2022 AMIA Board of Directors

Officers

Gretchen Purcell Jackson, MD, PhD, FACS, FACMI, FAMIA
Intuitive Surgical & Vanderbilt University Medical Center
Chair

Patricia C. Dykes, PhD, RN, FAAN, FACMI
Brigham and Women’s Hospital
Past Chair

Neil Sarkar, PhD, MLIS, FACMI
Brown University
Treasurer

Philip Payne, PhD, FACMI
Secretary

Directors

Julia Adler-Milstein, PhD, FACMI
University of California, San Francisco

Tiffany J. Bright, PhD
IBM Watson Health

William Brown III, PhD, DrPH, MA
University of California, San Francisco

Kenrick Cato, PhD, RN, CPHIMS, FAAN
Columbia University School of Nursing

Arlen Chung, MD, MHA, MMCi
University of North Carolina Chapel Hill

James J. Cimino, MD
University of Alabama at Birmingham

Laura K. Heermann Langford, PhD, RN, FAMIA
Intermountain Healthcare

Susan C. Hull, MSN, RN-BC, NEA-BC, FAMIA
CareLoop, Inc.

Deepti Pandita, MD, FACP, FAMIA
Hennepin Healthcare
Rachel Richesson, PhD, MPH, FACMI  
University of Michigan Medical School

Victoria L. Tiase, MSN, RN-BC  
NewYork-Presbyterian Hospital

Li Zhou, MD, PhD, FACMI, FAMIA  
Brigham and Women’s Hospital/Harvard Medical School

**Ex-Officio Board Members**

Genevieve Melton-Meaux, MD, PhD, FACMI  
American College of Medical Informatics President  
University of Minnesota

LaVerne Manos, DNP, RN-BC, FAMIA  
Academic Forum Executive Committee Chair  
University of Kansas Center for Health Informatics

Jonathan R. Nebeker, MD, MS  
Health Systems Chair  
Department of Veterans Affairs

Eileen Koski, MPhil, FAMIA  
Informatics Partnership Council Chair  
IBM Research

Jessica Ruff, MD, MA, MSPH  
Student Working Group Representative  
Case Western Reserve University/MetroHealth

Scott McGrath, PhD  
Working Group Steering Committee Representative

Tanya Tolpegin  
CEO
AMIA 2022 Informatics Summit Scientific Program Committee

**Chair**
Laura Wiley, PhD
University of Colorado

**Clinical Research Informatics Track**

**Vice Chair**
Yun Jiang, PhD, MS, RN
University of Michigan

**Members**
Thomas Campion, PhD
Weill Cornell Medicine

Licong Cui, PhD
The University of Texas Health Science Center at Houston

Vasa Curcin, PhD
King’s College London

Margo Edmunds, PhD, FAMIA
AcademyHealth

Eric Hall, PhD
Geisinger

Pey-Yun (Sabrina) Hsueh, PhD
Bayesian Health

Harold Lehmann, MD, PhD
Johns Hopkins

Abu Mosa, PhD
University of Missouri

Sanjay Purushotham, PhD
University of Maryland Baltimore County

Justin Rousseau, MD, MMSc
Dell Medical School

Hojjat Salmasian, MD, MPH, PhD
Harvard Medical School/Mass General Brigham

Neil Smalheiser, MD, PhD
University of Illinois School of Medicine

Ania Syrowatka, PhD
Brigham and Women's Hospital/Harvard Medical School

Kavishwar Waghokar, MD PhD
Harvard Medical School

Yanshan Wang, PhD
University of Pittsburgh

Asli Weitkamp, PhD
Vanderbilt University

Karmen Williams, DrPH, MBA, MA, MSPH, CPH
City University of New York

Adrian Zai, MD, PhD, MPH
University of Massachusetts Medical School

Yaguang Zheng, PhD
New York University

Data Science Track

Vice Chair
Mary Regina Boland, PhD, FAMIA
University of Pennsylvania

Members
Majid Afshar, MD, MS
University of Wisconsin-Madison

Theodoros Arvanitis, RT, DPhil, CEng
University of Warwick

Nrupen Bhavsar, PhD, MPH
Duke University

Daniel Capurro, MD, PhD
University of Melbourne
Translational Bioinformatics Track

Vice Chair
James Chen, MD
The Ohio State University

Members
Pawan Goyal, MD, MHA, MS, PGDHA
ACEP
Blanca Himes, PhD
University of Pennsylvania
Halil Kilicoglu, PhD
UIUC
Xia Ning, PhD
The Ohio State University
Arvind Rao, PhD
University of Michigan Ann Arbor
Dana Rollison, PhD
Moffitt Cancer Center
Nathan Seligson, PharmD
University of Florida
Vishakha Sharma, PhD
Roche Diagnostics
Casey Overby Taylor, PhD
Johns Hopkins University
Ece Uzun, PhD
Lifespan/Brown University
Kun-Hsing Yu, MD, PhD
Harvard Medical School
Zhongming Zhao, PhD, MS
University of Texas Health Science Center at Houston
<table>
<thead>
<tr>
<th>AMIA 2022 Informatics Summit Reviewers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbasi, Adeel</td>
</tr>
<tr>
<td>Abend, Aaron</td>
</tr>
<tr>
<td>Agarwal, Reita</td>
</tr>
<tr>
<td>Ahmed, Kamran</td>
</tr>
<tr>
<td>Ahmed, Zeeshan</td>
</tr>
<tr>
<td>Al Huneiti, Rasmeh</td>
</tr>
<tr>
<td>Alghamdi, Bushra</td>
</tr>
<tr>
<td>Alipanah, Neda</td>
</tr>
<tr>
<td>Allen, Katie</td>
</tr>
<tr>
<td>Amith, Muhammad</td>
</tr>
<tr>
<td>Anzalone, Alfred</td>
</tr>
<tr>
<td>Arbatti, Lakshmi</td>
</tr>
<tr>
<td>Arnaout, Rima</td>
</tr>
<tr>
<td>Arruda-Olson, Adelaide</td>
</tr>
<tr>
<td>Atnoor, Deven</td>
</tr>
<tr>
<td>Ator, Gregory</td>
</tr>
<tr>
<td>Baig, Furqan</td>
</tr>
<tr>
<td>Balasubramanian, Jeya</td>
</tr>
<tr>
<td>Balaji</td>
</tr>
<tr>
<td>Balyan, Renu</td>
</tr>
<tr>
<td>Banda, Juan</td>
</tr>
<tr>
<td>Bang, David</td>
</tr>
<tr>
<td>Bansal, Arvind</td>
</tr>
<tr>
<td>Barbour, Edward</td>
</tr>
<tr>
<td>Barrett, Laura</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></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></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Deshazo, Jonathan  Furo, Hiroko  Harris, Nomi
Dey, Vishal   Gabriel, Davera  Hasan, Md Mehedi
Dhond, Rupali  Gao, Grace  He, Linchen
Diederich, Catherine  Garcia-Milian, Rolando  Heider, Paul
Ding, Xiruo  Gartrell, Kyungsook  Hewitt, Kathleen
Doan, Son  Garza, Maryam  Hmwe, Susan
Doherty, Jennifer  Ge, Wendong  Ho, Joyce
Dorr, David  Gensheimer, Michael  Ho, Pei-Shu
Dos Santos, Fabiana  Ghalwash, Mohamed  Holl, Felix
Douthit, Brian  Giannaris, Pericles  Honcoop, Aubree
Drummond, Colin  Gold, Rachel  Hsiao, Allen
Edmiston, Scott  Gong, Yan  HU, Pengwei
Eilbeck, Karen  Gong, Yang  Hu, Ruifeng
Ezzeldin, Hussein  Gopalakrishnan, Vishrawas  Huang, Xiayuan
Falconer, Seanne  Gordon, Geoff  Huang, Yan
Fan, Yadan  Goeth, Rose  Hull, Susan
Feldman, Henry  Greene, Sarah  Hume, Sam
Fenton, Susan  Grove, Matthew  Hynes, Kelly
Fillmore, Nathanael  Guerrier, Christina  Israni, Sharat
Florez-Arango, Jose  Gupta, Samir  Jain, Nishant
Fort, Daniel  Hajagos, Janos  Jalloul, Nahed
Foster, Marva  Hamid, Zeyana  Jeanselme, Vincent
Fouladvand, Sajjad  Hammer, Richard  Jing, Xia
Foulis, Philip  Hannawi, Yousef  Joo, Jaehyun
Freimuth, Robert  Hanrahan, Larry  Jun, Inyoung
Fu, Sunyang  Harris, Debra  Jung, Hyunggu
Fultz Hollis, Kate  Harris, Marcelline  Jung, Jae-Yoon
Kamdar, Maulik
Kang, Tian
Kar, Reshma
Kayaalp, Mehmet
Kerns, Ellen
Khumrin, Piyapong
Kim, Min Soon
Kim, Yejin
Kim, Youngjun
King, Andrew
Klann, Jeffrey
Klasky, Hilda
Knosp, Boyd
Kogan, Yulia
Kohn, Martin
Kong, Sek Won
Kothari, Amit
Krichevsky, Spencer
Kumar, Sayantan
Kunney, Christopher
Kuo, Tsung-Ting
Kurc, Tahsin
Labilloy, Guillaume
Labkoff, Steven
Lai, Jiaying
Lee, Eunjung Sally
Lee, Younghee
Leftwich, Russell
Lenskaia, Tatiana
Li, Bin
Li, Dingcheng
Li, Dongkai
Li, Fang
Li, Qi
Li, Wentao
Li, Yanjun
Li, Yun
Liang, Jennifer
Liang, Lifan
Liang, Zhaohui
Liebman, Michael
Lim, Hansaim
Lim Choi Keung, Sarah
Lin, Rebecca
Liu, Hao
Liu, Mei
Liu, Nan
López-Campos, Guillermo
Lu, Dai-Yin
Lytle, Kay
Ma, Phillip
Ma, Rufeng
MacKelfresh, Andy
Madani, Sina
Magoc, Tanja
Mahajan, Satish
Mahendra, Malini
Maheshwari, Kamal
Makeda, Kai
Malec, Scott
Maleki, Sepideh
Manickam, Raj
Marcial, Laura
Maroille, Tatiana
Martins, Rute
Mason, Sarah
Mathew, Jomol
McCallion, Brian
McClay, James
McGilchrist, Mark
McPeek Hinz, Eugenia
Menser, Terri
Metke Jimenez, Alejandro
Mirmontes, Roque
Mironova, Maria
Mishra, Meenakshi
Mo, Huan
Mogharab Nia, Reyhaneh
Moldwin, Asher
Motiwala, Tasneem
Movahedi, Faezeh
Torii, Manabu
Trepp, Richard
Trinkley, Katy
Varner, Douglas
Vatani, Haleh
Wang, Lu
Wang, Peng
Wang, Zehai
Weir, Charlene
Weiskopf, Nicole
Weissenbacher, Davy

Wiley Jr, Ken
Williams, Marc
Wood-Wentz, Christina
Workman, T. Elizabeth
Wu, YiFan
Yan, Xiaowei (Sherry)
Yar, Wail
Ye, Jiancheng
Yee, Michelle
Yella, Jaswanth
Yoffie, Stephen

Yoon, Joo Heung
Yu, Zehao
Zawada, Stephanie
Zelle, David
Zhang, Jianqiu
Zhang, Lin
Zhang, Lingling
Zhang, Tianlin
Zhang, Wenhui
Zheng, Hua
Zitu, Md Muntasir
Automated and accessible prediction of progression in age-related macular degeneration (AMD): external validation from the perspectives of methods, populations, and users

Qingyu Chen, PhD¹, Tiarnan D. L. Keenan, BM BCh, PhD², Alexis Allot, PhD³, Michael F. Chiang, MD, MA², Michelle R. Hribar, PhD³, Emily Y. Chew, MD³, Zhiyong Lu, PhD²†

1. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
2. National Eye Institute, National Institutes of Health, Bethesda, MD, USA
3. Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, USA
† To whom correspondence should be addressed: zhiyong.lu@nih.gov

Introduction

Age-related macular degeneration (AMD) is the leading cause of blindness in developed countries. By 2040, 288 million patients are projected to have AMD worldwide [1]. Based on clinical features, it is classified into early, intermediate, and late stages. In some people with early/intermediate AMD, the disease progresses to the late stage slowly; in others, it advances faster and may quickly lead to loss of vision in one or both eyes. There is, therefore, a critical need to predict progression to late AMD, to provide information regarding secondary preventive therapy such as anti-oxidant vitamins and minerals for intermediate AMD and prediction of the wet form of late AMD for prompt therapy, and to recommend medical and lifestyle interventions in a personalized way.

Current clinical practice employs two primary methods for predicting risk of progression to late AMD. One method uses the Simplified Severity Scale (SSS), proposed following the Age-Related Eye Disease Study (AREDS), a primary clinical trial on AMD and other ophthalmic diseases for decades [2]. A score of 0-4 is generated for the individual by grading the presence of two risk factors – large soft drusen and pigmentary abnormalities – in the central retina of each eye, based on color fundus photographs (CFP). Each step on the scale is then mapped to a 5-year probability of progression to late AMD (in either eye). For example, if the individual has a score of 1 on the scale, the 5-year probability of progression to late AMD is 3.0%. The other method uses an online risk calculator, which is similar to the AREDS SSS [3]. However, both approaches require manual grading of the risk factors, which is time-consuming and requires highly trained graders (i.e., retinal specialists). It is thus critical to develop automated systems to facilitate AMD risk prediction.

Deep learning-based models have achieved remarkable progress for automatic ophthalmic disease diagnosis and prognosis over the last few years, which have potential to facilitate the clinical decision-making [4]. However, despite remarkable progress, most studies of deep learning models in ophthalmology have evaluated the performance on only local datasets. A recent review [5] on deep learning models in ophthalmology showed that only 29 out of 82 studies reported external validation performance, and only five out of 19 studies on AMD reported external results. In addition, existing studies in other clinical medicine domains have demonstrated that factors such as gender, age, and race may substantially decreased the generalizability of deep learning models [6-8]. The impact of these factors on the potential generalizability of deep learning systems in ophthalmology remain largely unknown. In response, in this study, we plan to conduct extensive external validations on an AMD progression prediction model that we developed previously. We herein summarize the methods, results, and proposed external validations.

AMD progression prediction model, evaluation results, and proposed validations

Methods. We developed an AMD progression prediction model, as shown in Figure 1(I) [9]. Given the color fundus photographs (CFPs) of both eyes as inputs, the model predicts the AMD progression risk at a given year. It firstly employs convolutional neural networks (CNNs) to classify the severity level of macular drusen and pigmentary abnormalities; then it applies a survival model using the classified macular drusen and pigmentary abnormality features and demographic information to predict the risk of progression to vision threatening late AMD in subsequent years. The model was trained and validated on the AREDS and AREDS2 datasets (consisting of ~80,000 images from ~3,000 patients) [2]. We also developed a software prototype for local use, as shown in Figure 1(II).

Figure 1. Overview of our evaluation pipeline. Stage I: we will conduct external validations on the AMD progression prediction model; Stage II: we will conduct usability studies on the prototype. Stage III: we will summarize and reflect the related health impacts from data, methods, and users.

12
Table 1. Representative evaluation results from [9]. The 5-year C-statistic (95% CI) results of models trained and tested on AREDS and AREDS2. Retinal specialists/calculator: the grades from 88 (AREDS) and 192 (AREDS2) retinal specialists are used as input for the risk calculator.

<table>
<thead>
<tr>
<th>Model Type</th>
<th>AREDS and AREDS2 Combined</th>
<th>AREDS</th>
<th>AREDS2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our model</td>
<td>86.6 (85.8,87.4)</td>
<td>71.0 (70.2,71.7)</td>
<td></td>
</tr>
<tr>
<td>Train and test on the combination of AREDS and AREDS2</td>
<td></td>
<td>83.1 (82.8,83.3)</td>
<td>63.9 (63.2,64.6)</td>
</tr>
<tr>
<td>Retinal specialists/calculator</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Evaluation results. Overall, the results demonstrate that the model achieved high prognostic accuracy that substantially exceeded that of retinal specialists using those two clinical standards. Table 1 shows the representative evaluation results in our previous study [9]. Specifically, the ground truth labels were assigned by expert human graders at the University of Wisconsin Fundus Photograph Reading Center; C-statistic was used as the main evaluation metric; and we compared the model performance with retinal specialists using the risk calculator, the standard clinical practice as mentioned above. We evaluated performance in two different ways of dataset combinations for training and testing: (1) using the two datasets combined, and (2) training on one dataset and testing on the other dataset to mimic external validation. Both settings demonstrated the model achieved better performance than that of retinal specialists/calculator; for instance, the model had a 5-year C-statistic of ~7% higher than that of retinal specialists using the online calculator approach on AREDS2. Second, the performance of the model trained on the AREDS only had substantially lower performance when testing on the AREDS2 (which is expected due to population differences); this also shows the importance of external validation.

Proposed validations. We commit to all three stages of the evaluation showcase; our proposed plan is summarized in Figure 1. For Stage I, we will focus on validating the effectiveness of the AMD progression prediction model on external data and analyzing performance bias on subgroups. Specifically, we will conduct external validations on the imaging repositories at the Casey Eye Institute at Oregon Health & Science University (OHSU) and the University of Wisconsin Fundus Photograph Reading Center. In addition, we used the Inception-V3 network [10] as the CNN backbone of the model, since it achieved state-of-the-art performance at that time. However, more advanced architectures such as DenseNet and Vision Transformers have since been proposed and shown superior performance. The frameworks for the implementations of deep learning models (such as TensorFlow and PyTorch) have also progressed to more mature versions. Therefore, we plan to redeploy the method using the latest model architectures and framework during Stage I as well.

For Stage II, we will focus on validating the efficiency and analyzing the usability of the prototype for users. Specifically, we will measure the inference speed using the prototype and user satisfaction. Potential users include both clinicians who will use the prototype to facilitate the decision-making process and trainees who will study artificial intelligence related topics. For Stage III, we will plan accordingly to the results of Stage I and II and reflect and summarize the impacts of data, method, and users. In addition, throughout the three-stage study, we aim to answer three central questions – from the perspectives of the methods, populations, and users:

At the methods level, how do different methods contribute to the generalizability? We aim to compare the performance of different architectures (e.g., DenseNet and Vision transformers, as mentioned above) and different training strategies (such as multi-task training and meta-learning) at the external setting. This will be primarily conducted in Stage I; we will also reactively update the model and prototype throughout all three stages.

At the population level, how does the model perform in different subgroups? We aim to compare performance on subgroups by gender, age, and race. As mentioned, few studies have conducted such analyses in ophthalmology. Its major part will be conducted in Stage I; we will also evaluate at later stages when datasets from other external sites are also available.

At the user level, how does the model facilitate the AMD progression prediction process? We aim to conduct user studies on clinicians and trainees to evaluate the usability of the prototype in terms of inference speed and user satisfaction. This will be primarily conducted in Stage II; we may also continue the evaluation or conduct follow-up analysis in Stage III when needed.

Conclusion

In this study, we introduce AMD progression prediction, explain the importance of conducting external validations, and propose a validation approach. The proposed validations are oriented to the perspectives of methods, populations, and users. It has the potential to assist decision-making in both clinical practice and research environments.

Acknowledgments. This work was supported by the Intramural Research Programs of the National Institutes of Health, National Library of Medicine and the National Eye Institute, NIH/NLM R01LM013426, and NIH/NLM R21LM013937.
Lessons from Developing and Deploying a Real-Time Sepsis Prediction Model for Hematopoietic Cell Transplant Recipients

Zahra Eftekhari, MS¹, Cameron S. Carlin, MS¹, Janet Munu, MSN, MBA¹, Tushondra Thomas, MBA, DNP¹, Sanjeet Dadwal, MD¹, Ryotaro Nakamura, MD¹

¹City of Hope National Medical Center, Duarte, California, United States

Introduction

Sepsis is a leading cause of death in HCT recipients post-transplant. It is difficult to detect sepsis with commonly used models in these patients, as sepsis presentation may differ from general patient population. A sepsis predictive model had to be trained and evaluated on this specific patient population. To deploy the model to help with decision making at point of care, a real time data infrastructure needed to be developed to identify HCT patients, capture data from Epic to generate predictions, send predictions to Epic and prompt notifications for clinicians in real time.

Methods

The machine learning model was trained on 1,806 patients who had inpatient HCT transplants at City of Hope National Medical Center from 2014-2017. The primary clinical event was sepsis diagnosis (n=132) during the transplant encounter, and time of sepsis was defined as the time sepsis treatment started. Model features included flowsheet vitals, administered medications, laboratory results, and historical treatment information. Missing data was handled by forward filling from previously observed values and filling unknown values to population mean. Data from 2014-2016 was used to train the model via cross-validation; data from 2017 was used as holdout data.

Most studies evaluated the performance of the model at a snapshot of time, such as time of sepsis, where the models were retrospectively applied to the data at time of sepsis\(^4\). If the models were able to detect sepsis at this time, it was defined as true positive, and if they did not detect sepsis, it was defined as false negative. The performance metrics from this evaluation method do not necessarily reflect how the models would perform in real world scenarios in inpatient settings, where models are not used or evaluated at snapshots of time.

Metrics such as lead time were developed to help simulate real world performance. Real world performance was simulated on holdout data by applying the model on an hourly basis from patients’ admission to sepsis diagnosis or discharge, whichever occurred earlier. For patients who developed sepsis, if the risk passed a certain threshold at any time, patient was flagged as a true positive; if risk never passed the threshold they were counted as a false negative. For non-septic patients, if risk passed the threshold at any time during their stay, patients counted as a false positive; if the risk never passed the threshold, they counted as a true negative.

Results

The random forest model achieved an AUC of 0.83 on training and 0.82 on the holdout dataset. Sepsis is relatively uncommon among HCT recipients (<10% incidence); even a low false positive rate can translate into many false alarms in practice. Therefore, precision-recall curve and Average Precision were considered for model evaluation. The model achieved 0.57 average precision on training and 0.36 average precision on holdout. The difference between average precision on training and holdout data was caused by the difference in sepsis prevalence in the two datasets. The median lead time at threshold of 0.6, with 40% sensitivity and 93% specificity was 35 hours (range: 1-2200, interquartile range: 275.75) for the training dataset and 47 hours (range: 1-1954) for the holdout data.

Discussions

The model was deployed in July 2020. Best-practice advisory (BPA) workflows were implemented, with different workflows for physicians and nurses. It is important to value provider experience and make sure notification fatigue is not outweighing decision support benefits. Post go-live surveys have been developed that periodically ask end users for feedback on the model. Technical metrics like ROC and AUC are monitored, but these metrics proved less useful in monitoring performance in production. One reason for this is the fact that the model affects the outcome through interventions. As an example, some patients are correctly identified as high risk, and an intervention is done, leading to the patient never developing symptoms. The model has served its purpose, but these patients will be perceived as false positives and hurt the model’s AUC score in prospective evaluation. Ideally, there should be statistically significant reduction in associated clinical metrics, such as ICU transfers, mortality, and length of stay. The issue with monitoring such metrics is that a causal relationship between improvements and the model cannot be assumed, as there are parallel efforts within the hospital to improve these metrics.


Hossein Estiri, PhD1,2, Zachary H Strasser, MD1,2, Jeffery G Klann, PhD1,2, Kavishwar B Wagholikar, MBBS, PhD 1,2, Shawn N Murphy, MD, PhD1,2
1Massachusetts General Hospital; 2Mass General Brigham, Boston, MA

Introduction
The healthcare research and industry have been increasingly progressive on the translation and implementation of artificial intelligence (AI) to improve outcomes and lower costs. Diligently identifying and addressing biases in AI/ML algorithms (hereafter, referred to as “algorithms”) have garnered widespread public attention as press[ing ethical and technical challenges.1,2 The Algorithmic Accountability Act of 2019[3] requires businesses to evaluate risks associated with algorithm fairness and bias. Nevertheless, regulating algorithmic bias in healthcare remains a tall task.

Deploying AI algorithms in healthcare carelessly could exacerbate the very health inequalities society is working to address. Algorithmic bias in healthcare may be caused by missing data (e.g., higher rates of missingness in minority populations due to decreased access to healthcare or lower healthcare utilization), observational error, misapplication, and overfitting due to small sample sizes or limited population and practice heterogeneity. Despite the eminent work in other fields, bias often remains unmeasured or partially measured in healthcare domains. Most published research articles only provide information about very few performance metrics -- mostly through measures of the algorithm’s discrimination power, such as the Area Under the Receiving Operating Characteristics Curve (AUROC). The few studies that officially aim at addressing bias, usually utilize single measures (e.g., model calibration[4]) that do not portray a holistic picture of bias. Proper evaluation of bias in medical AI requires a holistic evaluation framework that enables diligent search for unrecognized bias and invigorate follow-up investigations to identify the underlying roots of bias.

In this podium abstract, we present results from applying a comprehensive AI evaluation framework to a set of AI prediction models developed and validated retrospectively during the first six months of the COVID-19 pandemic. The models predict risks of mortality, hospitalization, ICU admission, and ventilation due to COVID-19 infection. We utilize our AI evaluation framework, which we have incorporated in MLHO[5] (an AI pipeline for modeling health outcomes), to identify unrecognized biases from multiple perspectives.

Methods
The goal of this study is to identify unrecognized bias in the four validated prediction models of COVID-19 outcomes to investigate whether (a) the models were biased when developed (we refer to this as a retrospective evaluation) and (b) the bias changed over time when applying the models to new COVID-19 patients who were infected after the models were trained (we refer to this as a prospective evaluation). To do so, we used data from 56,590 Mass General Brigham (MGB) patients, with a positive reverse transcription-polymerase chain reaction (RT-PCR) test for SARS-CoV-2 between March 2020 and September 2021 (Figure 1). To perform a comprehensive evaluation of the predictive models’ performance, we developed a module for MLHO that applies several standard model-level metrics for discrimination (e.g., area under the ROC/precision-recall curve), accuracy (e.g., Brier score), and reliability (a.k.a., calibration). In addition, the AI evaluation module also computes the Mean Absolute Error (MAE) of predictions for each patient to identify and illustrate model performance from an individual level perspective, when the variable of interest is
continuous (e.g., patient age). To visualize the MAE patterns, we plot the continuous variables of interest on the X-axis and the MAE on the Y-axis and fit a generalized additive model (GAM) with integrated smoothness.

Results

Data from 56,590 patients with a positive COVID test were analyzed -- >15,000 in the retrospective cohort (3/2020-10/2020) and >41,000 in the prospective cohort (11/2020-8/2021). Compared to the overall population, retrospectively and prospectively across time, the models marginally performed worse for male patients and better for Hispanic and female patients, as measured by AUROC and Brier scores (Figure 2). The range of delta between these performance metrics within demographics groups and the overall model was relatively small. For the rest of the demographic groups, the performances were more mixed. The models that were developed with data from March to September 2020, provided relatively stable predictive performance prospectively up until May-June 2021. Despite the increased variability, the prospective modeling performance remained high for predicting hospitalization and the need for mechanical ventilators. From the reliability/calibration perspective, except in the case of prospective evaluation of hospitalization predictions among Hispanic, female, and black patients, the diagnostic reliability diagrams did not show consistent bias towards or against a certain group. From the individual level, we found consistent bias in increasing error rates for older patients.

Discussion

The premise for evaluating these predictive models was to create a framework for discovering and quantifying the various types of biases towards different subgroups that were encoded unintentionally. Given that we face systemic bias in our country’s core institutions, we need technologies that will reduce these disparities and not exacerbate them. There are efforts from the larger AI community, (e.g., AI Fairness 360[6]) to develop open-source software systems for measuring and mitigating bias. These programs are often ad-hoc or work as standalone post-processing solutions. We plan to compare these model-independent methods and add relevant functionalities to our domain-specific approach. We have incorporated the presented bias measurement framework within the MLHO pipeline,[5] which is specifically designed for modeling clinical data. Providing means to evaluate and address unrecognized bias within a data-centric pipeline will enable the generation of medical AI that takes into account various biases while in production. Only a holistic evaluation framework that enables diligent search for unrecognized bias can provide enough information for an unbiased judgment of AI bias that can invigorate follow-up investigations on identifying the underlying roots of bias and ultimately make a change.

References

Emergency Department Wait Time Prediction based on Cyclical Features by Deep Neural Networks

Zhaohui Liang, PhD1, Jimmy Xiangji Huang, PhD2
1Dept. of Electrical Engineering & Computer Science, York University, Toronto, ON, Canada; 2School of Information Technology, York University, Toronto, ON, Canada

Study Objective:
An effective estimate of patient wait time at emergency departments (ED) significantly improves working efficiency and reduce unnecessary crowding in hospitals globally1. In this study, we present different types of deep neural network models to predict the ED wait time based on the cyclical features extracted from the time stamp of the electronic health records (EHR). The real application indicates that using cyclical features as time series patterns can effectively predict the ED wait time within a 30-minute window.

Current Technical Gap for ED Wait Time Prediction:
The current algorithms based on time series prediction mainly relies on empirical data from the healthcare systems with insufficient details that impacts on the wait time. As a result, the models such as linear and logistic regression, autoregressive integrated moving average (ARIMA), seasonal ARIMA (SARIMA), and recurrent neural network (RNN), etc. can predict the ED wait time well within the training and test datasets2,3. However, their performance is likely to be downgraded when these models are deployed for real-world setting.

Methods:
Unlike other EHR time series data such ECG (electrocardiography) or body temperature records, the ED wait time data usually does not have an identical time interval. Therefore, to generate time series patterns from these ED records, we extract the cyclical features from the timestamps of the ED visit records by applying the sine and cosine functions. At first, we convert all time stamp information into seconds. Then we generate six pairs of cyclical features given six time-intervals: 30 minutes, 1 hour (60 minutes), 1 day, 1 week, 1 month, and 1 year with all quantitative units in second. Then the cyclical patterns are generated based on the following formula:

\[ \sin(\text{timestamp} \times \frac{2\pi}{\text{time interval}}) \] or \[ \cos(\text{timestamp} \times \frac{2\pi}{\text{time interval}}) \]

After the cyclical feature are produced, we split the data with a window of 28 days with step=1 to generate the sequence data. Six deep neural network (DNN) models are used for training and comparison, together with a baseline model which uses the next step prediction as predicted value and a logistic regression model. The six DNNs include a two dense layers (128 neurons for each) network for single step inputs, two dense layers DNN with multiple step input, convolutional DNN, a long short-term memory (LSTM) DNN, and residual LSTM, and a LSTM with convolution layers multiple step input. All networks are optimized by the Adam optimization algorithm with initial learning rate of \( 2 \times 10^{-3} \) for 200 epochs with GPU.

Results:
After all models are optimized, we find the impacts of different cyclical features by visualizing the weights of from the logistic model. It indicates that though the sequence of the ED wait time values have the highest impact, the transferred cyclical features have different impacts toward the final prediction. These patterns can be illustrated by the Fast Fourier transfer of all cyclical frequencies (Figure 1). In addition, by observing the autocorrelation of the ED wait time pattern, we can observe the wait time has approximately weekly periodic patterns. These two findings both indicate the extraction of the cyclical features based on different time intervals are effective to reflect the periodic change patterns of the ED wait time.

By comparing the mean absolute error (MAE) of the prediction by different models, we find that the DNN model with convolutional layers, and LSTM model, and the LSTM with convolution layer have superior performance over the other models, in which the LSTM model has the lowest MAE (Figure 2). The MAE values are shown in Table 1.
Table 1. Mean absolute error (MAE) of predictions.

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE in validation</th>
<th>MAE in test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1.5341</td>
<td>0.7455</td>
</tr>
<tr>
<td>Linear (Logistic Regression)</td>
<td>0.9396</td>
<td>0.7010</td>
</tr>
<tr>
<td>DNN with single step input</td>
<td>0.8741</td>
<td>0.6942</td>
</tr>
<tr>
<td>DNN with multiple step input</td>
<td>0.8869</td>
<td>0.6937</td>
</tr>
<tr>
<td>Convolutional DNN (Conv)</td>
<td>0.8502</td>
<td>0.6703</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.8573</td>
<td>0.6584</td>
</tr>
<tr>
<td>Residual LSTM</td>
<td>2.0655</td>
<td>1.0173</td>
</tr>
<tr>
<td>Conv LSTM</td>
<td>0.8739</td>
<td>0.6895</td>
</tr>
</tbody>
</table>

Discussion of Results

The above results imply that using the cyclical feature extracted from time stamps can predict the ED wait time. In practice, the feedback from the clinical staff indicates the ensembled models combined with the four best DNN models from the experiment can effectively reduce the estimate error within a 30-minute window, and its performance is superior to the LSTM models only trained by directly retrieved wait time data.

Conclusion

The application of cyclical patterns extracted from EHR time stamps serve as useful features to optimized time series models such as DNNs to successfully predict ED wait time. The errors from uneven time interval can be reduced by transferring the observed features into cyclical feature. We believe this strategy will improve the AI performance in the healthcare systems.

Acknowledgment

This work is supported by the NSERC and the YRC program.

References

Leveraging Explainable Machine Learning to Predict New Onset of ICU Delirium

Siri Liu, PhD¹, Allison B. McCoy, PhD¹, Joseph J. Schlesinger, MD, FCCM², Adam Wright, PhD¹
¹Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN; ²Department of Anesthesiology, Division of Critical Care Medicine, Vanderbilt University Medical Center, Nashville, TN

INTRODUCTION

Delirium is an acute decline in cognitive function leading to confusion and emotional disorganization, which occurs in 29% to 65% of hospitalized older patients.¹ Delirium is a serious problem, resulting in higher mortality, in-hospital falls, need for long-term care, and other adverse outcomes.² Prevention is considered the most effective way to manage delirium,³ and more than two-thirds of delirium cases are preventable.⁴ Previous studies predicting delirium have used relatively small datasets of clinical trials, used less reliable prediction labels, or used traditional modeling techniques (e.g., logistic regression), contributing to an insufficiently ability for clinical use.⁵ The performance of explainable machine learning models that directly use electronic health record (EHR) data to predict the new onset of delirium remains unclear. We sought to more accurately and reliably predict new onset of delirium using a larger, EHR-based dataset and provide a visualization of explanation for individual prediction. This study was approved by the Vanderbilt University institutional review board.

METHODS

At Vanderbilt University Medical Center (VUMC), a large medical center in Nashville, TN, delirium is assessed every 12 hours by the validated Confusion Assessment Method for the ICU (CAM-ICU).⁶ We extracted data from Clarity, the Epic clinical data warehouse at VUMC, including adult patients with CAM-ICU assessments performed between 2018-01-01 and 2021-10-01. Extracted features were informed by previous studies regarding risk factor for delirium.⁷ We excluded assessments completed less than 12 hours after arrival to the unit and subsequent assessments after new onset of delirium. The prediction outcome was new onset of delirium, (i.e., a new positive CAM assessment result). We collected the latest values before three prediction periods: 6, 12, and 24 hours, to align for the shift change. The pre-processing process included three steps: 1) imputing missing values, 2) scaler, and 3) encoding categorical features. We split the dataset into training (80%) and testing sets (20%) at the patient level. We used 5-fold cross validation and reported the results using the hold-out testing dataset with 1000 rounds bootstrapping. We developed logistic regression (LR), random forest (RF), support vector machine (SVM), and LightGBM models. We reported outcomes in F1, accuracy, area under the receiver operator characteristic curve (AUC), recall, and precision. We applied the Shapley additive explanations (SHAP)⁸ to analyze feature importance globally and locally.

RESULTS

A total of 34,035 patients with 331,489 CAM-ICU assessments and 896 features were included in the final dataset: demographics (13), medications (195), vital signs (12), laboratory values (39), active problems (161), historical problems (84), surgical history (108), social history (5), and procedure history (279). The median age was 59 years with an interquartile range (IQR) of [44, 70]. The median length of stay was 3 [2, 6]. Among assessments, 37,246 were positive (11.2%). 5019 (14.7%) patients were Black/African American and 14140 (41.5%) patients were female. The LightGBM model outperformed the LR, SVM, RF, and NN models. The LightGBM model achieved the best performance when the prediction time window was 6 hours, which had an AUC score of 0.927 (95%CI: 0.924, 0.929) and an F1 score of 0.626 (95%CI: 0.618, 0.634). Additional results are reported in Table 1. In the global view, the top five predictors were lab values of ammonia level, the number of injectable anesthetics received, heart rate, Richmond Agitation-Sedation Scale (RASS) score, and fall risk score. Using the SHAP explainer, we can visualize the key contributed factors in an individual patient’s prediction.

Table 1. Results on the testing dataset using different time windows (h: hours).

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Precision</th>
<th>Accuracy</th>
<th>F1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR (12h)</td>
<td>0.759 [0.754, 0.773]</td>
<td>0.385 [0.379, 0.392]</td>
<td>0.825 [0.822, 0.828]</td>
<td>0.511 [0.506, 0.520]</td>
<td>0.879 [0.874, 0.883]</td>
</tr>
<tr>
<td>SVM (12h)</td>
<td>0.372 [0.368, 0.386]</td>
<td>0.62 [0.620, 0.646]</td>
<td>0.897 [0.896, 0.901]</td>
<td>0.465 [0.464, 0.481]</td>
<td>0.88 [0.878, 0.889]</td>
</tr>
<tr>
<td>RF (12h)</td>
<td>0.211 [0.203, 0.221]</td>
<td>0.797 [0.780, 0.812]</td>
<td>0.898 [0.896, 0.901]</td>
<td>0.334 [0.323, 0.346]</td>
<td>0.908 [0.902, 0.91]</td>
</tr>
<tr>
<td>NN (12 h)</td>
<td>0.405 [0.396, 0.416]</td>
<td>0.392 [0.382, 0.402]</td>
<td>0.853 [0.850, 0.856]</td>
<td>0.399 [0.390, 0.408]</td>
<td>0.772 [0.767, 0.778]</td>
</tr>
<tr>
<td>LightGBM (12h)</td>
<td>0.752 [0.742, 0.761]</td>
<td>0.526 [0.517, 0.535]</td>
<td>0.888 [0.886, 0.891]</td>
<td>0.619 [0.611, 0.627]</td>
<td>0.921 [0.918, 0.923]</td>
</tr>
<tr>
<td>LightGBM (6h)</td>
<td>0.75 [0.740, 0.759]</td>
<td>0.537 [0.527, 0.546]</td>
<td>0.892 [0.889, 0.894]</td>
<td>0.626 [0.618, 0.634]</td>
<td>0.927 [0.924, 0.929]</td>
</tr>
<tr>
<td>LightGBM (24h)</td>
<td>0.694 [0.684, 0.704]</td>
<td>0.48 [0.471, 0.489]</td>
<td>0.872 [0.870, 0.875]</td>
<td>0.568 [0.560, 0.576]</td>
<td>0.893 [0.890, 0.897]</td>
</tr>
</tbody>
</table>
DISCUSSION
We demonstrated that structured EHR data could be used to predict new onset of delirium via explainable machine learning models. The performance of the LightGBM model in a 6-hour time window outperformed the previous machine learning prediction models for delirium with reported AUC ranging from 0.82 to 0.86. Moreover, our use of the CAM as the prediction label is more accurate and clinically relevant than prior studies, which used billing codes and drastically underestimated the prevalence of delirium. In addition, we used mean values to impute missing values in other models; however, consistent with previous work, the LightGBM model performed better without the imputations.

Based on this high performance, we will implement the LightGBM predictive model into real-world clinical practice through a workflow-embedded clinical decision support (CDS) tool that will be used by nurses and the ICU care team (attending, fellow, residents, and advanced practice providers). The CDS tool will be delivered through an indicator column in the patient list of our Epic EHR system (Figure 1), which allows users to hover over to see more detailed information, including modifiable and unmodifiable factors contributing to the prediction. The model will be updated at least three times a day, one hour before each shift change. For high-risk patients, providers will follow the ICU Liberation Bundle to take relevant interventions.

In the next two stages, we will develop and evaluate the CDS tool. In Stage II, we will conduct a user-centered design with iterative prototypes to investigate clinicians’ needs for explainable machine learning, a usability testing and an 8-week theory-based implementation in the ICU settings at VUMC. An adapted technology acceptance model in the CDS context will be used to inform the design and implementation. In Stage III, we will perform a mixed-methods study to evaluate the performance through the pre-post interventional trial. The primary outcome is incidence rate of delirium. Secondary outcomes include the total number of days with delirium, the severity of delirium, recurrence rates, length of stay, falls, mortality, readmission rates, new discharges to long-term care, adherence rate, the days of mechanical ventilation, user’s satisfaction, the ability to prevent delirium when warned, and user responses.

We presented a high-performance model to predict the new onset of delirium and visualized risk factors for each prediction. It provides a solid technical basis for the development of a CDS tool and further implementation in clinical practice to efficiently improve the prevention of delirium.

ACKNOWLEDGEMENTS: The authors do not have conflicts of interest related to this study. This work was supported by NIH grant: R01AG062499-01.

References
Scaling Up and Evaluation of PhenoTagger for Deep Phenotyping on Two External EHR Datasets

Ling Luo, PhD1, Li Fang, PhD2, Shankai Yan, PhD1, Daniel Veltri, PhD3, Andrew Oler, PhD3, Rajarshi Ghosh, PhD3, Morgan Similuk, ScM3, Kai Wang, PhD2, Zhiyong Lu, PhD1
1National Library of Medicine, Bethesda, MD 20894, USA; 2Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children’s Hospital of Philadelphia, Philadelphia, PA 19104, USA; 3National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA

Genetic diseases are the leading cause of child mortality in the United States. Particularly for genetic diseases with a progressive course, genetic diagnosis must be equally fast to inform interventions. Unfortunately, a standard diagnosis pipeline involves complicated manual phenotyping, which is expensive and time-consuming. Recent studies have shown (semi-)automated analysis has the potential to greatly accelerate molecular diagnosis with significantly reduced labor intensity. Automated deep phenotyping is a crucial step in the AI-assisted disease diagnosis workflow.

Automated phenotype concept recognition remains challenging in clinical NLP research despite a few past studies where both dictionary-based and machine learning-based methods were attempted. The dictionary-based methods can achieve high precision in general but suffer substantially in recall. The lack of comprehensive annotated training data greatly limited the performance of machine learning approaches. In response to these problems, we recently developed PhenoTagger, a hybrid phenotype recognition method that combines dictionary and BioBERT-based methods to recognize Human Phenotype Ontology (HPO) concepts in unstructured biomedical text. PhenoTagger is an innovative ontology-driven method without requiring any manually labeled training data and it was validated on two separate gold-standard datasets. Experimental results show that PhenoTagger compares favorably to all existing methods on both datasets and improves the previous best F1-score by over 10%.

Despite PhenoTagger’s promising results, the complexity of the BioBERT model (https://github.com/dmis-lab/biobert) may be a limiting factor for its application at scale, an important and general issue that affects many existing clinical NLP tools that are based on large-scale pre-trained models in practical use. Hence, we herein propose improving the efficiency of PhenoTagger by experimenting with a smaller model while retaining its accuracy. Compared with the original PhenoTagger, we use Bioformer (https://github.com/WGLab/bioformer/) instead of BioBERT as it is a lightweight BERT model that was recently created from scratch—using only biomedical vocabulary and datasets—without transfer learning.

We first evaluated the improved PhenoTagger on its previous benchmarking datasets and found that the new PhenoTagger can maintain its accuracy with approximately three times faster processing speed. Moreover, two distinct EHR datasets were used to assess its generalizability and robustness on real-world clinical application scenarios: (1) an EHR dataset from National Institute of Allergy and Infectious Diseases (NIAID) that includes all longitudinal clinical notes from ~4,500 pediatric and adult patients with suspected inborn error of immunity who were seen at the NIH Clinical Center and subsequently underwent genomic evaluation. HPO terms for a subset of 82 of these patients have been manually compiled. (2) an EHR dataset from the Children’s Hospital of Philadelphia (CHOP) that includes longitudinal clinical notes of ~100 patients with a diverse range of Mendelian diseases. HPO terms for each note have been manually compiled. Benchmarking on the NIAID and CHOP EHR subsets shows that PhenoTagger achieves an average score of 0.82 and 0.77 in sensitivity, respectively. These new results are consistent with those previously reported in paper1 and as in paper1 exact precision cannot be obtained due to incomplete reference annotations.

In summary, we improved our PhenoTagger with a newly developed scalable language model, allowing it to be applied to two external EHR datasets with high accuracy and efficiency. These results further suggest that automated deep phenotyping using PhenoTagger holds great potential to assist decision-making in clinical diagnosis and patient care. In subsequent studies, we plan to measure the health impact of our AI-based tool by quantifying whether it leads to improved and faster diagnosis and treatment.

Acknowledgment: This work was supported by the Intramural Research Program, National Library of Medicine.

Technical Performance Evaluation of a Deep Learning-based Vancomycin Therapeutic Drug Monitoring Model

Bingyu Mao, MA1, Masayuki Nigo, MD1,2, Hong Thoai Nga Tran, MS3, Ziqian Xie, PhD1, Han Feng, PhD3, Laila Rasmy, PhD1, Hongyu Miao, PhD3, Degui Zhi, PhD1,3

1School of Biomedical Informatics, Univ. of Texas Health Science Center at Houston, TX
2Division of Infectious Diseases, UTHealth McGovern Medical School, Houston, TX
3School of Public Health, Univ. of Texas Health Science Center at Houston, TX

Introduction
Vancomycin (VAN) is a commonly used antimicrobial in hospitals and therapeutic drug monitoring (TDM) is required to optimize the efficacy and avoid toxicities. Bayesian models are currently recommended to predict VAN serum levels. However, these models were often developed in limited patient populations using carefully designed lab observations. We developed a deep-learning-based pharmacokinetic prediction model for vancomycin (PK-RNN-V) which takes the patient’s real-time sparse and irregular observations from structured electronic health record (EHR) and offers dynamic predictions. The main objective of this study is to evaluate the prediction accuracy of the PK-RNN-V model and its variants using standard benchmark measures and compare it with the Bayesian models.

Methods
5,483 patients with 9,504 encounters who received VAN from Memorial Hermann Hospital System were included in this study. The dataset was split 70:15:15 for training, validation, and test sets. We chose a population-level Bayesian model VTDM to be our baseline. The VTDM feedback model can adjust the PK parameters according to the first measurements of VAN level. The original PK-RNN-V models predicted individual patient volume distribution (v) and VAN elimination (k) at each time step using an irregular timesteps GRU model. To have a fair comparison with VTDM, one variant of PK-RNN-V (PK-RNN-V feedback) only uses the first measurement to adjust its hidden state. The variant that takes all measurements available as PK-RNN-V full feedback. An ablation study (PK-RNN-V without kidney function) was completed by excluding the kidney function-related features. Additionally, we use model ensembling (PK-RNN-V E) to get a posterior estimate of the vancomycin concentration. We chose to use three ways for evaluation: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE).

Results and discussion
All PK-RNN-V models exhibited better RMSE, MAE, and MAPE compared to any of the VTDM models; The baseline PK-RNN-V already showed good performance, PK-RNN-V: RMSE of 5.86, MAE of 4.09, MAPE of 37.57, VTDM model: RMSE of 8.58, MAE of 6.54, MAPE of 41.81. PK-RNN-V performed better at predicting the first measurement of VAN level, compared to VTDM (RMSE: 5.22 vs. 7.09, MAE: 3.87 vs. 4.85, MAPE: 32.95 vs. 65.18, respectively). The evaluation results revealed the superior performance of PK-RNN-V models when compared with traditional Bayesian VTDM models (Table 1).

Optional illustrations

Table 1. Model Performance Comparing Different Types of PK-RNN and Bayesian Models

<table>
<thead>
<tr>
<th>Model Type</th>
<th>RMSE (mg/L)</th>
<th>MAE (mg/L)</th>
<th>MAPE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian Model - VTDM</td>
<td>8.58</td>
<td>6.54</td>
<td>41.81</td>
</tr>
<tr>
<td>Bayesian Model - VTDM with Feedback</td>
<td>6.29</td>
<td>4.26</td>
<td>29.15</td>
</tr>
<tr>
<td>PK-RNN-V model - PK-RNN-V</td>
<td>5.86</td>
<td>4.09</td>
<td>37.57</td>
</tr>
<tr>
<td>PK-RNN-V model - PK-RNN-V E with Feedback</td>
<td>5.39</td>
<td>3.64</td>
<td>25.41</td>
</tr>
<tr>
<td>PK-RNN-V model - PK-RNN-V E with Full Feedback</td>
<td>5.37</td>
<td>3.62</td>
<td>25.05</td>
</tr>
</tbody>
</table>

References
Abstract
Currently, there is a lack of standardized methods to assess artificial intelligence models for their implementation in healthcare. To address this issue, an evaluation framework, termed "Translational Evaluation of Healthcare AI (TEHAI)" was developed with capability, utility and adoption as the three main components and fifteen other associated subcomponents. The authors envisage this framework can be used by regulatory agencies, vendors, AI developers and researchers to assess the translational aspects of AI in healthcare models.

Introduction
The application of artificial intelligence (AI) algorithms to big data has opened new opportunities to monitor and respond to the COVID-19 epidemic. However, many AI systems have been developed on potentially biased data sets. Also, variability in quality and structure of many healthcare datasets makes it arduous for them to be combined and be used as big data required for developing advanced machine learning systems. Further, technology-supported clinical decisions require a robust ethical framing which is lacking for many available AI systems. Utilizing and integrating AI systems in clinical settings can be potentially expensive and disruptive, thus necessitating strong justification for their deployment. Also, there is a lack of standardized information on what algorithms are available and how well they are suited to specific situations and datasets. As a result, adopters of AI cannot benefit from previous experience and face unnecessary roadblocks on the path to an effective healthcare response.

Methods
To address this issue, the Australian Alliance for AI in Healthcare (AAIH) have developed an evaluation framework, termed "Translational Evaluation of Healthcare AI (TEHAI)". TEHAI has been developed to assess the capability, utility and adoption components of a curated list of COVID-19 AI models. These three components were chosen as we wanted to focus on the translational of AI in healthcare i.e., how and how often does AI get applied in healthcare? The emphasis on translational and ethical features of the model development and deployment distinguishes TEHAI from other evaluation instruments. The evaluation framework (along with its components and subcomponents) is outlined at a high level in figure 1a. In the subsequent sections, we will detail each component and sub-component of TEHAI. All together there are three main components (capability, utility and adoption) and fifteen sub-components in TEHAI (Figure 1b).

Figure 1a. Overview of TEHAI
Figure 1b. Reviewers evaluation of TEHAI subcomponents

Capability: assesses the intrinsic technical capability of the AI system to perform its expected purpose by reviewing key aspects as to how the AI system was developed. Utility: evaluates the usability of the AI system across different dimensions including the contextual relevance and safety and ethical considerations regarding eventual deployment into clinical practice. It also assesses the efficiency of the system. Adoption: assesses the translational value of the AI system by evaluating key elements that demonstrate the adoption of the model in real life settings.

AAIH initially conducted a pilot application of the TEHAI framework on a list of 25 randomly chosen entries from the COVIDENCE review system. Each reviewer read through the papers independently and assigned points according to a guideline document and arrived at the findings as summarized in Figure 1b. (1)

Conclusion
The authors envisage evaluation of AI in healthcare becoming de rigueur in the near future, necessitating an appropriate evaluation framework like TEHAI. The framework can be used by regulatory agencies, vendors, AI developers and researchers to assess the translation aspects of AI in healthcare models.

References
A Technical Performance Study and Proposed Systematic and Comprehensive Evaluation of an ML-based CDS Solution for Pediatric Asthma

Shauna M. Overgaard, Ph.D. 1, Kevin J. Peterson, Ph.D., M.S. 1 Chung II. Wi, M.D. 2,3, Bhavani Singh Agnikula Kshatriya 1,4, Joshua W. Ohde, Ph.D. 1, Tracey Brereton, M.S. 1, Lu Zheng, Ph.D. 1, Lauren Rost, Ph.D. 1, Janet Zink, Ph.D. 1, Amin Nikakhtar, Ph.D., M.S. 1, Tara Pereira 1, Sungwan Sohn, Ph.D. 4, Lynnea Myers, Ph.D., R.N., Ph.N. 3, Young J. Juhn, M.D. 2,3

1 Center for Digital Health, Mayo Clinic, Rochester, Minnesota; 2 Precision Population Science Lab, Mayo Clinic, Rochester, Minnesota; 3 Department of Pediatric and Adolescent Medicine, Mayo Clinic, Rochester, Minnesota; 4 Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester

Abstract
Achieving optimal care for pediatric asthma patients depends on giving clinicians efficient access to pertinent patient information. Unfortunately, adherence to guidelines or best practices has shown to be challenging, as relevant information is often scattered throughout the patient record in both structured data and unstructured clinical notes. Furthermore, in the absence of supporting tools, the onus of consolidating this information generally falls upon the clinician. In this study, we propose a machine learning-based clinical decision support (CDS) system focused on pediatric asthma care to alleviate some of this burden. This framework aims to incorporate a machine learning model capable of predicting asthma exacerbation risk into the clinical workflow, emphasizing contextual data, supporting information, and model transparency and explainability. We show that this asthma exacerbation model is capable of predicting exacerbation with an 0.8 AUC-ROC. This model, paired with a comprehensive informatics-based process centered on clinical usability, emphasizes our focus on meeting the needs of the clinical practice with machine learning technology.

Introduction
Clinical care of chronic diseases requires a personalized treatment approach, and the heterogeneity of variables contributes to the challenges of managing chronic diseases such as asthma, the most common chronic condition in children and adolescents in the United States (1-3). A major impediment to optimal asthma care is the lack of efficient and effective clinical decision support (CDS) systems that enable high-value asthma care and quality outcomes while reducing clinician burden (4-6) and costs (7-9). While risk stratification is challenging at the point of care because it relies on the synthesis of large heterogeneous information, there is strong potential for the use of machine learning, natural language processing, and deep learning algorithms to identify and stratify patients by symptom severity and socioeconomic status (10, 11). In pediatric asthma management, efficient and effective review of electronic health records (EHR) and timely clinical decisions are impeded by the varied location, contents and formatting of features characterizing pediatric asthma, given that asthma may serve as an endpoint to a variety of underlying pathologies (12). Furthermore, relevant patient data for asthma management is often stored as unstructured free-text narratives or coded as non-uniform (standardized) inputs into multiple EHR workflows (13). If providers are to better manage uncontrolled asthma and to optimize asthma care, efficient and effective use of EHRs is critical.

Our aim is to develop an AI solution to support clinicians through 1) predicting future risk of asthma exacerbation for their patients (risk stratification and resource management), 2) providing this risk evaluation in the context of a summary of relevant information for asthma management (reduction of EHR review burden), and 3) offering actionable intervention options. We also aim to evaluate the developed algorithm in the context of a comprehensive AI implementation and evaluation framework developed by Park et al. 2020 (14), and further bolster this framework by applying our internally developed model documentation framework.

Methods
Through the support and collaboration of the Mayo Clinic Center for Digital Health, Precision Population Science
Lab and the Department of Pediatric and Adolescent Medicine Artificial Intelligence (DPAM-AI) program, a multidisciplinary team of practicing clinicians, data scientists, and translational informaticians was formed to develop, evaluate, and execute a machine learning-based CDS solution to offer augmented artificial intelligence that optimizes asthma care.

User Needs and Initial Workflow Assessment
The tool will be used by physicians, nurse practitioners, nurses, and care coordinators/asthma managers from the Division of Community Pediatric and Adolescent Medicine and Department of Family Medicine for outpatient asthma management in children. Providers aim to efficiently identify pediatric patients with poorly controlled asthma or with moderate to severe asthma and improve their asthma care by applying precision asthma clinical care in both the personal and social context. To represent the primary user group of the CDS tool, clinical stakeholders are central to all developmental processes and engaged on a weekly basis to participate in shaping and prioritizing translational informatics deliverables that are people-centric, provide value, and are evidence based. Currently, the clinical workflow involves the clinician reviewing the EHR prior to their upcoming appointments to better understand the clinical background and current health status for each patient. For patients with asthma, this requires searching multiple locations within the EHR to determine the patient's asthma status and what potential interventions, if any, need to be considered. The process of locating all asthma related data is challenging and time consuming, reducing time with patients and their families. The translational informatics team reviewed feasibility study data previously conducted by the DPAM-AI group (15-17). Follow-up interviews were then conducted to create an initial list of user requirements for workflow integration, visual presentation, and explainability of the model output. The user requirements specification was evaluated against the NASSS (Non-Adoption, Abandonment, Scale-up, Spread, and Sustainability) framework (18) to assess requested features for sustainability, and scalability, a method employed to increase the likelihood of CDS tool value and broad adoption. The consensus was to create an AI-assisted CDS tool imbedded within the current workflow to help clinicians review the EHR efficiently and effectively, prioritizing resources for high-risk patients. Importantly, the clinical stakeholders required that the custom AI/ML algorithms identify patients at high risk for asthma exacerbation and further curate a list of monitorable asthma-related measures for comprehensive and efficient tracking and review by providers.

System Description
The purpose of the comprehensive CDS tool is to offer augmented artificial intelligence by reducing the amount of time a provider spends locating any single patient's features in the EHR that may contribute to asthma evaluation and intervention. By requirements of clinical stakeholders, the ML/AI algorithm is one feature in the CDS tool and will present a risk score to indicate the likelihood that a patient will experience an asthma exacerbation within the next twelve months. This comprehensive information would provide insight for the autonomous clinician to use in preparing a diagnosis and care intervention. Therefore, the Mayo Clinic’s Asthma-Guidance and Prediction System (A-GPS) is being developed as an AI-assisted CDS tool that extracts patient-level data to a) predict the likelihood that the patient will experience an asthma exacerbation (AE), b) explain the predicted risk by detailing the relevant factors contributing to the predicted likelihood of AE, c) contextualize the algorithm by placing the outcome within a visualization of relevant clinical information pertaining to a patient’s asthma status, and d) operationalize the algorithm by offering actionable, timely interventions (precision asthma care).

We anticipate that the improved workflow will involve reviewing the A-GPS tool by clinicians immediately before each patient appointment. Prior to entering the room, clinicians often review the EHR of the patient in the upcoming appointment. This process currently requires significant time to locate and review the patient’s background, recent visits, medications, and current diagnosis. To improve this process, A-GPS will be integrated into the current workflow allowing clinicians to access asthma-related information in one location, enabling an efficient review of the patient’s clinical information prior to an appointment. The ability to use the A-GPS tool without changing the clinician’s current workflow is a high priority among users and stakeholders.

Technical Performance Evaluation
Technical performance evaluation will be assessed and validated through the data card system developed by Mayo Clinic. We also conducted an evaluation of the predictive model component of A-GPS. Development of this machine learning model focused on the prediction of AE risk for a patient, where AE risk is defined as the probability of having at least one exacerbation within the next 12 months. Patient is labelled as having exacerbation based on three specific criteria: a) An inpatient/hospitalization visit for asthma diagnosis, b) An ED visit for asthma diagnosis, c) An outpatient visit of a patient with asthma diagnosis along with usage of oral corticosteroids medications. If a patient meets any one of the three criteria, then the patient is labelled as having asthma.
exacerbation. A rule-based NLP algorithm was developed and validated to identify exacerbation patients with any of the above criteria as described in Model Features section.

**Cohort Selection**
Mayo Clinic Rochester Campus patients between the ages of 6 and 17 with active asthma were selected for this study, where active asthma is defined as at least one previous Mayo Clinic clinical visit with an associated asthma diagnosis in the past twelve months.

**Model Features**
Currently, there are 28 variables which include both structured and unstructured data that are used to train the models. All of these features are identified as clinically significant factors by the domain experts which assist in determining the probability of having an asthma exacerbation. Both unstructured and structured variables have Yes, No, and/or Unknown labels. The structured data are retrieved from the Mayo database using customized SQL queries, but the unstructured features are extracted using rule-based natural language processing algorithms developed under MedTagger IEarchitecture (19) (developed specifically to address corresponding asthma condition). For example, NLP was used to identify pre-determined the asthma criteria of the patient, which is one of the 28 features in the A-GPS study. Each NLP algorithm was developed using hand-crafted rules with clinical concepts suggested by domain experts and was validated against theManual Chart reviewed dataset as a gold standard.

**Data Quality Evaluation**
Many machine learning metadata systems are focused on experimental aspects (hyperparameters, F1 score for a particular run, etc.) (20, 21). While these are important considerations, healthcare data can be complex/challenging, which in turn determines performance of AI systems and provides insights into explainability, transparency, and accountability of AI systems. The Mayo Clinic Data Card is a component of the organization’s comprehensive model documentation tool, with a goal to incorporate a system that can be shared, re-used, and clinically validated. The data card includes relevant project model information, is organized by feature, and contains both structured and unstructured data. It also consists of a definition and category for each feature, data provenance such as sample query and data source, distribution of data availability, distribution of comorbidities aligned with clinical criteria, clinical description, and interpretation. A data card for each of the 28 model features was processed for the interdisciplinary team’s evaluation and feedback.

**Model Evaluation**
Model evaluation was conducted using 5-fold cross-validation. Evaluation metrics selected for the binary dependent variable included the area under the receiver operating characteristic curve (AUC-ROC) and F1 score. For F1 evaluation of the binary output, a 0.5 positive/negative threshold was initially chosen. The final threshold will be determined via analysis of the clinical application and through feedback from care providers.

**Results**

**Model Selection**
The following candidate models were developed from several different architectures: logistic regression, support vector machine, random forest, Gaussian Naive Bayes, and multilayer perceptron. Each model was implemented using the scikit-learn Python machine learning framework (22).

Overall, logistic regression model outperformed the other candidates producing a 0.8 AUC-ROC. F1-score is shown in Table 1 (Figure 1).
Table 1. Logistic Regression model output

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Mean F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.421527</td>
</tr>
<tr>
<td>0.2</td>
<td>0.522789</td>
</tr>
<tr>
<td>0.3</td>
<td>0.531575</td>
</tr>
<tr>
<td>0.4</td>
<td>0.484955</td>
</tr>
<tr>
<td>0.5</td>
<td>0.456748</td>
</tr>
<tr>
<td>0.6</td>
<td>0.430991</td>
</tr>
<tr>
<td>0.7</td>
<td>0.4161</td>
</tr>
<tr>
<td>0.8</td>
<td>0.388956</td>
</tr>
<tr>
<td>0.9</td>
<td>0.339465</td>
</tr>
</tbody>
</table>

Figure 1. Comparison of AUC score of various models

Final parameters chosen for the logistic regression model were: (C=0.23357214690901212, max_iter=10000, penalty='l1', solver='liblinear'). All other parameters of logistic regression were applied with default values defined in Scikit-learn package (22).

Discussion

We have described our proposed machine learning-based CDS system to alleviate the clinician burden of consolidating relevant information for optimal care of pediatric asthma patients and presented the methods and results from our study of algorithm performance to predict future risk of asthma exacerbation. To deploy our tool safely and effectively in the clinical setting, we further propose a systematic and comprehensive evaluation plan.

AI Evaluation Plan

Our plan is to conduct an AI evaluation using the framework described by Park et al. 2020, which includes five
phases (Figure 2) to address studies of technical performance, usability and workflow, and health impact (14). We iteratively follow our model documentation steps throughout the AI evaluation to ensure dissemination based on our experience and a literature study, which will accompany the tool throughout its life cycle and into maintenance.

**Model Documentation**

We followed our standardized model documentation framework developed internally, which includes five steps: (1) Prepare, (2) Develop, (3) Validate, (4) Deploy, (5) Maintain. Components from the TRIPOD, ML Test Score, CONSORT-AI, Model Card reporting guidelines, and internal best practices prompt model documentation items (23-26). Our model documentation framework serves as a reference to inform future phases and stakeholders, and accompanies the AI/ML tool for knowledge continuity through evolution and maintenance. By promoting transparency and standardizing documentation requirements, we position our ML-based CDS tool for translation, address and mitigate bias and risk proactively, and communicate use cases and deployment strategies across all stakeholder groups.

![Model Documentation Framework](image)

**Figure 2.** Phased research framework for evaluation of AI applied to the A-GPS project based on Park et al. (2020), with the internally developed model documentation framework (14).

**Phase 0: Discovery and Intervention**

Clinical stakeholders are central to the CDS tool and AI/ML developmental processes and are currently engaged on a weekly basis to participate in shaping and prioritizing translational informatics and data science deliverables. While we have completed Phase 0, user needs and workflow assessment will continue to inform model development and translation. In addition to traditional aspects of model development, collaborative translational informatics and data science processes include design thinking, process improvement, and implementation science to ensure the AI/ML-augmented CDS tool is people-centric, provides value, and is evidence based (27).
**User Needs and Initial Workflow Assessment**

As a tool to support asthma management in outpatient practice, the CDS tool users are composed of a variety of clinician roles, such as pediatricians, primary care providers, pulmonologists, asthma care coordinators and nurses. Preparation and evaluation of user requirements specifications through workflow assessment and clinician interviews indicated a consensus to create an AI-assisted CDS tool (i.e., A-GPS) to help clinicians review the EHR efficiently and effectively, prioritizing resources for high-risk patients. Importantly, the clinical stakeholders required that the custom AI/ML algorithm identify patients at elevated risk for asthma exacerbation and further curate a list of monitorable asthma-related measures for comprehensive and efficient tracking and review by providers in a single dashboard view. Moreover, the current clinical workflow will not change with the integration of A-GPS. Clinicians require access to the A-GPS tool at the same time they would typically review the EHR.

**Identification of Target Outcomes and Key Performance Indicators**

To prepare for clinical evaluation, model outcomes were included in a broader prospective evaluation to measure clinician satisfaction and the feasibility of the AI-CDS tool implementation. The plan will assess the capacity to execute the workflow triggered by the model and estimate the potential impact of healthcare delivery factors and work capacity constraints. In a future version of the tool, we plan to integrate a remote patient monitoring spirometry device and embed asthma patients’ home-based data in the CDS tool for proactive asthma management. As primary users of the spirometry device, the satisfaction and requirements of pediatric patients and their caregivers will be fully evaluated through in-depth interviews.

**Model Documentation: Prepare**

We explained the medical context (including whether diagnostic or prognostic) and rationale for developing the AE prediction model, specifying the objectives and key elements of the model through the use of TRIPOD (23). TRIPOD was also used to establish eligibility criteria for participants, data sources, extraction dates, and description and distribution of cohorts. The intended use of the AI intervention in the context of the clinical pathway and intended users was explained through CONSORT-AI (25). We identified out-of-scope use cases using model cards (26).

**Phase 1: Technical Performance and Safety**

Continued algorithm modeling performance optimization will be scheduled before medical use setting evaluation. UX design specialized translational informaticians will engage with practice components to iteratively construct prototype and evaluate usability and workflow to determine effectiveness, efficiency, satisfaction, ease of use, explainability, and utilization.

**IRB and FDA Submission**

Organization-specific Institutional Review Board (IRB) processes will be followed in accordance with standard protocols. Mayo Clinic’s legal department will be engaged to determine whether the algorithm falls under the purview of Food and Drug Administration (FDA) oversight and fits within regulatory definitions.

**AI/ML Model Explainability**

Literature surrounding the current state of AI/ML-based CDS documentation recognizes inconsistencies in standards and governance surrounding explainability. The ineffective balance of CDS tool intelligence with explainability, creates gaps in translation, implementation, and accountability (28). The lack of sufficient documentation further proves challenging for stakeholder adoption calling for the development of enhanced strategies to effectively translate AI-CDS tool integration into clinical settings (29).

The translational informatics and data science teams will engage clinical stakeholders in creating materials to ensure appropriate interpretation, contextual information, and educational support for use of the model. When it is known that the ML-based CDS tool functions technically, its fit into the clinical workflow must be evaluated and education and documentation must be provided to explain the algorithm and its limitations to effectively translate between the perspectives of experts that create and support the technology and the perspectives of experts that employ the solution to patients. Evaluating the interpretation needs of the clinicians, preferences for display of model output (e.g., percentage vs. binary threshold), and feature contributions will be assessed. Concurrently, the teams will engage clinician stakeholders in the development of model documentation to support explainability(30). Strategic efforts to promote explainability include the application of a ML-based CDS tool documentation framework grounded in scientific research addressing known challenges by encompassing interdisciplinary best practice reporting.
requirements that follow phases of model development (Prepare, Develop, Validate, Deploy, Maintain) for knowledge continuity throughout the solution's life cycle.

**UX Research and Prototype Design**
Continued UX design research includes clinician shadowing and interviews to understand user requirements and facilitate optimal workflow integration, estimate potential impact of healthcare delivery factors, and work capacity constraints on achieved benefit. UX design research will be conducted through collection of qualitative and quantitative data to identify all user groups, understand user needs, probe for optimal tool design to support clinical decision making and routine workflow for each group. The data will be collected from a minimum of 12 clinicians through interviews. Based on research goals, four formats of UX data collection methods will be used: (1) in-person shadow, (2) virtual stakeholder interview, (3) virtual user interview, (4) screen capture of EHR walkthrough. These four methods provide comprehensive contextual data from different facets. The shadowing will allow for in-situ investigation of facility workflow, patient flow and physician-patient interaction for decision-making during visits. The interviews and EHR walkthrough will focus on problem-probing and user needs consolidation.

**Feasibility Analysis**
The interdisciplinary team will plan and document how the model will be deployed into the environment. This will include review of architecture requirements, a diagram of architecture supporting the model, technical specifications of model features, clear determination of variable extraction, mapping, and methods of incorporation into model. In addition to the technical review of the model, the team will review potential bottlenecks and solutions, identify success metrics, establish expected output of discovery and usability tests, evaluate feasibility of the algorithm’s implementation, and identification of elements required for capture once algorithm has been deployed in end-user practice domain.

**Data Quality Evaluation: AI Metadata Management**
While the Mayo Clinic Data Card framework is an initial step towards transparent and comprehensive, consolidated model documentation management, the longer-term vision includes a fully vetted model-agnostic tool for managing all Mayo AI-related artifacts. As a bottom-up system model, it will house reusable components, or cards, at the feature, model, and project level. These components all have a one-to-many relationship with one another and are designed to be accessed via a user-friendly interface with version control, and to be searchable, discoverable, executable, and auditable. Data governance and provenance are also defining characteristics of the tool, and it is rooted in the FAIR initiative principles of digital asset findability, accessibility, interoperability, and reusability (31).

**Model Documentation: Develop**
We defined the outcome that is predicted by the model and described how predictors were used in developing and validating the model (TRIPOD). We specified model type, model-building procedures, methods for internal validation, and all measures used to assess model performance. Performance measures for the model were reported as well. Relevant and evaluation factors, model performance measures, decision thresholds, variation approaches, preprocessing, and training data were all identified. We provided a lay summary of model development based on Mayo Clinic Translational Informatics Best Practices. Intersectional results investigating algorithm discriminatory behavior against subgroups and intersectional subgroups were detailed. Participant characteristics, including demographics, clinical features, and available predictors were described. Detailed testing was also conducted for inclusion and reproducibility considerations. In addition, we described the results of risk assessment tests, detailed a risk management plan, and outlined user requirements based on Mayo Clinic Translational Informatics Best Practices.

**Phase 2: Efficacy and Side Effects**
The CDS tool interface will be designed iteratively with clinical stakeholders. The algorithm will undergo a controlled performance and efficacy evaluation by intended users in the medical setting. This will first incorporate model evaluation on EHR data trained in silent mode. Following validation and approvals, a pilot study will be administered on a subset of pediatric providers to assess usability and workflow integration of the CDS tool.

**Interface Design**
UX design specialized translational informaticians will employ agile methodology to iteratively design the AI/ML-augmented CDS tool with the clinician focus group. The team will build the interface and test workflows to support
the seamless integration of the algorithm in end-user practice. Relevant systems and back-end workflow will be mapped to identify key systems impacted by CDS tool integration. The user workflow and risks will be designed and validated to compare current and future state, given CDS tool impact. A change management strategy will be prepared to assist in clinicians’ transition in order to minimize workflow impact and maximize value of the CDS tool.

**Pre-Pilot Model and Business Plan Validation**
Following initial statistical assessment of model performance, the model will be configured to train on EHR data extracts in silent mode. Model performance after training will be validated. Following model completion and the verification of infrastructure needs, the team will construct the business integration plan, review the implementation timeline, training, and communication materials, and seek approvals from the department and enterprise governance committees. Results of model performance, validation, and proposed implementation timeline will be reviewed by the interdisciplinary team, practice proponents, and EHR analysts for consensus approval to initiate pilot in a clinical setting. Adhering to internal protocols, the team will submit a change request for implementation into the EHR.

**Pre-Post Implementation Evaluation**
Study items for clinician satisfaction and time saved with current asthma management tools/interface will be drafted based on findings from analysis of interview data and information extraction time. To evaluate the impact of the new A-GPS patient summary tool implementation systematically, a baseline evaluation will be captured prior to any changes made to interface or workflow, and repeated following implementation of the A-GPS summary tool. Outcomes will be measured and organized using Proctor’s Implementation Outcomes. Implementation outcome metrics include use frequency and prevalence of the tool. Service outcome metrics can be clicks, screen transitions, and time of information retrieval process. Client outcome metric would be satisfaction level of users. Success in meeting user requirements will be evaluated by tool users themselves and recorded following implementation.

**Pilot Model and Business Plan Validation**
A pilot study evaluating the effectiveness of the algorithm, its ability to correctly process input data to generate accurate, reliable, and precise output data, and the association of the algorithm’s output and targeted clinical indicators will be conducted. Meanwhile, validation utilities will provide insight into model performance for the organization’s patient population. Important model performance metrics like true-positive, true-negative, false-positive, false-negative rate, and AUC-ROC will be calculated against testing data. The ML-based CDS tool will undergo an internal evaluation for potential commercialization.

**Model Documentation: Validate**
We will present the full prediction model to clinicians in order to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline values), demonstrate how to use the prediction model, specify the output of the AI intervention, specify whether there was human-AI interaction in the handling of the input data and what level of expertise was required of users, and explain how the AI intervention’s output contributed to decision-making or other elements of clinical practice (23, 25).

**Phase 3: Therapeutic Accuracy**
A clinical validation study will evaluate the algorithm in its intended purpose to produce clinically meaningful, precise, accurate, and reliable output for the target clinical indicators. A randomized controlled trial will be executed and will identify adverse events, effectiveness, efficiency, satisfaction, ease of use, learnability, and utilization.

**Prospective Study/Randomized Clinical Trial**
A randomized controlled trial will demonstrate the clinical validity, efficacy, and safety of the algorithm. A parallel-group, non-blinded, dual-site, 2-arm pragmatic randomized controlled trial will be planned for a period of 9 months (2-month enrollment period followed by a 6-month trial) to assess the feasibility of implementation of the CDS tool within primary care practices 30 dyads of primary care clinician-patient with asthma will be recruited from Mayo Clinic Rochester and Mayo Clinic Health System primary care sites. Clinicians under intervention group after randomization will access the CDS tool that can support asthma management whereas those under control group will follow usual care for asthma. Primary end point is the change of clinicians’ and caregivers’ satisfaction score with asthma care during the trial. Also, we will assess if the CDS tool may improve efficiency and quality of asthma care by measuring time to review patient’s electronic health records as well as direct interaction time with patients at the point of care. The CDS tool can automatically extract and synthesize large amounts of pertinent patient data related to asthma management from electronic health records to support clinicians and care teams to optimize asthma care
through the following: 1) provision of a one-page summary of the most relevant clinical information for asthma management for the past 3 years, 2) prediction for future asthma exacerbation risk within a year from a machine learning algorithm, and 3) offering actionable, timely and guided interventions (precision asthma care). After the trial, we will review model performance, evaluate usability data, consider, and document adjustments and needs.

**Model Documentation: Deploy**

Even in the era of electronic health records, there is lack of effective and efficient CDS tool for streamlining asthma management (7-9). While many CDS tools using machine learning algorithms have been developed and even deployed within electronic health records, few CDS tools have been tested via randomized controlled trial (32). We plan on conducting a clinical trial in which the results may include, but not limited to, safety, clinical outcomes, care quality, care cost, and satisfaction of end users (e.g., clinicians, patients, and caregivers). Once clinical validation is demonstrated and the tool is deployed to health care institutions, the model performance overall as well as in subgroups needs to be monitored on a regular basis to prevent potential bias.

We aim to give an overall interpretation of model results, considering objectives, limitations, and results from similar studies and other relevant evidence, as well as discuss the potential clinical use of the model and implications for future research (23). Methods of integration for the AI intervention into clinical settings, including any onsite or offsite requirements and training, will be described (25). Model limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses, will be detailed (25). We will detail a prospective study for technical validation of the model, aiming to communicate whether the model correctly processes input data to generate accurate, reliable, and precise output data and validates model performance for clinical application (Mayo Clinic Translational Informatics Best Practices).

**Phase 4: Safety and Effectiveness**

Post-deployment surveillance and post-production management will include an evaluation of adoption and sustainability of the AI algorithm utilization.

**Post Deployment & Maintenance Risk**

The performance of the algorithm and relevant processes will be tracked and compliance with relevant regulatory standards and requirements ensured. The algorithm performance will be monitored for quality assurance testing, patient safety, and incorrect results. Conditions will be re-evaluated to determine whether the algorithm is exposed to external factors that may affect its accuracy including but not limited to evolving societal norms and values, covariate shift, downtime risk, and degradation of model. The need for algorithm retraining, algorithm improvement, and algorithm retirement will be assessed at a minimum of every six months.

**Quality Monitoring and Audit**

Tracking mechanisms for optimal and high-quality performance will be established, including frequent status updates to leadership, meta-data monitoring, algorithm failures, and distributional shift. The internal audit team will be engaged for quality monitoring and regulatory compliance.

**Maintenance**

The interdisciplinary team will troubleshoot algorithm related issues or failures, the addition of new features, and versioning, updating the model documentation as needed. The internal operations team will review technical issues related to algorithm performance or user feedback.

**Model Documentation: Maintain**

We aim to discuss any limitations of the model (such as nonrepresentative sample, few events per predictor, missing data) and provide information about the availability of supplementary resources associated with the model. We will ensure and document considerations taken in regard to data, human life, bias mitigation, risks and harms, and appropriate use cases. Contact information communicating where to send questions, comments, or additional
concerns about the model will be provided. We aim to describe a quality assurance plan, address the potential of external and internal drift, develop incident monitoring logs, and provide contact information for accountable teams.

Conclusion

Our aim was to describe an ML-based CDS tool for pediatric asthma management developed by our multidisciplinary team of clinicians, data scientists, translational informaticians, and UX experts at Mayo Clinic, present the results of our algorithm study, and propose our comprehensive evaluation framework. This tool has the potential to reduce the clinician’s EHR review burden by gathering features to optimize precise pediatric asthma management and to predict the likelihood of asthma exacerbation with an 0.8 AUC-ROC. The goal is to provide this risk evaluation in the context of a summary of relevant information for asthma management and offer actionable intervention options. The construction and application of the phased research approach requires strategic partnership across our multidisciplinary team and stakeholders, and a commitment to rigorous evaluation. As we progress this work, we will continue to leverage a published comprehensive evaluation framework on which we will infer our problem-specific studies. This evaluation plan will be further enhanced through application of our internally developed model documentation framework to ensure rigorous evaluation, transparency, and knowledge continuity at each step: 1) Prepare, 2) Develop, 3) Validate, 4) Deploy, 5) Maintain. These efforts are intended to support the potential for future adoption of this tool on a national and international level.

References

Predictive Modeling of Periodontal Disease Risks Using Electronic Dental Records and Explainable Machine Learning

Jay Patel BDS, PhD\(^1\), Chang Su, PhD\(^1\), Marisol Tellez PhD\(^1\), Jasim Albandar PhD\(^1\), Huanmei Wu PhD\(^1\)

\(^1\)Health Informatics, Department of Health Services Administration and Policy, College of Public Health, Temple University, Philadelphia, PA 19112

Introduction

Advances in periodontal disease (PD) research and treatments showed that PD could be prevented. However, 42% of the U.S. population has PD, leading to tooth loss, poor quality of life, and increased healthcare costs\(^1\). Limited studies show the effectiveness of periodontal treatments in preventing disease progression and tooth loss based on patient characteristics. One primary barrier is the difficulty of conducting randomized controlled trials with adequate numbers of patients over a long time because of several reasons, such as ethical reasons, expenses, and difficulty in enrollment and patient retention. PD preventive care needs to know the risk factors responsible for PD progression could be controlled by assessing patients' disease risk. Some preliminary prediction models have been developed to determine patient-specific PD disease risks. However, studies have shown that these tools do not represent the current patient population and fail to provide patient-specific disease risk and treatment recommendations.

There are several reasons behind these models' sub-optimal performances. First, the models are one size fit for all for treatment recommendations. Thus, the risk score provided by these models is not patient-specific. Second, it does not provide new information to the clinicians that they do not already know, as the models are developed using evidence generated two decades ago and simple rule-based systems from experts' opinions. The increased availability of longitudinal electronic patient data through the electronic dental record (EDR) offers an opportunity to characterize the patient demographics, disease profiles, PD outcomes, medical history, and other health information\(^2\). We can build advanced predictive models with up-to-date multimodal information. Moreover, advanced machine learning (ML) and artificial intelligence (AI) provide an opportunity to develop sophisticated data-driven models for periodontal disease. This study aims to create a data-driven predictive model for PD risks using advanced explainable ML methods through a retrospective study using EDR data.

Methods

Study cohort: We have performed the preliminary study on a cohort of 27,138 dental patients who received at least one comprehensive oral examination (COE) at the Temple University School of Dentistry between 01/01/2017 to 08/31/2021. The raw data retrieved from the EDR consists of >100 attributes, including the patient’s demographics, medical history, periodontal findings, teeth conditions, PD diagnoses, and treatment information from the EDR.

Data preprocessing and missing value imputation: Before building the predictive model, the exclusion criteria are applied to filter patients with missing PD diagnoses. Patients who have >65% candidate predictors missing are excluded from the study also. Second, we performed a data quality check and missing value imputation. For example, one candidate predictor, ethnicity, was dismissed for analysis because it has a very high missing rate across all patients (93%). According to the data types, the remaining missing values were imputed using two different strategies for phase I of our study. For continuous attributes, such as teeth number, missing values were imputed using the median values in the training set. For categorical attributes, such as medical histories, missing values were imputed using the most frequent values in the training set. We let the model select variables of interests as per their feature importance.

During phase II of the study, we will perform missing value imputation based on patient similarity. Various types of similarity metrics can be applied to describe the similarity between patients, including Neighborhood-Based Algorithms, Distance-Based Similarity Metrics, and Correlation-Based Similarity Metrics to retrieve the k-nearest patients. Missing data will be imputed based on the k-patients.

Model training & evaluations: We first grouped PD diagnoses into three groups: healthy control, mild PD, and severe PD. The predictive model was built on the XGBoost\(^3\) framework, a powerful ensemble ML model. We used the one-vs-rest strategy in Phase I for this multi-label classification and prediction, using 80% for the training and 20% for the validation with a 5-fold cross-validation strategy to identify the optimal hyperparameters of the model. After the hyperparameters had been determined, we trained the final model using the entire training set. Prediction performance was measured by receiver operating characteristics curve (ROC), area under ROC curve (AUC), and confusion matrix of prediction accuracy (Fig. 1).
Model interpretability: To enhance model interpretability, we engaged the SHapley Additive exPlanation (SHAP) values to assess contributions of the predictors in distinguishing each class from the others (Fig. 2).

For the initial data sets of 27,138 unique dental patients, there are more females (57%) than males (42%). Most are in the 58-67 years age group (19%), followed by 48-57 years (18%) and 28-37 years. The race information of 57% of patients is missing. Among the remaining 43% with reported races, black was the most dominant race (28%), followed by the White race (12%). The performance showed that our initial predictive model could differentiate healthy patients vs mild PD cases and mild PD vs severe PD cases. We have achieved an average AUC of 0.72 (Fig. 1). Identifying patients with minimal gingival inflammation and bone loss has the most accurate results, followed by mild and severe PD cases.

Our model has moderate performance because we imported a total of 74 features. No studies have used 74 features towards predictions. During this process, we also identified new associations such as renal diseases, mental conditions, and cancers as driving factors for the PD risk. We will expand this information in the next phase of this study.

Conclusions: Our pilot study demonstrated promising results on utilizing EDR, ML, and AI models to predict the risk of PD. This model would provide new information to the clinicians about the PD risks and the factors responsible for the disease progression to take preventive approaches. We will utilize more sophisticated imputation and phenotyping methods to improve the quality of the EHR data for superior risk predictions.

A Deep Understanding approach for Case Identification and AutoCoding System of Cancer Pathology Reports for a Population Cancer Registry.
Jon Patrick, PhD, MSc, BSc (Psych), Grad Dip Behav Psych, Dip L.Surv.
Health Language Analytics Global LLC, Sydney, NSW, Australia

Abstract

This report describes the process of developing the HORIZON system for Case Identification and AutoCoding of pathology reports to ICD-O-3 for the California Cancer Registry (CCR) using a Deep Understanding methodology. The pipeline structure of 6 components and the process of creating an industrial quality processing system are described, rather than answering research questions of Clinical NLP (CNLP).

Introduction

The objective of the project was to build a multi-unit processing pipeline using an admixture of supervised machine learning and NLP technologies. Initially it separates reportable cancer pathology reports from non-reportable reports. Subsequently, it encodes the reportable reports for 5 cancer attributes, namely, {Tumour Site, Histology, Behaviour, Grade and Laterality}. Immunohistochemistry and Genomics reports were defined as out-of-scope for the project, so only Histopathology and Haematology reports were to be fully processed.

Results

The architecture of the Deep Understanding pipeline consists of 6 components: Document Classifier, NLP Pre-processor, Clinical Entity Recogniser, Coding Inference Engine, Active Learning Feedback, Quality Control process. The accuracy of cancer attributes of the final system varied from 97.5% to 99.95% while the core coding system was accurate to 97.5% overall. The system represents an improvement over manual coding of 20-80% across the various cancer attributes. The system delivered Reportability classification of: Reportable 48%, Non-Reportable 36%, Manual 15% Unusable 1%. Coding would appear to initially have the potential for about 72% automated, with further improvement attainable. Reports computed to be too complex or incomplete were routed automatically to a manual processing workflow. Accuracies over 3165 reports were Site (4 digits) 95.23%, Histology (4 digits) 96.37%, Grade 98.77%, Behavior 99.53%, Laterality 97.82%.

Figure 1 shows the progressive improvement in accuracy for each of the 5 data items over 12 months of development as the machine learning elements were delivered an increasing amount of manually compiled gold standard classified and annotated documents.

![Figure 1. Coding accuracy of all 5 attributes from RUN 17 to 46.](image)

Conclusions

Specific tests confirmed the generalizability of the computational model by finding and correctly coding content that had not been seen in the training set. The delivered system has been installed on a Virtual Machine in the CCRs data center and operates on an around the clock schedule.
Augmenting the Quality Assurance Process with Human-Centered Machine Learning in Radiation Oncology

Malvika Pillai¹, John Shumway, MD², John Dooley, BA², Karthik Adapa, MPH, MBBS¹,², Ross McGurk, PhD², Shiva Das, PhD², Bhishamjit Chera, MD², Lukasz Mazur, PhD¹,²
¹Carolina Health Informatics Program, University of North Carolina, Chapel Hill, NC ²Department of Radiation Oncology, University of North Carolina, Chapel Hill, NC

Introduction
Radiation therapy (RT) treatment plans are created using a complex and iterative process by a multi-disciplinary team of professionals including physicians, physicists, and dosimetrists. To ensure high quality of the RT treatment plan dosimetrists and physicists manually perform chart checks of multiple (>65) metrics for each patient’s RT plan before it can be made deliverable. This is a time intensive process, involving a high cognitive workload [1, 2]. At our institution, there are number of automated second-check tools used to augment the QA process, thus improving efficiency, and reducing cognitive workload [2]. The research objective of this work is to further augment QA processes of RT treatment plans by applying human-centered ML solutions for QA plan difficulty evaluation and work prioritization.

Methods
We aim to use machine learning to identify and flag difficult cases that require additional scrutiny by the physicist. The analysis demonstrates a classification algorithm that categorizes the radiation treatment plans as difficult or not difficult to QA via the physics pre-treatment chart check. Data were retrieved from 973 physics pre-treatment chart checks for treatment plans from July 2018 to October 2020, encompassing all cancer sites. The outcome variable, the degree of difficulty of treatment plans, was collected from sixteen physicists as a subjective rating on a scale of 1-10, normalized to the physicist performing the chart check. The top 30% were labeled as difficult and the bottom 70% as not difficult. The dataset was split such that 778 cases were used as a training set, and 195 cases were used as the test set. The test set was set aside to evaluate performance after all model training and parameter tuning. The training set was further split into a train set (622 cases) and a development set (156 cases). An iterative data selection process was conducted by one clinician, one physicist, and a software development team to select attributes along the radiation therapy workflow [1]. The attributes selected were based on clinical relevance, contribution to plan complexity, and quality assurance metrics. Prior to developing the ML system, feature extraction methods were evaluated using the train and development sets to improve transparency and interpretability [3].

A voting classifier, consisting of support vector machine (SVM), decision tree, random forest classifier, adaboost, and neural network, was used for binary classification to predict the degree of difficulty of radiation treatment plans. Sensitivity analysis was also conducted on the test set to assess how much each feature contributed to predicting degree of difficulty.

Algorithm Results
The five algorithms were trained using the oversampled train set and development set, and performance was evaluated on the test set. Overall accuracy, positive predictive value (PPV) (i.e., precision) and sensitivity (i.e., recall) were used as metrics for evaluation (Table 1). Sensitivity analysis showed features associated with plan complexity and clinical relevance were sensitive across at least three algorithms.

Conclusion
Through iterative development cycles, key stakeholders were included throughout development of the ML solution, and transparency and interpretability were heavily considered to improve physicists’ adoption. The proposed ML solution can be used to guide the physicist in prioritizing their activities to reduce cognitive workload, potentially improving QA process effectiveness. We will use the Systems Engineering Initiative for Patient Safety (SEIPS) 2.0 model [4] to evaluate the proposed ML solution. Our next effort includes using the proposed evidence-based theoretical model rooted in human-centered systems engineering to support our evaluation plan.

Table 1. Overall accuracy, PPV, and sensitivity for all algorithms and the voting classifier across development and test sets.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Development Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall Accuracy</td>
<td>PPV</td>
</tr>
<tr>
<td>SVM</td>
<td>85.26%</td>
<td>83.00%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>80.13%</td>
<td>77.40%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>83.33%</td>
<td>80.97%</td>
</tr>
<tr>
<td>Adaboost</td>
<td>80.77%</td>
<td>78.80%</td>
</tr>
<tr>
<td>Neural Network</td>
<td>75.64%</td>
<td>72.96%</td>
</tr>
<tr>
<td>Voting</td>
<td>87.18%</td>
<td>81.76%</td>
</tr>
</tbody>
</table>

References
Towards an Optimal Policy of Mass Casualty Trauma Triage

Alexandrea K. Rammarine, BS¹, Nathaniel D. Bastian, PhD, MEng, MS¹, Anne M. Stey, MD, MSc²

¹Northwestern University School of Professional Studies, Chicago, Illinois, USA; ²Northwestern Medicine, Chicago, Illinois, USA

What might the attendee be able to do after being in your session?

Quantitatively address clinical considerations of primary and secondary trauma triage to standardize mass casualty incident and retrieval policies in U.S.-based urban healthcare systems.

Description of the Problem or Gap

Healthcare systems, including hospitals, have been slow to incorporate artificial intelligence (AI) for patient-and care-centric use cases despite stringent regulations against the misuse of data under the Health Insurance Portability and Accountability Act (HIPAA) and Health Information Technology for Economic and Clinical Health Act (HITECH). Emergency rooms (ERs) in the United States (U.S.) generate massive quantities of data annually, captured by a network of hospital trauma centers contributing to a standardized database curated by the American College of Surgeons Committee on Trauma (ACSCOT), the National Trauma Data Bank (NTDB). NTDB data can be leveraged to model and prescribe triage prioritization of emergency centers responsive to simulated trauma case influx over a period of time due to distinct scenarios. The proposed prescriptive research is intended to correct a “one-size-fits-all” clinical triage procedure given traumatic and emergency situations, addressing major concerns of retrieval and optimizing to avoid burdening bed availability of the highest-level trauma centers (ie, bypass). Additionally, there is no “gold-standard” trauma triage protocol adapted throughout the U.S., nor is there clear policy indicating that trauma triage protocols are necessarily distinct given the heterogeneous features and outcomes of emergency situations.

Methods: What did you do to address the problem or gap?

A discrete event simulation (DES) was developed using SimPy Python package to create a synthetic urban healthcare system environment complete with simulated trauma patients from a pre-defined mass casualty event. Patient parameters were simulated based on 2002-2016 data from the National Trauma Data Bank. Patient data were subject to feature engineering for data dimensionality reduction using the Sci-kit learn Python package. A baseline computational agent traverses the DES following a Markov Decision Process (MDP) using stochastic Monte Carlo draws. Pandas python package is used to translate the MDP to Q-learning in order to track action-state values and subsequent optimal policy generation as the agent completes multiple simulated episodes of the DES. Current work will use TensorFlow and Keras Python packages to enable artificial intelligence for agent decision making.

Results: What was the outcome(s) of what you did to address the problem or gap?

This study is ongoing, set to complete in Jan 2022. The final results are pending. The preliminary results indicate that a naive reinforcement learning agent (currently without use of artificial intelligence) is able to triage three trauma patients across three health centers, a non-trauma Hospital, a low-level trauma center, and a high-level trauma center in less than 17 maximum steps to the terminal state. Additionally, the naive agent is capable of performing retrieval on the most severely injured patients without burdening high level trauma centers (ie, bypass).

Discussion of Results

Final discussion is pending.

Conclusion

The results of the simulation optimizations may provide evidence-based scoring and prioritization processes for U.S. trauma centers, leading to policy change surrounding ER utilization given high influxes of situation-specific trauma cases in mass casualty and disaster scenarios. National adaptation of standardized, data-driven trauma responses would lead to a reduction is U.S. healthcare spending. The ultimate goal is to preserve as many lives as possible by capitalizing on data-driven organized treatment rather than time spent on triage given the critical, vital seconds that distinguish trauma case survival rates from others.

Attendee’s Take-away Tool
The algorithm, supporting documents, and saved TensorFlow models will be **publicly** available in Jan 2022 via github.com/ramnadime/TraumaTriage_ProtocolOptimization. Currently the GitHub repository is private, only viewable to the aforementioned authors, until final work is complete in Jan 2022.

**Use of Knowledge Acquired at Previous AMIA Events**

N/A
A multi-aspect technical performance evaluation of deep learning based models for predicting COVID-19 patients outcomes on admission

Laila Rasmy, MSc1, Masayuki Nigo, M.D2, Bijun Sai Kannadath, M.B.B.S, M.S3, Ziqian Xie, Ph.D1, Bingyu Mao, MA1, Angela Ross, PhD3, Hua Xu, PhD3, Degui Zhi, PhD1

1School of Biomedical Informatics, UTHealth, Houston, TX; 2McGovern Medical School, UTHealth, Houston, TX; 3College of Medicine, University of Arizona – Phoenix, AZ

Introduction

While a comprehensive evaluation plan needs to be agreed on during the early phase of prediction task definition, we propose six factors that need to be considered during the evaluation of the implementability of a predictive model. Those factors are prediction performance, transparency, generalizability, data mechanics, efficiency, and data privacy. Those factors should be considered starting from phases 0, 1 as described in the AI evaluation framework1 and continuously monitored and improved as we go through all phases till phase 4.

Methods

Pytorch_EHR is a framework to train and evaluate deep learning (DL)-based models to predict different types of clinical events and it includes utilities to facilitate multi-metric technical performance evaluation. The framework starts with the data preparation and preprocessing flow, then followed by the DL model architecture. The first layer of the model is the embedding layer, which can be randomly initialized or initialized from a pre-trained static or a contextualized embedding, such as Med-BERT2. The feature representations trained/fine-tuned in the embedding layer, along with time information or other continuous data variables, are then fed into the core model architecture, which can be as simple as a linear layer or a more complicated RNN model, to achieve a binary classification or a survival prediction. In addition to the basic predictive model training module, we included three ancillary modules. The first is the explainability module, which currently uses the integrated gradient technique. The second is a multimetric performance evaluation module which includes several informative plots, such as the calibration plot and stratified Kaplan-Meier (KM) curves, based on the predicted survival probabilities. The third is a subgroup analysis module that can be used to calculate the model performance for different subgroups based on their demographics, location, or common comorbidities. We utilized the Pytorch_EHR framework in our most recent work (CovRNN)3 and for reproducibility and further evaluation by researchers, we share our codebase as an open-source repository at https://github.com/ZhiGroup/pytorch_ehr.

Results and Discussion

Using Pytorch_EHR we reported multi-metric performance results for CovRNN for example for mechanical ventilation prediction, we reported the AUROC (92.9%), AUPRC (79.5%), and specificity at 95% sensitivity (63.5%), as well as the sensitivity (83.4%), specificity (85.1%), and F1 score (59.7%) at the recommended threshold of 10% that achieve the best balance between the models’ sensitivity and specificity using the validation set. We also plotted the calibrations curves as well as the stratified KM curves for low, medium, and high-risk groups, using the survival models3. To evaluate the transparency of the proposed CovRNN models, we first report our study design, including the cohort definition and labeling criteria as well as our results following TRIPOD standards, then we used the Pytorch_EHR explainability module to explain the DL model prediction at patient-visit-event. To evaluate the generalizability of CovRNN models, we demonstrated the high discriminative accuracy in four different test sets from different sources, as well as the model consistent results between different patients’ subgroups. For efficiency evaluation, we sketched the data flow and calculated the running time for each step. We found that the explainability module is the most time consuming with duration of 56 seconds, which we are working on mitigating as part of our interface design and usability evaluation stage. For privacy concerns, we never use PHI data as an input features for our models.

Acknowledgments L.R. is supported by the UTHealth Innovation for Cancer Prevention Research Training Program Pre-Doctoral Fellowship (CPRIT Grant No. RP160015).

References

Prospective Evaluation of a 90-day Mortality Prediction Model: From Silent Pilots to Real Time Deployment in the EHR

Lorenzo A. Rossi, PhD; Laura Roberts, MS; Finly Zachariah, MD
City of Hope National Medical Center, Duarte, CA, USA

Introduction

Prognostication in oncology is increasingly difficult due to the rapid evolution of therapies with significant improvement of survival. Accurate prognostication is essential to provide optimal, value-driven end of life care for cancer patients, and can promote goals of care (GOC) conversations with the potential to minimize chemotherapy or ICU utilization in the last weeks of life, and possibly increase hospice admission and length of stay. There are several recent publications on the application of machine learning for prognostication.

We developed a 90-day mortality prediction model trained with data in the Electronic Health Records (EHR). After a non-interventional pilot stage, we deployed the model in February 2021 in the real-time Electronic Health Record Epic infrastructure of our cancer center. Here we present the model and evaluate its overall performance for the first 7.5 months since the go-live and outline our evaluation process for the next stages.

Methods

We trained a gradient boosted tree classifier (via the XGBoost library) with observations from 28,484 patients (less than 10% non-oncologic) and 493 features from demographics, lab test results, flowsheets and diagnoses collected from the EHR in our medical center between 2014 and early 2019. We extracted several hand-crafted features from the time series of labs and flowsheet data in the 180-day temporal window preceding the time of each prediction. We imputed missing values for features only in obvious cases, because tree-based classifiers can handle missing data. Table 1 lists the clinical variables used by the model and the type of features extracted from lab and flowsheet time series. The features associated with the diagnoses consisted of aggregations of Word2Vec embeddings of the ICD-9 codes. A portion of the observations were used for retrospective evaluation based on a temporal split to mimic real world deployment, where past observations are used to train a model to predict in the present. For inclusion in training and evaluation sets, patients needed to have at least two encounters and alive patients needed to have an active encounter in the EHR at least one year after the prediction date. To limit risk of leakage, we picked dates of prediction to exclude encounters within 7 days of death. We also avoided over-representing observations with prediction dates within 30 days of death to avoid training the model with a disproportionate number of near-term deceased patients.

Table 1: Types of clinical data used to extract the features for the machine learning models.

<table>
<thead>
<tr>
<th>category</th>
<th>clinical variable</th>
<th>feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laboratory Tests</td>
<td>Lymph%, Alb (Blood), Calcium (Blood), WBC, LDH (Blood), Hgb (Blood), Platelet Count, ALP, Creatinine (Blood), Bilirubin (Blood), RBC, B12 (Serum), Segmented Neutrophil%</td>
<td>count, median, min, max, SD, slope, intercept, early/late difference to l/h ref., % normal results, % very abnormal results % lab late night orders</td>
</tr>
<tr>
<td>Flowsheet</td>
<td>weight, BMI</td>
<td>count, median, min, max, SD, slope, intercept, early/late difference to l/h ref.</td>
</tr>
<tr>
<td>Demographics</td>
<td>age, gender</td>
<td>value</td>
</tr>
<tr>
<td>Diagnoses</td>
<td>ICD codes</td>
<td>embedding aggregations</td>
</tr>
</tbody>
</table>

After hyperparameter tuning via cross-validation and retrospective evaluation, we retrained a version of the 90-day mortality model for deployment in a silent (invisible to clinicians) prospective pilot, inclusive of all City of Hope patients with accessible EHR data. The resulting model consisted of an ensemble of 413 decision trees with 6 maximum depth. Since October 2019, the model made batches of predictions from observations automatically queried once a day from our electronic data warehouse (EDW). The presence of select new lab results triggered a prediction. We ran
the pilot for the entire 2020, retraining the model several times and then deployed the model in the medical center’s EHR real time infrastructure (RTI) in February 2021. In the RTI, the model receives a stream of Health Level 7 (HL7) messages. Pilot and production included predictions for both inpatients and outpatients.

Results

In the retrospective evaluation \((n = 5,037)\), areas under receiver operating characteristic (AUROC) and precision-recall (AUPRC) curves were 0.82 and 0.20 respectively with 5% prevalence. Figure 1 reports ROC and PRC curves, and performance metrics for pilot (year 2020) and real time production implementations (February 15 deployment - September 30, 2021). The prevalence is the 90-day mortality rate. Predictions are triggered by new recordings of certain lab results. Hence, there may be multiple predictions for a patient. The median age of the patient population at the time of first prediction in production was 52 years (18-101 years range). The performance for pilot and production implementations were very consistent: 0.86 (95% CI: 0.860 – 0.864) vs. 0.85 (95% CI: 0.844 – 0.849) AUROC and 0.41 (95% CI: 0.41 – 0.417) vs. 0.39 (95% CI: 0.39 – 0.402) AUPRC, respectively. The lead days are the time between a correct mortality prediction and the death date of a patient.

Figure 1: ROC and PRC plots; Table with pilot and production performance metrics.

Discussion and Outline for Next Stages of AMIA AI Evaluation Showcase

A common pitfall for deployments of machine learning models in production is a drop in performance with respect to previous evaluations. In our production deployment, overall performance was consistent with the 2020 pilot and the retrospective evaluation at development stage. Pilot and production relied on different data infrastructures: EDW vs. real time streams of HL7 messages.

The production model is being employed in an increasing number of clinical workflows: e.g. to help Clinical Social Work to prioritize patients for advance directive completion, to identify patients who would benefit from GOC discussions, to order Supportive Care consults for ICU transfers of high risk patients, and to identify patients for trial recruitment.

In Stage II, we will present in detail the different clinical decision support applications of the model with related workflows and specific performance results. We will include also results from clinician surveys on utility and usability.

In Stage III, we will report results on our quality measures (e.g. completion rates of advance directives and GOC conversations), as well as on national end of life quality measures (e.g. time in the ICU in the last 30 days of life, chemotherapy use in the last two weeks of life, and hospice utilization metrics).

References

Comparing Multiple Implementable Predictive Models for Pediatric Clinical Deterioration

Sameh N. Saleh, MD, Lingyun Shi, MS, Gerald P. Shaeffer, MS, Yujie Sun, MS, Lusha Cao, PhD, Meghan Galligan, MD, Brianna C. Weeks, MSN, Jennifer Jacob-Freese, BSN, Sanjiv Mehta, MD, Nathaniel Sergay, MD, Vinay Nadkarni, MD, Robert Grundmeier, MD, Christopher P. Bonafeide, MD, MSCE, Robert P. Sutton, MD, MSCE, Fuchiang R. Tsui, PhD, Naveen Muthu, MD

1Children’s Hospital of Philadelphia, Philadelphia, PA
2University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

Introduction: Early recognition of and response to clinical deterioration in hospitalized children are key national patient safety priorities for both health systems leaders and parents. The development of proactive critical care outreach teams (CCOT) has complemented identification and evaluation of patients deemed at risk of clinical deterioration by bedside teams (e.g., Watcher programs). Early warning scores (EWS) like the BedsidePEWS leverage clinical data to estimate risk of impending deterioration, but performance is highly variable and utility for a proactive CCOT team has not been evaluated. For the Stage 1 objective, we developed a data-driven predictive model for predicting critical deterioration events from routinely collected electronic health record (EHR) data and will compare algorithm performance to the Bedside PEWS.

Methods: We conducted a retrospective cohort study of pediatric inpatient ward hospitalizations (>48 hours) at a quaternary children’s hospital in Pennsylvania between January 2014 and December 2018. The primary outcome was a critical deterioration event (a ward-to-ICU transfer with subsequent initiation of vasopressors or positive pressure ventilation within 12 hours), a validated measure of deterioration associated with increased mortality. There were 72 raw readily available EHR variables used, including demographics, labs, vital signs, and nursing assessments. We constructed 14 time-series features for each numerical variable (labs and vital signs). Semi-structured nursing assessments were parsed and one-hot encoded. We developed a generalized linear model (GLM), gradient boosting model (XGBoost), and deep neural network (DNN) and performed nested cross-validation (CV) to optimize hyperparameters and evaluate model performance. We compared model discrimination using the receiver operating characteristic (ROC) and the precision-recall (PR) curve.

Results: For 57,233 hospital admissions across 35,998 patients, 3.6% (n=2,069) experienced a critical deterioration event during the admission. Case admissions had proportionally more males (55.8% vs. 50.3%) and primary Medicaid insurance (50.3% vs. 43.9%) than control admissions. After feature engineering, the 72 variables were transformed into 3,943 unique predictors. Figure 1 shows the ROC and PR curve for the three models compared to patients identified as Watchers by the bedside team. The XGBoost model, which included 587 features (of which 427 were lab-derived), had the best discrimination with an area under the curve (AUC) of 0.951. The GLM model had a similar AUC of 0.946, but used only 190 features. Watchers (patients at risk identified by bedside teams during usual clinical care) had a positive predictive value of 0.261 and a sensitivity of 0.197 sensitivity within 2-24 hours prior to the event.

Conclusions: We developed a machine learning model that had excellent discrimination in predicting critical deterioration events and a more than sufficient clinical lead time of 24 hours for critical care outreach. While the XGBoost model had the best performance, the GLM model was comparable and required a third of the features. Therefore, it may be more easily implementable and interpretable. The clinical team did not identify most patients experiencing deterioration in real-time clinical care as Watchers, suggesting the value of added risk prediction to inform CCOT. Limitations include development within a single hospital, the need for 48 hours of historical data, and the use of many types of EHR data (although all are commonly available). Verifying calibration, assessing model performance by subgroups, completing out-of-time-validation, and comparing to the current BedsidePEWS score will be key to move forward with a model implementation framework. We are currently prospectively comparing the model to the BedsidePEWS and Epic Deterioration Index to assess model performance relative to the implementation effort as this model cannot be natively implemented in the EHR. In Stages 2 and 3 of the AMIA AI Showcase, the utility of the risk prediction will be assessed by 1) evaluating the appropriateness of the display and workflow integration from a sociotechnical lens using EHR simulations (Stage 2) and 2) conducting a pragmatic alternating intervention trial to measure the clinical impact of the tool in identifying patients for CCOT (Stage 3).
Developing a Chest Radiography Generative Adversarial Network (CXR-GAN) for Medical Education Simulation

Parth K. Shah¹, Harshat Kumar, MSE², Eduardo Jose Mortani Barbosa Jr., MD¹, Gary E. Weissman, MD, MSHP¹

¹Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; ²Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA

Introduction

Timely recognition of common pathologies and clinical emergencies on chest radiographs (CXRs) is an essential skill taught during medical education. A generative adversarial network (GAN) that produces realistic CXRs could generate precise and novel images for clinical simulations and augment atlases of rare cases. This project aimed to train and validate CXR-GAN—a GAN capable of generating a wide variety of CXRs, both normal and with pathology.

Methods

To develop CXR-GAN, we collected 2,119 CXRs from the MIMIC-CXR database containing normal images and images spanning 12 different pathologies. We trained a StyleGAN2-ADA network using one NVIDIA Tesla P100 graphics processing unit until it had seen 2 million images (~44 hours). The network learns to generate CXRs from a latent space representation, which is a 512-element vector that informs image features. However, to modify existing real images, the latent space representations for the real images must be discovered such that, when passed through the model, would produce an approximation of the real images, i.e. the images must be embedded within the latent space of the model. Using an optimization algorithm, we embedded ten real CXRs. We then computed linear interpolations between pairs of embeddings and visualized the resulting generated images to demonstrate how CXR-GAN can be used to morph between two real CXRs, which is one potential application.

Results

Linear interpolations between pairs of embeddings demonstrate gradual accumulation of pathology, simulating the evolution of pathology over time; however, the model struggles to completely remove lung markings in pneumothorax, i.e. collapsed lung (Figure 1). Images produced by CXR-GAN reveal preservation of normal anatomy including continuous ribs and expected soft tissue contours.

Conclusion

We find that CXR-GAN can produce novel normal and pathological CXRs and demonstrates a promising approach for creating realistic image libraries and interactive simulations to support medical education efforts.
I-WIN: an Intensive care Warning INdex system for early prediction of clinical deterioration and intervention using data-driven machine learning modeling of electronic health record data

Fuchiang (Rich) Tsui, PhD1,2,3, Victor Ruiz, PhD1, Michael Goldsmith, MD2,3, Lingyun Shi, MS1, Matthew Scharf4, Maryam Naim, MD2,3, and Allan Simpao, MD2,3

1Tsui Laboratory, 2Department of Anesthesiology and Critical Care Medicine, Children’s Hospital of Philadelphia; 3Department of Anesthesiology and Critical Care, University of Pennsylvania, PA

Introduction: Intensive care units (ICUs) face key challenges and needs despite continuous investment in medical technology and personnel training. Common challenges faced in ICUs include high mortality rates compared with other hospital units (10%-29% in adult ICUs; 2%-6% in pediatric ICUs), increased admissions, medical errors, intensivist shortage, and alert fatigue.1 Despite state-of-the-art approaches showing promising advantages of using AI in ICUs, current research is limited in several aspects: the use of high dimensional electronic health record (EHR) data, intervention recommendation, and translational science to clinical practice mainly due to infrastructure barriers.2 In this study, we built and evaluated an AI system, the Intensive Care Warning Index (I-WIN) system1 at the Children’s Hospital of Philadelphia (CHOP) in two clinical aspects: deterioration prediction4 and intervention prediction.

Methods: Cohort Data: The CHOP Institutional Review Board (IRB) approved this study (IRB 18-015520). The study period was between 2014 and 2019. The study cohort included infants with palliated or unrepaird single-ventricle, ductal-dependent, or shunt-dependent congenital heart diseases admitted to a cardiac intensive care unit (CICU) at CHOP from birth until one of the following censoring events: age 6 months, conversion to biventricular circulation, Stage 2 palliation, heart transplant, or death. Additional inclusion criteria included required stage-1 single-ventricle palliation, pulmonary-artery banding followed by single-ventricle palliation, surgical systemic to pulmonary shunt, trans-catheter stenting of the ductus arteriosus, or maintenance of ductus arteriosus patency with prostaglandin-E as the primary source of pulmonary or systemic blood flow until the time of a censoring event.3 Exclusion criteria were admission in the CICU for <24 hours or died within 24 hours of CICU discharge. Cases were defined as a composite of critical deterioration events: emergent intubation, cardiopulmonary resuscitation (CPR), extracorporeal membrane oxygenation (ECMO) cannulation, or CPR with refractory cardiac arrest requiring ECMO (ECPR). Four data types were used in predictive models: vital signs from nursing flowsheets, laboratory test results, administered medications, and coded diagnoses.

Predictive modeling of deterioration events: For deterioration prediction, we developed an ensemble of 5 machine learning models (1,028 variables) using extreme gradient boosting (XGB) machine algorithm, and each of the models was developed from one of the 5 prediction horizons ranging from 1 hour to 8 hours before the onset of an event.4 Predictive modeling of intervention (medications): We used random forest for intervention prediction of medication categories in the same cohort. Six commonly administered medication (intervention) categories used in emergency situations were calcium gluconate, chlorothiazide, fentanyl, furosemide, morphine, and sodium bicarbonate. Model evaluation We performed a nested 10-fold cross validation for modeling and evaluation. The area under the receiver operating characteristic curve (AUC) was used for evaluation metric.

Results: Deterioration prediction: The cohort included 488 infants with single ventricle physiology with 203 critical events in 134 infants. At 4 hours before deterioration, the model achieved an AUC of 0.92 (95% CI: 0.84-0.98), 0.881 sensitivity, 0.776 positive predictive value, 0.862 specificity, and 0.571 Brier skill score. Performance remained high at 8 hours before deterioration with AUC 0.815 (0.688-0.921). Intervention prediction: In the window (8 hours to 30 minutes before deterioration onset), we observed AUCs of 0.857(0.767-0.922) for calcium gluconate, 0.853(0.763-0.923) for fentanyl, 0.741(0.583-0.886) for sodium bicarbonate, 0.718(0.632-0.802) for furosemide, 0.697(0.58-0.801) for morphine, and 0.675(0.511-0.808) for chlorothiazide.

Current Status and Plans: I-WIN currently is deployed at CHOP as an independent homegrown tool with pilot users from CICU intensivists for patient monitoring and case review. I-WIN uses secure web-based graphical user interfaces and adopts organizational user authentication to enhance usability. I-WIN collects and monitors high-speed waveform and EHR data form over 200 bedside monitors. We plan to add waveform data to the predictive models and embed the models’ outputs with explanation to I-WIN following event-driven approach. IRB approval for conducting user feedback has been obtained. We will evaluate usability, workflow, and the functional effectiveness of the I-WIN system.

References:
Predicting Thirty-day Unplanned Cancer Readmissions using Machine Learning and Artificial Intelligence

Danny T.Y. Wu, PhD, MSI\textsuperscript{1,2,3}, Hoang Vu\textsuperscript{1}, Tripura M. Vithala\textsuperscript{1}, Adam Alexander\textsuperscript{1,3}, Amy Roberto, MBA\textsuperscript{4}, James J. Lee, PhD\textsuperscript{1,2}, Thomas J. Herzog, MD\textsuperscript{1,3},

\textsuperscript{1}University of Cincinnati College of Medicine, \textsuperscript{2}University of Cincinnati Digital Scholarship Center, \textsuperscript{3}University of Cincinnati Cancer Center, Cincinnati, \textsuperscript{4}University of Cincinnati Health, OH, 45229.

Introduction
Unplanned cancer readmission is one of the primary outcomes of cancer treatments and a key indicator of the reduced quality and length of cancer survivorship.\textsuperscript{1,2} It can cause significant burdens to patients and increase hospital costs.\textsuperscript{3} Moreover, cancer patients readmitted into intensive care units in an unplanned manner face significantly higher mortality rates during their hospital stay or shortly thereafter.\textsuperscript{4,5} It is therefore imperative to identify intervention strategies to reduce unplanned cancer readmissions. Our first step to approach this problem was to develop a predictive model to identify patients likely to have a 30-day unplanned readmission using electronic health records with machine intelligence. This abstract reports the technical development process and the model performance.

Methods
A human-in-the-loop machine learning (ML) approach was adopted in this research, where domain experts were heavily involved in feature selection, label definition, and error analysis. Thirty-six months of clinical data (2017-19) were extracted from the EHR system at the University of Cincinnati (UC) Health through the Center for Health Information (CHI) at the UC College of Medicine. The dataset contained 2,701,620 admissions and 46,671 patients. The dataset contained 80 variables, 21 of which were selected by domain experts at the UC Cancer Center as key independent variables (Table 1). These variables were further narrowed down based on the risk factors identified in the literature. Specifically, a search (keywords: Cancer AND "Unplanned readmission") was conducted in PubMed. The retrieved papers were screened by two research team members, resulting in 81 papers and 8 risk factors. After considering the domain experts’ feedback, a filtered set of 13 variables was created. Both the full set (N=21) and the filtered set of variables (N=13) were used as our model input. The definition of an unplanned readmission was surveyed amongst the literature, and it was found to be quite varied. Tallying up the results and working with the domain experts resulted in our working definition: an unplanned readmission constitutes of a readmission that was not part of the patient’s treatment plan and/or was a readmission admitted to the emergency department. Following this definition, 28.5% of admissions (N=769,962) had an 30-day unplanned readmission, which were labelled as positive cases. The data were split into training and testing based on the months using 24-to-12 months or 30-to-6 months. Missing values in the numerical and categorical variables were filled with the mean and null, respectively. The numerical variables were scaled using mean and standard deviation; the categorical data were turned into binary variables using one-hot encoding. Five classic ML classifiers were selected, including logistic regression (LR), decision tree, random forest, Gradient Boosting Machine (GBM), Support Vector Machines with Linear Kernel (linear SVM). In the training process, a 5-fold cross validation with hyper-parameter tuning were employed. The Area under the Receiver Operating Characteristic Curve (AUROC) was used to determine the model performance. Moreover, Sensitivity (recall), Specificity (true negative rate), Precision (positive predictive value), Accuracy were reported to describe the model performance. Literature was surveyed to benchmark our model performance. The best AUROC for unplanned readmissions after a cardiac surgery was 0.79 using a deep learning model.\textsuperscript{6} The best AUROC for 30-day cancer readmissions in neuro-oncology was 0.76 with a specificity of 0.996, and a sensitivity of 0.07.\textsuperscript{7}

Results
Figure 1 shows the ROC curve for all models on the testing set. Our best model used 24 months of data and full variable set for training and the GBM algorithm with 400 trees. Our model achieved 0.8067 AUROC, which outperformed the best models in the literature. Moreover, the model performed at 0.9 sensitivity, 0.53 specificity, 0.41 precision, and 0.63 accuracy. Figure 2 shows the monthly prediction performance of our best model. Table 2 lists the top features, all of which were information known prior to the readmissions. At the time of submitting this abstract, we were preparing for the error analysis and had selected 50 false positive and 50 false negative admissions using clustering and stratified random sampling for domain expert review. We were also experimenting a deep learning algorithm (i.e., long-short-term-memory) that remembers the admission history of a patient to improve the predictive power.

Discussion and Conclusion
We successfully developed a ML model to predict 30-day unplanned cancer readmissions with good performance in our institution. We will continue improving the model by including more variables, experimenting artificial intelligence (AI)-based methods, and evaluating model calibration (e.g., Brier Score). Meanwhile, we will explore the context of use of the ML model in clinical routines, and further conduct a health impact study to demonstrate the effectiveness of the best-performed model as a clinical decision support tool to reduce unplanned cancer readmissions and improve cancer survivorship.
Evaluating vendor-derived pediatric sepsis predictive model in acute care settings
Feliciano Yu, Jr., MD1, Daniel Liu, MD1, Shengfan Zhang, PhD2
1University of Arkansas for Medical Sciences, Little Rock, Arkansas, 2 University of Arkansas, Fayetteville, Arkansas

Introduction
Healthcare institutions are leveraging their electronic health records (EHR) systems to implement clinical decision support (CDS) tools for identifying sepsis. (1,2) Recently, healthcare institutions have begun augmenting early sepsis identification with vendor-derived machine learning (ML) models using their own EHR data. (3) Our project aims to evaluate the performance of our existing early sepsis decision support models against the EHR-derived ML model and our own self-developed ML algorithms.

Vendor-derived Pediatric Sepsis predictive model
We adopted the vendor-derived Early Detection of Pediatric Sepsis (EDPS) model using Non-linear Reduced Gradient method, a specificity-focused model that runs on high-risk patients who are likely to develop sepsis, and then a Scaled Linear Model, which is a sensitivity-focused model that rules in patients from the general pediatric population. The EDPS model patient population included patient admitted to our hospital who are less than 18 y/o who were admitted in the inpatient setting, excluding those who were admitted in the ICU settings, and those seen in the emergency department. The Target/Label for the EDPS are patient encounters with a transfer to ICU for severe sepsis or septic shock in the inpatient (EHR diagnosis data), or ED encounters with sepsis alerts in the workflow. The EHR-derived Variables or Features include vital signs (temperature, pulse, respiratory rate, and systolic blood pressure), nursing assessments (capillary refill time, skin exam, peripheral pulse exam, mental status exam, PEWS), laboratory results (ALT, band neutrophils, WBC, lactate, procalcitonin, C reactive protein), and select high-risk condition categories. There are 211,915 encounters included in our hospital’s validation dataset, reflecting hospital visits from January 1, 2019, to November 1, 2021. About 1,193 of those visits met the criteria for our Target population. The EHR vendor provides a “validation utility” tool that calculates model scores on a retrospective population to estimate the model’s performance for our hospital. This validation utility looks for sepsis diagnoses in the hospital account or problem list.

Model Performance of Vendor-derived Pediatric Sepsis Model
The overall AUC performance of the model using our institution’s own EHR data is 0.81, accuracy is 0.85, true positive rate is 0.67, false positive rate is 0.151, positive predictive value is 0.024, and the negative predictive value is 0.99. We also calculated for the F1 score (0.05) and Matthews Correlation Coefficient, or MCC (0.11) to provide more insight into the model performance given the imbalanced class distribution of the data set. Although our AUC is 0.81, indicating a strong model, our F1 score and MCC resulted in less-than-ideal performance. We have begun data collection existing early sepsis warning decision support activities and other EHR healthcare data and will be able to analyze the datasets as soon as they are complete.

Conclusion
Our project will compare the performance of 1) our existing early sepsis warning decision support tools, 2) EHR-derived ML models for pediatric sepsis, and 3) self-developed ML models for predicting sepsis using healthcare data from our organization. We present the model performance from our EHR-derived pediatric sepsis model. We are currently collecting operational and EHR data to be able to evaluate our existing early sepsis models and the self-developed ML models. We will determine whether newer, more robust, machine learning models will perform better than existing models.

References
Fast-Tracking the Path from FHIR Implementation Guide to Implementation

Sara Armson, MS1 (Organizer); Bryn Rhodes2; Laura Haak Marcial, PhD1; Daniel Vreeman, PT, DPT, MS, FACMI1; Roland Gamache, PhD, MBA, FAMIA3

1RTI International, Research Triangle Park, NC; 2Alphora, Orem, UT; 3Agency for Healthcare Research and Quality, Rockville, MD

Abstract:
The interoperable exchange of data is facilitated by the data exchange standard HL7 FHIR® (Fast Healthcare Interoperability Resources)2. As FHIR technologies are integrated into health systems, health technology experts are tasked with becoming versed in the necessary baseline knowledge4,5 to implement these solutions. FHIR Implementation Guides (IGs), sets of rules for using FHIR resources to enable data exchange, build upon each other to create a complex yet powerful framework for interoperability3. However, for many systems integrators, the FHIR IG is the first introduction to the world of interoperability standards, and the current FHIR IG model contains gaps that make this introduction challenging. By supporting system integrators with FHIR IGs enriched with implementation support materials, interoperability teams will move beyond demonstration projects into the production environment to achieve long term sustainability.

Significance:
Leveraging and exchanging structured health data within and between disparate health record systems is the foundation of ground-breaking, outcome improving technologies, like clinical decision support (CDS)1. This interoperable exchange of data is facilitated by the data exchange standard HL7 FHIR® (Fast Healthcare Interoperability Resources)2, and FHIR has enabled an explosion of new interoperability capabilities. As FHIR technologies are integrated into health systems, health technology experts are tasked with becoming versed in the necessary baseline knowledge4,5 to implement these solutions. With a focus on lowering the barrier of entry for new FHIR implementers, a panel of technical experts (TEP) of interoperability pioneers was convened to develop best practices in utilizing pilot implementation materials (such as use cases and operations manuals) to facilitate future implementation by enriching FHIR Implementation Guides. Spanning diverse perspectives, including developers, researchers/academics, and a patient advocate, this panel sought to analyze and define the gap between sandbox and production implementation and present an accessible and feasible support approach for implementers.

IT departments within healthcare systems have varied levels of experience with interoperability standards, and there is a steep learning curve for a non-standards implementer to stand up new FHIR tooling. Additionally, adherence to interoperability standards presents unique challenges for health systems and often results in critical trade-off decisions. These departments are often at capacity focusing on maintaining their vendor systems and adding the ramp up and implementation of new FHIR tooling in this environment can overwhelm their resources if they are already at capacity. Innovative pilot projects that move standards-based FHIR technologies into practice may end up decommissioned if they overtax the implementation and support teams.

FHIR Implementation Guides (IGs), sets of rules for using FHIR resources to enable data exchange, build upon each other to create a complex and powerful framework for interoperability3. For an interoperability standards developer, a FHIR IG comprises the building blocks of an implementation and details the extensible, reusable components to leverage across projects. These developers have in-depth FHIR experience and can utilize these interoperability levers. While standards developers are the authors and most frequent audience of FHIR IGs, recent implementation experience indicates that these IGs serve various stakeholder groups along the interoperability continuum.

A FHIR IG is an important final product of any pilot of a shareable interoperable solution especially if the plan is to reuse the solution. The IG serves as a reusable shareable resource for other interested parties to extend implementation and adoption. For a system integrator at an implementation site, the FHIR IG provides a blueprint to make the connections between an application and corresponding EHR data via FHIR and attempts to help make
sense of harnessing the messy nature of EHR data. Because systems integrators commonly implement and support proprietary APIs (e.g., vendor-specific EHRs), the FHIR IG is often an initial introduction to less commonly used open, standards-based APIs, and the current FHIR IG model contains gaps that make this introduction hard. The panel presenters will detail these gaps and provide some suggestions for bridging them effectively.

The HL7 (Health Level 7) FHIR Clinical Decision Support Work Group has published a methodology for the relationship between distinct categories of IGs (Foundational, Model, Specification, and Content) and proposes that conceptually, the FHIR Content IG is the primary interoperability tool for system integrators. With this methodology in mind, the Technical Expert Panel (TEP) used lessons learned from an ongoing FHIR implementation pilot, the Agency for Healthcare Research and Quality (AHRQ) Clinical Decision Support for Chronic Pain Management (CDS4CPM) project. As a part of this project, the TEP considered a set of best practices for enriching FHIR Content IGs to better serve stakeholders along the implementation continuum. As interoperability improves, these IGs will increasingly serve as building blocks for standards literacy for an ever-broadening group of integrators. These best practices include providing detailed examples of pilot experiences with FHIR tooling and revealing the work done to ramp up and implement a FHIR solution. This documentation includes:

- Use cases that provide the first opportunity to look at data modeling and movement and act as a starting point for the data element crosswalk and examination of system requirements.
- Test scripts from the pilot help sites to run explicit tests to ensure the system produces the expected outcome or uncovers flaws preventing the proper execution of scripts.
- Wireframe models that can be reused and updated by the implementing sites serve as the connection between the system integrators and the rest of the project.
- Training and implementation resources that guide a site through the steps needed to realize the full potential of an implementation.

These types of site-oriented enriching materials provide a step-by-step guide to navigating the toughest points of data exchange. By supporting system integrators with these augmented FHIR Content IGs, interoperability teams will more quickly and easily move beyond demonstration projects into the production environment. In addition, as these tools become more prevalent in production environments, the path toward achieving long term sustainability of these interoperable solutions, the real promise of a learning health system, becomes more viable.

Description of the Panel

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>10'</td>
<td>Armson (moderator)</td>
<td>Introduces the panel and provides an overview of the CDS4CPM project, aims, and goals—then leads discussion during the audience Q&amp;A</td>
</tr>
<tr>
<td>12'</td>
<td>Marcial</td>
<td>Describes the implementation challenges and barriers encountered when launching the FHIR-based apps developed for the CDS4CPM project</td>
</tr>
<tr>
<td>12'</td>
<td>Rhodes</td>
<td>Discusses the utility of a Content IG and the role of IG developers in reducing the barriers to entry for system integrators</td>
</tr>
<tr>
<td>12'</td>
<td>Vreeman</td>
<td>Discusses the future state of FHIR IGs and how they should evolve to build stakeholder support on the path towards a plug-and-play future</td>
</tr>
<tr>
<td>12'</td>
<td>Gamache</td>
<td>Discusses how the IG development process is an important component of CDS development and how these tools are incorporated into the AHRQ CDS portfolio</td>
</tr>
<tr>
<td>30'</td>
<td>All</td>
<td>Discussion, Q&amp;A with audience</td>
</tr>
</tbody>
</table>

Learning Objectives

1. Identify the diverse users of FHIR IGs and consider the varying perspectives and needs of different users
2. Describe the challenges encountered by system integrators when deploying new FHIR technologies, especially when site system integrators are FHIR novices
3. Contemplate strategies to lower the barrier to entry for health systems to integrate FHIR-based technologies and the role of FHIR IG developers and researchers in supporting implementations

Individual Speaker Contributions
Sara Armson, MS – Ms. Armson is a clinical terminologist at RTI International who will serve as the panel moderator and introduce the panel participants.

Laura Haak Marcial, PhD – Dr. Marcial is a health informaticist at RTI International who will describe the specific CDS4CPM implementation challenges encountered during the application integration into the site’s health IT systems. This will include detailing how the project team arrived at documenting and addressing the needs implementers were identifying throughout the implementation process.

Bryn Rhodes – Bryn Rhodes will review the emerging categories of HL7 FHIR IGs and go into detail about the development and utility of Content IGs, implementation guides that are focused on knowledge artifacts such as decision support rules, quality measurement specifications, or more generally, computable clinical guidelines. This emerging category of FHIR content supports the authoring, representation, distribution, and evaluation of computable clinical knowledge, building on the semantic interoperability afforded by FHIR implementation guides to enable knowledge interoperability.

Daniel Vreeman, PT, DPT, MS, FACMI – Dr. Vreeman is a health informaticist at RTI International who will discuss a future state in which research translates stakeholder feedback into real-world solutions, ultimately reaching for a FHIR technology plug-in-play future.

Roland Gamache, PhD, MBA, FAMIA – Dr. Gamache is a Staff Fellow at AHRQ who will discuss how the FHIR IG development process is incorporated into the ARHQ CDS portfolio of work.

Expected Discussion and Discussion Questions

We expect the audience will engage the panelists in a discussion on solutions for overcoming implementation barriers in FHIR-based pilot projects. We anticipate the audience sharing unique challenges from their own experiences and deepening the discussion about solutions. The audience is expected to learn from the CDS4CPM project team’s approach to adding implementation support materials to Content IGs, and that this discussion will lead to future strategies for embedding FHIR technologies into practice environments. These discussion questions include:

1. How have your FHIR technology implementation experiences been similar or different than the CDS4CPM experience?
2. What types of implementation resource support would be helpful to include in a Content IG?
3. What are some different or additional avenues for supporting system integrators in the implementation of FHIR technologies?

Urgent Topics for Intended Audiences

This panel addresses pressing issues around developing, implementing, and managing the challenges of moving critical FHIR technologies into real-world practice. We will present a feasible approach to lowering the barrier of entry for FHIR-novice system integrators and invite discussion on extending and enhancing future solutions.

Attestation

The panel moderator has assurances from all participants that they will be available to participate at the AMIA Informatics Summit 2022.

References

How distributed clinical observations become big-data COVID research publications, using the N3C

Christopher G. Chute, MD, DrPH1, Emily Pfaff, PhD, MS2 Davera Gabriel, RN1, Charisse Madlock-Brown, PhD, MLS3 Melissa Haendel, PhD4

1Johns Hopkins University, Section of Biomedical Informatics and Data Science, Baltimore, MD; 2UNC Chapel Hill, Department of Medicine, Chapel Hill, NC; 3Tennessee Clinical and Translational Science Institute, The University of Tennessee Health Science Center, Memphis, TN; 4University of Colorado Anschutz, Denver, CO

Abstract

The National COVID Cohort Collaborative (N3C) has been a technical and social experiment to integrate electronic health record (EHR) data from academic and rural medical centers, across four data models. In this panel, we trace the path of data from ingestion to analyses, to publication - a year in the life of COVID research. The community is organized into multidisciplinary domain teams, we highlight the Social Determinants of Health team to illustrate this end-to-end process. We address data ingestions, data harmonization, project formation, collaborative analytics, attribution, and publication within the collaborative. We consider N3C a model for large-scale translational science across multiple institutions, engaging diverse investigators from many organizations self-assembling as a team, with attribution fairly shared over the consortium.

Introduction - Christopher Chute (moderator)

COVID-19 has precipitated a worldwide public health emergency requiring responsive action from all branches of medical science. Coordinating multidisciplinary teams to generate innovative solutions that address the most pressing health issues to the US population is a mission of the NIH National Center for Advancing Translational Sciences (NCATS). The National COVID Cohort Collaborative (N3C) is a partnership sponsored by NCATS that applies translational and open-science principles focused on providing pioneering, collaborative research methods on leading-edge technology to discover disease risk and progression, effectively inform treatment decisions, and equally optimize the health outcomes of those infected with COVID19 among diverse populations nationwide. N3C reflects “partnership, inclusivity, transparency, reciprocity, accountability, and security” by enlisting collaboration among a diverse set of clinical and technical experts from leading academic institutions in a highly effective public-private research partnership(1). The result of this collaboration is a research environment that addresses technical, legal and policy barriers to rapid discovery and dissemination of actionable clinical findings to address acute and long-term effects of COVID-19(2).

Ingestion and Data Quality - Emily Pfaff

N3C has engaged in a precedent-setting endeavor to centrally harmonize the four major common data models (CDMs; OMOP, PCORnet, ACT, TriNetX) into OMOP. This enables N3C’s EHR dataset to span 61 institutions while putting minimal burden on sites themselves. In order to ensure a consistent cohort definition across sites, N3C’s COVID phenotype(3) is translated into all four CDMs and four database dialects. Though the syntax differs, the definition remains the same, granting downstream users assurance that the larger N3C cohort shares a single set of inclusion criteria. Each participating site regularly submits patient, encounter, diagnosis, procedure, laboratory, medication and vital sign data in their CDM of choice; these data “payloads” are then picked up by N3C’s harmonization pipeline, transformed to the OMOP model, and evaluated for quality. N3C’s quality metrics cover both COVID-specific (e.g., number of COVID tests, plausible COVID test results, presence of COVID diagnosis codes) and disease-agnostic checks (e.g., plausible proportion of inpatient visits, adherence to source CDM conventions). These checks function as a way to catch critical data errors that could compromise data accuracy, but also serve to ensure as much inter-site data consistency as is reasonable. (Certainly, each site’s data contains local idiosyncrasies and inconsistencies that are acceptable so long as they are known.) The N3C team iterates with each site until all critical checks are passed, at which point the site’s data is included in the weekly N3C release set for use by researchers.
Enabling Investigators via Liaisons - Davera Gabriel

Collaborative N3C research projects involve multi-disciplinary teams whose members bring expertise in data engineering, analytics and statistics, and clinical care. Conducting robust and reproducible science in the N3C enclave requires that teams make effective use of OHDSI tools for cohort and codeset definitions, understand the structure, strengths, and weaknesses of both the data and software environment, and utilize community-built analysis pipelines and other shared resources. To support project needs N3C has developed both community-based training and 'liaison' support models, bootstrapping and linking research teams to core N3C resources while emphasizing knowledge transfer and cross-project interactions. Data Liaisons, who are also involved in core N3C ingestion and harmonization pipelines, are subject matter experts on N3C data and its appropriate use. Logic Liaisons consult on efficiently and effectively utilizing the high-performance, distributed-computing environment of the enclave; together they help researchers distill billions of rows of raw data into meaningful insights while ensuring maximal reuse and discoverability of code, data, and information. Additionally, the liaisons achieve cross-domain team visibility to concept sets and code workbooks developed in the enclave that have common focus or achieve the same analytic goals. In this way, the liaisons play an important role in realizing a key strength of the centralized analytic environment by promoting reuse of common “knowledge objects” throughout the widely diverse and wholly independent investigator community. While both liaisons and core N3C personnel develop training materials in the form of tutorials, documentation, and office hours, the wider community additionally contributes training resources, wiki-style documentation, and participates in regular cross-domain and user-group meetings organized by N3C. Fostering a collaborative and supportive research environment is thus a crucial component of N3C addressing the diverse skills and needs required by large-scale team science.

Publication and Attribution - Melissa Haendel

In the face of the pandemic, it has been necessary to work as efficiently and effectively together as possible. This requires a strong dedication to the development of shared research artifacts, such as the aforementioned codesets, algorithmic workflows, data visualizations, etc. However, such a shared environment also therefore requires the ability to attribute all those that contribute. In some cases an artifact, such as the “Covid positive” codeset, may be used by hundreds of downstream analyses. The N3C Enclave allows the transitive association of all who have contributed to be tracked. The N3C has developed an inclusive attribution and publication policy, which focuses on a three tier model of tracking authorship contributions: Masthead authors that have contributed substantially, N3C consortial authors that meet ICMJE guidelines but are included as consortial authors for smaller contributions such as key artifact reuse, and acknowledgements for contributors that do not meet the ICMJE criteria for authorship. Once analyses are ready for publication, the authors submit a download request with the extracted relevant summary tables and figures for review to ensure that cell sizes are not less than 20 and that the data meets privacy and sharing requirements. The authors also submit their proposed manuscript to a Publication Committee, that in addition to reviewing that the data policies are adhered to, also helps the authors collect information from the different communities of contributors and determines the nature of each contribution and authorship. This process ensures that even those that have been tangentially involved in the writing or analyses but have otherwise contributed substantially are acknowledged at the correct tier of authorship. We have found that this process ensures that even in the large scale of the N3C collaborative analytics community where much of the work is shared and transitive, that all contributors are properly acknowledged and incentivised to work together.

Social Determinants of Health Domain Team - Charisse Madlock-Brown

Studies indicate that older adults, males, and minority populations are at greater risk of infection, morbidity, and mortality from COVID-19 in the United States(6–8). There is an urgent need to understand the root causes of these disparities. Social determinants of health (SDoH); the conditions in which people live, work, and age; are significant factors associated with health disparities for a wide variety of health disparities(9). The N3C SDoH domain team focuses on public data analysis and N3C patient-level data related to SDoH across multiple dimensions and COVID-19 trends and outcomes. The team focuses on data quality and comprehensiveness issues, interoperability of patient-level SDoH data collection tools, and racial disparities. The team has one paper accepted by the AMIA 2021 Symposium on harmonizing SDoH variables in EHRs(10), one journal article in development in response to an invitation from the Advanced Genetics journal, and an invited book chapter under review for the upcoming Springer Book, "Reimagining Personal Health Informatics for Precision Medicine and Healthcare." We have additional research in progress related to data quality of race/ethnicity data variables, county-level trends, and racial disparities in outcome severity.
Conclusion

The N3C has succeeded in its aspiration to 1) define a governance framework for centralized EHR data from multiple organizations, 2) ingesting and harmonizing that data into a single common data model, 3) maintaining a secure enclave that protects privacy while enabling analytics, 4) fostering multidisciplinary domain teams across participating organizations to efficiently collaborate on sophisticated analytics, 5) creating a culture and publication framework that attributes contributions fairly, 6) generating a steady stream of high-impact publications that would be difficult to produce absent N3C infrastructure. We demonstrate N3C as a model for large-scale data science on clinical records within the Clinical and Translational Science Award (CTSA), the IDeA-CTRs, and related communities.

Sample Questions

1. What precedents exist for similar projects providing research access to large-scale EHR data?
2. How does data and project governance play a role in sustaining such a program?
3. What will become of N3C in the post-COVID era?
4. How well has shared attribution and collaboration played out in N3C?
5. How many other domain teams, like the Social Determinants of Health, have been created, and how productive are they?

All panel participants named above have agreed to take part in this panel. The project described was supported by the Center for Data to Health (CD2H) under NIH/NCATS U24TR002306, and NIH/NIGMS U54GM104942. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Building Data Capacity for Patient-Centered Outcomes Research (PCOR) for the Next Decade

Nancy De Lew¹, MPA, MA, Susan C. Hull MSN, RN², Katherine K. Kim, PhD, MPH, MBA², Scott R. Smith, MSPH, PhD¹, and Sara Wei, MHA¹
¹Office of Health Policy, Office of the Assistant Secretary for Planning & Evaluation, Department of Health and Human Services, Washington D.C.; ²The MITRE Corporation, McLean, VA.

Abstract
Data capacity is critical for patient-centered outcomes research (PCOR) to produce valid findings and new knowledge that can inform health care decisions. PCOR studies are designed to address questions that are important to patients and their families about the outcomes and effectiveness of treatments, services, and other health care interventions. The Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) portfolio team works with HHS agencies and research networks to build data capacity that enables researchers to generate new scientific evidence on patient outcomes. This panel will discuss the 2021-2030 OS-PCORTF Strategic Plan, which provides a framework to address four high-priority areas to advance data capacity for PCOR studies. Participants will be able to (1) identify key issues and challenges that represent opportunities to strengthen data capacity, and (2) describe the role of the OS-PCORTF in continuously improving the capacity for collecting, linking, and analyzing data for research.

A general description of the panel and the issue(s) that will be examined and a brief description of each panelist’s presentation
The Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) portfolio team works with Health and Human Services (HHS) agencies and research networks to build data capacity that enables researchers to generate new scientific evidence on patient outcomes. This panel will discuss the 2021-2030 OS-PCORTF Strategic Plan which provides a framework to address four high-priority areas to advance data capacity. It will highlight what has been learned from agencies, researchers and the public regarding research and data needs; review trends and drivers in the scope, focus, technology and governance of PCOR; and chart the course for HHS and other federal health agencies to work collaboratively and in partnership with other entities to build data capacity for PCOR. Participants will be able to describe the role of the OS-PCORTF portfolio in building data capacity and infrastructure for PCOR; and identify anticipated changes on data capacity for PCOR studies.

The 90-minute panel will be timed to allow one-half of the session for audience participation. Questions for the audience will be posed for discussion with two fifteen-minute intervals, and participant feedback on the Strategic Plan will be welcomed during an additional fifteen-minute interval.

Panel organizer and moderator: Nancy De Lew, MPA, MA, Associate Deputy Assistant Secretary for Health Policy, the Office of the Assistant Secretary for Planning and Evaluation (ASPE) at HHS

Background and Introduction
Established by Congress in 2010, the Patient-Centered Outcomes Research Trust Fund (PCORTF) supports the efforts of the Patient-Centered Outcomes Research Institute (PCORI), the Agency for Healthcare Research and Quality (AHRQ), and the Office of the Secretary of HHS, to conduct, disseminate, and expand data capacity for PCOR and comparative effectiveness research (CER).¹ The Office of the Assistant Secretary for Planning and Evaluation (ASPE), under delegation of authority by the Secretary of HHS, coordinates across relevant federal health programs to build data capacity for PCOR, including administering the OS-PCORTF. This coordination involves partnerships with agency leaders, scientists, research programs, and data stewards to develop and implement an extensive array of projects that improve the availability and suitability of data, as well as the analytic resources for answering questions that are important to patients, caregivers, clinicians, and policymakers.

Following the reauthorization of the PCORTF in 2019², ASPE led the development of a strategic plan for the next decade of the OS-PCORTF. The Strategic Plan was developed through a comprehensive review of the literature; interviews with agency leaders, program officials, and data stewards; a series of public meetings; and extensive engagement, input and close collaboration with a committee of HHS agency representatives. The Strategic Plan is responsive to evolving data infrastructure needs, priorities, and relevant developments, including legislative or
policy changes and advances in health care, data science, and the needs of decision makers who use the findings from PCOR studies.

**Panel Member 2: Katherine K. Kim, PhD, MPH, MBA, FAMIA,** Principal, Consumer Health Informatics and Health Science, The MITRE Corporation

**Trends for the 2021-2030 OS-PCORTF Strategic Plan**
HHS agencies routinely collect, link, and analyze data that can be used to generate new scientific evidence that expands knowledge about the outcomes and effectiveness of healthcare treatments and interventions. Although the potential is great for secondary uses of data, the current use of clinical, person-generated, and administrative data for research on nationally important issues is limited. Additionally, trends in current environment impacting the scope and focus of PCOR, in how PCOR is conducted, in research data governance models, in federal health programs’ roles and data capacity, and government and regulatory environment for fostering innovation in relation to PCOR will be discussed.

**Panel Member 3: Susan C. Hull MSN, RN-BC, NEA-BC, FAMIA,** Principal, Consumer Health Informatics, The MITRE Corporation

**Foundations for the 2021-2030 OS-PCORTF Strategic Plan**
In consideration of the broad community of producers and users of health data for PCOR, the Strategic Plan was developed with the participation of internal and external stakeholders and the public. Engagement activities invited the perspectives of researchers, policy analysts, external government and non-governmental organizations, HHS agency leaders, OS-PCORTF participants, and other PCOR producers and users. ASPE and the Health Federally Funded Research and Development Center (Health FFRDC), operated by MITRE, conducted an environmental scan, observations and interviews with clinical registries and health outcomes research data networks stakeholders, interviews with HHS stakeholders, and observations from recently funded COVID-19 PCOR projects. The compilation of findings can be found in multiple reports, including Challenges and Improvements for PCOR Data Infrastructure: Results from a Stakeholder Prioritization Activity, Research Data Networks and Patient-Centered Outcomes Research Trends and Opportunities: Scan and Interviews with Key Informants, and Building Data Capacity for Patient-Centered Outcomes Research (PCOR) for COVID-194,5,6.

**Panel Member 4: Scott R. Smith, MSPH, PhD,** Director of the Division of Health Care Quality and Outcomes, Office of Health Policy, the Office of the Assistant Secretary for Planning and Evaluation (ASPE) at HHS

**Public Engagement for Setting Strategy**
ASPE also engaged the National Academies of Sciences, Engineering and Medicine to convene a study committee and series of public workshops to identify issues critical to building data capacity and generating new evidence to inform health care decisions for PCOR into the next decade. Findings and conclusions to help guide a future course will be presented and published in: Building Data Capacity for Patient-Centered Outcomes Research: An Agenda for 2021 to 2030.7

**Panel Member 5: Sara Wei, MHA,** Public Health Analyst, Division of Health Care Quality and Outcomes, Office of Health Policy, the Office of the Assistant Secretary for Planning and Evaluation (ASPE) at HHS

**Elements of the OS-PCORTF’s Strategic Plan (2021-2029)**
Robust data capacity and infrastructure are necessary to improve evidence generation, decision making, and health outcomes for all Americans. To achieve this vision, the OS-PCORTF Strategic Plan targets high-priority opportunities to address critical data challenges for PCOR in four areas:

- Building and strengthening data capacity related to national health priorities.
- Advancing data harmonization, privacy-preserving linkage methods, and other approaches to enhance the availability, quality, accessibility, and suitability of linked data for PCOR.
- Leveraging advanced technology solutions to improve the utilization of large volumes of data as well as the variety and timeliness of data availability for PCOR, thereby increasing the richness and the robustness of the evidence generated.
- Expanding the collection and use of socioeconomic, environmental, and other data to support a more comprehensive understanding of health, identify and address disparities, and examine economic outcomes.
Because the Strategic Plan will be implemented through a set of strategies that allow for adaptation based on evaluations from both internal and external efforts, it will remain a consistent guide for HHS in carrying out its mission to strengthen data capacity for PCOR through coordination across agencies.

A list of possible discussion questions to enhance audience participation.
1. How do you see your work fitting within the goals, objectives, and planned efforts of the OS-PCORTF Strategic Plan?
2. As the OS-PCORTF portfolio continues to support HHS to work collaboratively and build data capacity for PCOR, what are your priority needs to (a) use existing data or link datasets across Federal and State sources, and (b) improve data accessibility and interoperability to strengthen longitudinal data resources?
3. What are the most pressing technical and/or non-technical challenges that create barriers to improving data capacity for PCOR (e.g., data sources, standards, methods, resources, policies, etc.)?
4. What challenges do you anticipate will be most significant in adopting advanced informatics research approaches to actively participate in a comprehensive, interoperable ecosystem for PCOR outcomes and effectiveness?
5. What are the most effective strategies for increasing dissemination and uptake of products to end-users? Describe by type of product and/or end-user.

An explanation why the topic of this session is timely, urgent, needed, or attention-grabbing.
Answering PCOR questions require the ability to link and aggregate vast amounts of data from an increasingly more varied set of sources. Healthcare data in the United States (U.S.) are held by a wide variety of distinct and independent entities and in many different formats. It is projected in 2020, the U.S. healthcare system generated more than 2,314 exabytes of data, and the volume is increasing exponentially because of breakthroughs in digital health. Furthermore, new technologies like AI, ML, and blockchain are being applied to large data sets for researchers to find new ways to cure or manage costly health conditions. In the U.S., health care spending reached $4.1 trillion in 2020 and it is projected to reach $6.2 trillion by 2028, and there is growing recognition that big data and new technologies will transform the diagnosis, treatment, and prevention of disease in the 21st century healthcare system.

A statement from the panel organizer that all participants have agreed to take part on the panel.
All panel participants named have agreed to take part in the panel.

References
2. Patient Protection and Affordable Care Act of 2010 (US)
Generating Synthetic Health Data to Accelerate Patient-Centered Outcomes Research (PCOR) and the Evaluation of an Open-Source Synthetic Data Platform for Simulation Studies

Stephanie Garcia, MPH¹, Daniella Meeker, PhD², Casey Thompson, MSN, RN-BC³
¹Office of the National Coordinator for Health Information Technology, Washington, DC; 
²Keck School of Medicine, University of Southern California, Los Angeles, CA; 
³Clinovations Government + Health, Washington, DC

Abstract

Researchers and health information technology developers often depend on data from health records to test theories, models, algorithms, or prototype innovations. Due to legal and privacy concerns, data must often be aggregated or de-identified before it can be released, often introducing costs and delays. Synthetic health data helps address these issues and speeds the initiation, refinement, and testing of innovative health and research approaches. Capitalizing on this opportunity, the Office of the National Coordinator for Health Information Technology funded an effort to evaluate and enhance an open-source synthetic data engine, known as Synthea™, to accelerate research. The Synthea platform, originally designed for generating realistically coded health records for software testing, implements publicly contributed clinical modules representing life cycle and disease and treatment progression. A Synthea module was iteratively developed using published parameters from a recent study of acute myeloid leukemia; we replicated the initial conditions and simulation endpoints of demographics, health events, costs, and mortality. We compare Synthea’s Generic Model Framework to platforms designed for simulation and show that Synthea can be used, with modifications, for simulation studies similar to prior work.

Introduction

The Office of the National Coordinator for Health Information Technology (ONC) has led and collaborated on numerous projects that inform policy, standards, and services specific to the adoption and implementation of a patient-centered outcomes research (PCOR) data infrastructure. Projects funded by the Patient-Centered Outcomes Research Trust Fund (PCORTF), which is administered by the Assistant Secretary for Planning and Evaluation (ASPE), support the development of data capacity and infrastructure that engage patients in health care decision-making and incorporate their responses into research. This work supports the development of an integrated electronic environment that continuously supports evolving demands for health information and adapts to novel and promising technologies.

Clinical data are critical for the conduct of PCOR, which focuses on the effectiveness of prevention and treatment options and care delivery models. Realistic patient data are often difficult to access because of cost, patient privacy concerns, or other legal restrictions. Researcher or health information technology (health IT) developers will typically need to aggregate, de-identify, and analyze data before testing the effectiveness of algorithms and modeling approaches used in matching and disease-modeling techniques. Synthetic health data can help relieve some of this burden and be used to initiate, refine, or test innovative research approaches more rapidly. Synthetic health data can be generated using a variety of mechanisms, including algorithmic approaches that use real patient data and synthetic health data that is generated based upon rules that reflect real-world patient data.

Synthea™, a synthetic health data engine developed by the MITRE Corporation, employs an open-source development model, uses publicly available data to generate synthetic health records, and can export information in multiple standardized formats. Synthea generates realistic patients, simulates their entire life, and outputs synthetic but realistic health record data. Synthea generated data have also been used in academic institutions for educational and research purposes. This type of synthetic data engine can support the greater PCOR data infrastructure by providing researchers and health IT developers with a low risk, readily available synthetic data source to provide access to data until real clinical data is available.

Simulations are used in healthcare research when desired empirical data is not available, particularly when uncertainty and unexplained variation might impact outcomes. ONC funded a demonstration study (demonstration) to evaluate Synthea as a platform for simulation studies of patient healthcare utilization and outcomes. Such studies produce a simulated data set for statistical analysis comparing alternative treatment or policy scenarios or forecasting alternative future circumstances, including outcomes such as disease progression, quality of life, and mortality. This project
evaluated whether Synthea is suitable for simulation studies by creating a Synthea module that replicates a recently published cost-effectiveness simulation. We assessed whether features of Synthea were sufficient for replicating without modification to the source code. As discussed in prior work,7 models using only pathways from guidelines and quality measures, which is recommended on the Synthea wiki,8 does not produce realistic distributions of outcomes. For these reasons we developed and tested a more complex model using parameters from a peer-reviewed simulation study.

Panel Objectives and Presenters
Panelists will discuss the types of synthetic data available for researchers and how this project used Synthea to model a simulation study. Panelists will also describe their experiences using synthetic data, including benefits and limitations of using synthetic data for research.

This panel will 1) present an overview of the project’s purpose and ONC’s efforts to increase the availability of synthetic health data for research and other uses; 2) describe synthetic health data and how Synthea can be used, with only configuration modifications, to replicate certain types of simulation studies; and 3) discuss generating synthetic data and its limitations (e.g., dependency on underlying models) and opportunities for future research in method development.

Ms. Stephanie Garcia (moderator) is the ONC PCOR Portfolio Manager. Ms. Garcia will moderate this session, provide an overview of ONC’s relevant portfolio of work, and describe the goals of the Synthetic Health Data Generation project.

Dr. Daniella Meeker (panelist) is an Associate Professor at the University of Southern California. She serves as head researcher on the demonstration study and will discuss the types of simulation software currently available to researchers and how Synthea was repurposed for a simulation study. She will also discuss limitations to synthetic data, including its dependency on underlying models, and opportunities for future research.

Ms. Casey Thompson (panelist) is a Clinical Informaticist at Clinovations Government + Health. She serves as a project lead and will provide an overview of the project objectives and synthetic data generation approach, solicit feedback on project resources for others to generate their own data, and discuss how a Synthea module was developed to replicate an existing simulation study.

Panel Learning Objectives
1. Participants will learn about the different types of synthetic data and understand current and prior validations and limitations of Synthea data in outcomes research.
2. Participants will understand how to generate synthetic data based upon project findings, including accomplishments to-date and opportunities for researchers and health IT developers to validate use cases for Synthea-generated synthetic health records.
3. Participants will learn about the benefits and limitations of using synthetic data for simulation studies and outcomes research.
4. Participants will learn about tools and resources to generate their own synthetic health data using the open-source Synthea data generation engine.

Panel Discussion Questions
1. What types of synthetic data are available to researchers and how are they used? What are the benefits and limitations to each type?
2. How does such Synthea data compare and contrast to data generated with commercial microsimulation software?
3. What improvements to the Synthea data generation engine and its data outputs are essential to support the use of synthetic data by researchers?

Statement of Participation
Each of the panelists and the moderator have confirmed that they will participate if this submission is accepted, at the assigned timeslot during the AMIA Informatics Summit.
Conclusion

Synthea was not designed for simulation studies, but the Synthea Generic Module Framework includes many features of commercial microsimulation software. A module was iteratively developed using published parameters from the original study; the module design was adapted to replicate the initial conditions and simulation endpoints of demographics, health events, costs, and mortality. Synthea’s Generic Model Framework was then compared to platforms designed for simulation to demonstrate if Synthea can be used, with modifications, for some types of simulation studies and to support PCOR research.

References

Using Clinical Informatics to Improve Chronic Pain and Opioid Management in Real-World Clinical Practice

Panelists
Christopher A. Harle, PhD, Laura Haak Marcial, PhD, Meredith C. B. Adams, MD, MS, FASA, FAMIA, Neda Laiteerapong, MD, MS

Moderator
Roland Gamache, PhD, MBA, FAMIA

1University of Florida, Gainesville, FL; 2RTI International, Rockville, MD; Wake Forest Baptist Health, Winston-Salem, NC; 4University of Chicago, Chicago, IL; Agency for Healthcare Research and Quality, Rockville, MD

Abstract
Chronic pain affects an estimated 50-100 million Americans. In clinical practice, effective management of chronic pain is complex when incorporating each person’s pharmacologic considerations, and comorbid physical health conditions. Best practices integrate patient goals and preferences to support shared decision-making around risks and benefits of treatment options. Tied to pain management, The opioid epidemic is closely tied to pain management, supporting the need for risk mitigation when prescribing opioids. One key strategy to address the complex problem of chronic pain management is to provide clinicians with clinical decision support tools for opioid and non-opioid treatment options. For this panel, we present four different investigator-initiated clinical informatics solutions using clinical decision support tools to improve management of chronic pain.

Introduction
Chronic pain is highly prevalent and negatively impacts millions of Americans. Effective management of chronic pain is complicated by the shifting perspectives on effective treatment options. Specifically, opioids, historically a mainstay of chronic pain treatment, has led to overdose deaths exceeding 93,000 in 2020, reflecting a 30% increase from 2019. As a result, while some providers want to decrease or taper opioids, without safe alternative pain treatments, weaning opioids has been associated with higher rates of mental health comorbidities, including suicidality and mortality. Additionally, treatment for patients with chronic pain on opioids has been highly stigmatized. Optimal chronic pain treatment requires interdisciplinary care through the right modality to the right person via the right channel at the right time in workflow. Clinical informatics tools can enable the complex integration of medical information, patient-reported data, and the full spectrum of treatment options to facilitate informed shared-decision making. In this panel, a multidisciplinary group of presenters, will discuss their research and technology development projects aimed at improving pain care and ensuring safe opioid prescribing through clinical decision support and other health information technology. Their projects shared similar goals and objectives, yet highlights their different approaches in the design, development, implementation, and evaluation of their proposed solutions.

Why this Panel is Timely and Urgently Needed
Biomedical informatics tools have a significant role to play in supporting effective pain treatment, reducing opioid misuse or opioids-related adverse events, and treating opioid use disorder. In settings where providers had limited resources to offer comprehensive care for patients with chronic pain and opioid use, clinical decision support (CDS) tools proved to increase provider adherence to treatment guidelines. This panel is very timely and urgent because chronic pain and opioid use are highly prevalent and rates of overdose have reached epidemic proportions, now exacerbated by the COVID-19 pandemic. Clinical informatics solutions meets the challenge of an increasingly complex healthcare environment, helping a heavily burdened primary care system meet the needs of patients in an evidence-based, patient-centered manner. The composition of this panel includes diverse scientific backgrounds and approaches to pain informatics and brings both clinical and research expertise to these solutions.
General Description of the Panel
In this panel, four researchers from different disciplines will describe several novel efforts to use data and informatics tools to support safer and more effective pain management and opioid prescribing. The panelists will describe efforts that span multiple institutions, and primary and specialty care. The panelists will each discuss successes, challenges, lessons learned, and informatics tools developed and evaluated to advance knowledge and care delivery. These presentations will be used to stimulate an extensive audience discussion about the informatics challenges and solutions being explored, implemented, and studied around the country to improve pain treatment and respond to the opioid crisis.

Panelists
Dr. Harle is a Professor of Health Outcomes and Biomedical Informatics and the Chief Research Information Officer at University of Florida Health. Dr. Harle will discuss more than ten years of federally-funded research to understand clinical information needs and decision making for primary care treatment of chronic pain, and the translation of that understanding to EHR-based CDS tools. Specifically, he will describe several key results of user-centered design studies, and the incorporation of those results into the Chronic Pain OneSheet. COPE, “Improving Chicago Older Adult Opioid and Pain Management through Clinical and Self-Care Opioid, Pain Management, and Evaluation” (COPE) includes several components: provider education, a patient pre-visit questionnaire, and an electronic health record. Additionally, Dr. Adams is the Pain, Substance Use, and Behavior Domain Informatics lead for the University of Pittsburgh Hub and Spoke Pain Clinical Trial Network (EPPIC-Net) serves as part of the NIH’s Helping to End Addiction Long-term (HEAL) Initiative. She will discuss progress and opportunities for scaling and disseminating key design features of OneSheet via interoperable CDS in EHRs.

Dr. Adams is Assistant Professor of Anesthesiology and Public Health Sciences at Wake Forest Baptist Health and the only NIH funded researcher board certified in Anesthesiology, Pain Medicine, and Clinical Informatics. Dr. Adams is the PI of the NIH HEAL Networks’ IMPower data and coordination center for this combined chronic pain and opioid use disorder clinical trial network. She also serves as the site PI and Informatics lead for the University of Pittsburgh Hub and Spoke Pain Clinical Trial Network (EPPIC-Net) EPPIC-Net serves as part of the NIH’s Helping to End Addiction Long-term (HEAL) Initiative. She will discuss her NIH National Institute of Biomedical Imaging and Bioengineering ongoing work on chronic pain and the electronic health record. Additionally, Dr. Adams is the Pain, Substance Use, and Behavior Domain Lead for the NIH NCATS National Covid Cohort Collaborative (N3C) Data Warehouse. She will discuss the extension of this work in the NIH N3C Data Warehouse and the relationship between chronic pain and COVID-19/PASC (long covid).

Dr. Laiteerapong is a board-certified general internist and clinical informatician and Associate Director for the Center for Chronic Disease Research and Policy at the University of Chicago. She will discuss the AHRQ-funded I-COPE project. I-COPE, “Improving Chicago Older Adult Opioid and Pain Management through PCCDS and Project ECHO®,” is a primary care, clinic-based toolkit for more effective management of chronic pain, opioid prescribing, and as needed, OUD, in older adults. The aim of I-COPE is to improve clinical and self-management of chronic pain, opioid use, and opioid use disorder (OUD) in older adults living in Chicago. I-COPE includes several components: provider education, a patient pre-visit questionnaire, an
adaptive EHR order set with patient-centered clinical decision support, a shared decision-making conversation tool, and a customized patient visit summary.

Dr. Gamache is a staff fellow in the Digital Healthcare Research Division in the Center for Evidence and Practice Improvement at the Agency for Healthcare Research and Quality. His areas of focus include the effective use of information technology that suggests interventions to improve health outcomes and to enhance operational processes and tools that improve the utilization of health data. Dr. Gamache will moderate this panel, discuss overlapping themes in the panelist presentations, and lead an audience discussion of key informatics challenges and opportunities to improve pain care. All participants have agreed to take part on the panel.

Learning Objectives
At the conclusion of this panel, audience members will be able to:

1) Identify 3 current clinical informatic challenges to appropriate chronic pain treatments
2) Propose informatics solutions for gathering patient-reported pain information in clinical practice.
3) Solve technical challenges and solutions for improving the clinical management of chronic pain.

Discussion Questions
The panelists will encourage in-depth discussion from the audience. Several discussion questions are listed below:

1) What challenges do institutions and clinicians face in accessing and using data to assess pain and opioid-related risks and outcomes? What solutions have you identified?
2) What information exchange and application standards (e.g., FHIR, SMART) are being used in decision support for pain care, opioid prescribing, and substance use disorder care? What successes and challenges have been faced in implementation and dissemination? How are they being evaluated?
3) What informatics tools are being implemented and used to support more patient-centered pain care? How are patient-reported outcome measures being used? What successes and challenges have been faced?
4) How challenging was it to get and sustain stakeholder buy-in to build and implement this system?
5) Which experts did you need on your team to successfully build and implement your system?
6) What elements of your system are directly transferrable to another system vs. locally contextual?
7) How can the lessons learned from your projects be transferred to other topics in medicine?

References
Ensuring Quality: A Core Competency of Federated EHR Data Networks

Jeffrey G. Klann, PhD\textsuperscript{1,2,3}; Darren W. Henderson\textsuperscript{4}; Shyam Visweswaran, MD, PhD\textsuperscript{5};
Hossein Estiri, PhD\textsuperscript{1,2}; Shawn N. Murphy, MD, PhD\textsuperscript{1,2,3}

\textsuperscript{1}Lab of Computer Science, Massachusetts General Hospital, Boston, MA; \textsuperscript{2}Harvard Medical School, Boston, MA; \textsuperscript{3}Research Information Science and Computing, Massachusetts General Hospital, Boston, MA; \textsuperscript{4}University of Kentucky, Lexington, KY; \textsuperscript{5}University of Pittsburgh, Pittsburgh, PA

Abstract

The growing prevalence of Clinical Data Research Networks supports nation- and even world-wide research on electronic health record data. Although national programs like PCORnet and ACT have developed network resources over years, COVID-19 accelerated development of innovative new approaches, including the Consortium for Clinical Characterization of COVID-19 by EHR (4CE), an innovative decentralized network with multiple source data formats. The panelists have been involved in methods and tool development for both ACT and 4CE and will identify core network components that ensure data quality and interoperability. These include: consistent organization and term naming (i.e., a biomedical ontology); multi-site evaluations of data completeness; and, deep harmonization across sites through chart review and validation. Together, these enable large-scale data analytics including machine learning in these networks. This panel will discuss these components (naming, completeness, deep harmonization, and impact on analyses) with practical examples that allow these networks to transcend previous accomplishments.

Introduction and Background

The promise of Clinical Data Research Networks is the ability to harness EHR data to make generalizable discoveries using data on large swaths of the population. Some networks, like the Patient Centered Outcomes Research Network (PCORnet) \cite{1,2} and the Accrual to Clinical Trials Network (ACT) \cite{3}, have been in development for years and have yielded many methodological improvements in data quality and interoperability. Others, such as the Consortium for Clinical Characterization of COVID-19 by EHR (4CE) have been built in response to COVID-19 and respond with great agility to the continuing pandemic. \cite{4} With the establishment of data research networks as an emerging commodity, we believe it is possible to identify core components needed to ensure network data quality and the generalizability of findings.

To ensure quality, a network must include the following three components. First, sites need to adopt consistent naming in an organized manner across the network. Network sites must agree on the biomedical ontology used for data in the network which must be adapted to each site through mappings. \cite{5} Second, data must be evaluated for completeness. Many existing data quality checks focus on completeness \cite{6} (e.g., “what percent of diabetic patients with a prescription have an A1C test?”). A newer approach is to identify subsets of patients who are ‘loyal’ - i.e., have complete data because they receive most of their care at an institution that is a member of the network. \cite{7} Third, data must be harmonized across sites. We call this “deep harmonization” because it goes deeper than just mapping codes. We advocate comparing data distributions across sites to ensure consistent mapping, and validation through both chart review and comparison of the underlying data and codes when applying phenotyping algorithms. Code mappings can dramatically vary from site-to-site. \cite{8}

Panel Description

Dr. Shawn Murphy will lead the panel discussion on the core components of quality in data research networks, and in doing so will help the audience better understand the methodologies that can enable high-quality clinical data research within the U.S. and internationally. Dr. Murphy is a Professor of Neurology and Biomedical Informatics at Harvard Medical School, Chief Research Information Officer at Mass General Brigham, and the Associate Director of the Lab of Computer Science at MGH. He is one of the Principal Investigators for the Data Resource Core of the large NIH RECOVER program that focuses on the study of post-acute sequelae of COVID-19 infection.

The panelists will discuss these core quality-enabling components with perspectives and tools from the 4CE Consortium and the ACT network. The ACT network is a well-established national research network embodying established best practices \cite{3}. In contrast, 4CE is an agile decentralized consortium developed at the beginning of the COVID-19 outbreak that has pushed the envelope of new methodological approaches and tools. \cite{4} ACT has
demonstrated the ability to interchange data at scale through mappings to standardized ontologies. [5] The 4CE philosophy is that researchers must stay as close to the data as possible at each site. This way, studies can respond to the actual complexities of the data in implementation. In 4CE, research questions and methodologies are refined by going back to those who know the data best, directly involving the researchers and analysts at each site in network-wide analytics.

This panel is timely and needed because of the growing urgency to utilize the burgeoning wealth of EHR data, especially in response to the public health needs of the ongoing pandemic. This group of panelists have multi-institutional experience with two prominent flagship data networks. Given the national scope of ACT and international scope of 4CE, the proposed panel is well suited for the AMIA members and audience.

Presenters

Naming the Data (Visweswaran)

Dr. Shyam Visweswaran is an Associate Professor of Biomedical Informatics and is the Director of the CTSI’s Biomedical Informatics Core at the University of Pittsburgh in Pennsylvania. He leads the Data Harmonization Workgroup that develops the ontologies for the ACT network. He will discuss biomedical ontologies in i2b2/SHRINE data research networks, with a focus on the ACT network and the enhanced ontology functionality developed for COVID-19 research. Ontologies offer not only a way to query the data but also an information model to organize the data. The i2b2/SHRINE ontologies have a tree-like hierarchical structure in which concepts closer to the root are more general than concepts located near the leaves. The tree-like structure enables the investigator to navigate the concepts and construct succinct queries by using the most general concepts that are applicable. The ACT network rapidly developed and validated a specialized COVID-19 ontology and deployed it on the network, and ACT sites augmented their data to support the ontology. The latest version of the ACT COVID-19 ontology, Version 4, consists of 52,476 codes in the domains of diagnosis, procedures, medications, and laboratory tests. The COVID-19 ontology has several unique features. It has categorized emerging terms from ICD-10, CPT-4, HCPCS, and LOINC terminologies that were introduced in response to SARS-CoV-2. It includes computable phenotypes to characterize the course of illness and outcomes in COVID-19 that include illness severity, respiratory therapy management, and level of care. And it contains harmonized value sets for the growing number of SARS-CoV-2 nucleic acid antigen and antibody tests. [9]

Assessing Completeness by Finding Loyal Patients (Henderson)

Darren W Henderson is a Database Administrator at the University of Kentucky focused on data warehouse development and performance tuning. He collaborated on the development of a tool to identify whether patients are “loyal” - meaning they get most of their care within a healthcare system and are thus likely to have longitudinally complete data. This tool implements a previously-validated algorithm which relies on the presence of proxies for loyalty such as prostate-specific antigen (PSA) tests, pap smears, and recent visits. [7] This algorithm has been shown to enhance the performance of machine learning methods by choosing this “enriched” cohort. Mr. Henderson will discuss the loyalty methodology and related approaches presently being used to enhance data network completeness.

Deep Harmonization Across Sites (Klann)

Dr. Klann is an Assistant Professor of Medicine in the MGH Laboratory of Computer Science and Harvard Medical School. Data harmonization must include data mapping to common terminologies. But even when this is accomplished, coding differences across sites can still abound. There are many LOINC codes for the same test, and if institutions choose different (but equally valid) codes, it could cause phenotyping algorithms to fail. For data to be truly harmonized, the underlying codes in the data must be examined. To ensure network harmonization, it is essential to explore the network-wide distribution of codes used to represent a given concept. Additionally, chart review on a subset of patients must be performed to challenge assumptions and discover intricacies in data encoding practices. Dr. Klann will discuss his work in two related projects: comparing count distributions in the ACT network, and enhancing an international phenotype of COVID-19 severity in the 4CE network. [8]

Impact of Data Completeness on Downstream Machine Learning (Estiri)

Dr. Hossein Estiri is an Assistant Professor of Medicine at MGH Lab of Computer Science and Harvard Medical School. Dr. Estiri develops novel computational phenotyping and predictive algorithms using clinical data. In his
presentation, Dr. Estiri will discuss ways in which longitudinal completeness of patient records in EHRs may impact the performance of machine learning (ML) algorithms measured by discrimination power, reliability, and bias. In binary classification tasks, which are widely exercised in healthcare research, standard ML performance evaluation is often focused on discrimination metrics, such as the confusion matrix, area under the receiver operating and/or precision-recall curves. Reliability of predictions are also evaluated from time to time, especially in the context of predictive models. Dr. Estiri has developed the MLHO pipeline [10], an end-to-end ML pipeline, that outputs a comprehensive set of ML performance metrics, including discrimination and reliability, as well as algorithm-level and patient-level measurements of bias. Dr. Estiri will showcase how different levels of longitudinal data completeness (quantified through the loyalty methodology) can prospectively alter discrimination, reliability, and bias in MLHO’s predictions of mortality, ventilation, and ICU and hospital admission due to COVID-19.

Discussion Questions

1. How can biomedical ontologies aid in data quality? What are the unique features of the COVID-19 ontology that enable phenotyping?
2. Even when sites are working in the same data model, differences in set size, patient population disparities, and hardware/database differences can stymie progress in developing reproducible algorithms. What techniques can mitigate these issues?
3. How can chart review be used to augment the structured data coming from EHRs?
4. How can longitudinal data completeness impact machine learning performance in terms of discrimination, reliability, and bias?

Panel Organizer Statement: All participants have agreed to take part in the panel and discuss the topics as outlined above.

References

Identifying and Addressing the Evolving Privacy and Security Risks Lurking in Translational Science

Luke V. Rasmussen, MS¹, Chris Lunt², Shelley Rusincovitch, MMCi³, Justin B. Starren, MD, PhD, FACMI²
¹Northwestern University Feinberg School of Medicine, Chicago, IL; ²National Institutes of Health, Bethesda, MD; ³Duke University, Durham, NC

Abstract

The tools used to manage and analyze translational research data are evolving rapidly. Security and privacy are critical components of translational science, as researchers and institutions have been entrusted by research participants to both safeguard their data and ensure broader benefit is achieved from its use. Newer tools are increasing the availability and utility of scientific data, however, managing security and privacy risks requires tradeoffs between scientific utility as well as equitable access. New tools and capabilities can shift those tradeoffs in subtle and non-obvious ways. Even with comprehensive policies, there remain lurking risks that institutions may not be aware of. In this panel, we will highlight security and privacy risks that may be overlooked and describe balanced strategies to mitigate them. Panelists will engage the audience to discuss additional risks and considerations that are needed for the larger translational science ecosystem, and how informatics can contribute to the solutions.

Background and General Description

Data is well accepted as the fuel driving health and biomedical research. This makes software the vehicles driven by that fuel. From cloud-based solutions, commercial packages, open-source software to home-grown solutions, there are myriad options for data acquisition, management, analysis, and dissemination. With the use of both electronic health record (EHR) data as well as study-specific data collection, maintaining security and privacy of the data is paramount to maintaining the trust of patients and research participants.

Organizations approach risk management differently for privacy and security. Some enforce strict policies that restrict access to websites, or tightly control who is able to install software (including analytic packages) on their devices. With tighter constraints comes lower risk, but at the expense of the utility and flexibility of those using the software. Even when software is vetted (which can be a lengthy process), there are ongoing risks to security and privacy that may easily go unrecognized.

For example, Jupyter Notebooks are becoming a standard tool, especially with increases in reproducible research and open science. However, it is not always recognized where and how Jupyter notebooks embed data – and how that data may inadvertently be leaked. Likewise, tools that are cloud-based or use cloud-based services may be transmitting data outside of approved data boundaries that an institution has established. These are more benign examples than malicious attacks making cybersecurity headlines but are a realistic threat to research security and privacy that must be taken seriously.

Similarly, software packages may provide increased functionality by sending data through a cloud-hosted service in ways that are not obvious to a user. Furthermore, risk is no longer just about balancing security, privacy, and utility – equitable access is of equal importance (Figure 1). Increasing open science and citizen science means data and tools must be accessible, and that solutions to maintain security and privacy should be easily deployed and not require hefty infrastructure or monetary costs for all users.

In addition, as data science and analytics are suffusing all areas of biomedical research, traditional biomedical research teams are increasingly being supplemented by broader membership, including computational experts from non-biomedical fields, who bring critical analytics skills but who are unaccustomed to working with highly regulated human subjects’ data, and students and trainees who are learning about their responsibilities and procedures to appropriately use and safeguard data. The cultural difference among the different research groups can further complicate the development of uniform policies and practices within a project.
A comprehensive approach to managing risk is needed to support translational research. This includes institutional leaders, large consortia, infrastructure providers, educators, software developers, and individual researchers. From these stakeholders, policy, software tools, education, and awareness are all important to the solution.

In this panel, we convene multiple experts who support translational science activities and have considered the balance of utility, equitable access, security and privacy. Panelists will share their experiences and illuminate classes of risk that they have identified. More importantly, panelists will not just identify concerns, but illustrate how they have approached risk mitigation and developed practices to support such activity. Audience participation will be crucial to share additional experiences, and to help identify current best practices as well as shape a future research agenda.

Panel Importance and Target Audience
With a growing number of headlines detailing cybersecurity incidents such as data leaks and ransomware attacks, there is an increased scrutiny on both security and privacy.\(^1\)\(^-\)\(^3\). However, security and privacy risks exist far beyond these mainstream attacks, and all risks need to be taken seriously. How we react and respond to these risks can in turn impact the ability to conduct research, and so careful planning of risk tradeoff and risk mitigation is needed.

This panel is important as it will illuminate a number of valid security and privacy risks that exist today within the translational research ecosystem, but that institutions and researchers may not be considering. Additionally, it highlights the need to provide equitable access to data and tools, which requires additional considerations in risk analysis.

Given the breadth of experiences of the speakers, this panel will provide insights to attendees of all levels: from researchers wanting to ensure security and privacy in their own work, to organizational leaders who need to develop and enforce security and privacy policies. It will also appeal to developers who create or maintain software to support translational science.

Learning Objectives
By the end of this panel, attendees will be able to:

1. Identify privacy and security risks outside of major cybersecurity incidents that can impact translational science if not addressed.
2. Describe considerations for balancing privacy and security, scientific utility, and equitable access to software and data.
3. Plan educational activities for researchers/informaticians/data scientists that support ongoing application of best practices for security and privacy.

Panelists
Justin Starren – Dr. Starren is the Director of the Center for Biomedical Informatics and Data Science in the Institute for Augmented Intelligence in Medicine, Chief of Health and Biomedical Informatics in the Department of Preventive Medicine, and Professor of Preventive Medicine and Medical Social Sciences at the Northwestern University Feinberg School of Medicine. Dr. Starren directs the Data Management and Biomedical Informatics core of the Successful Clinical Response In Pneumonia Treatment (SCRIPT) systems biology center. In this role, he has personally encountered and addressed many of the challenges described above. He will provide an introduction to the topic and the panelists and will moderate participation between the panel and the audience.

Luke Rasmussen – Mr. Rasmussen is a Clinical Research Associate in the Department of Preventive Medicine at the Northwestern University School of Medicine. Leveraging his experiences in software engineering and architecture as well as data management and reproducible research, he will describe classes of risks that exist in common software platforms and describe how researchers can balance access to tools and data while maintaining utility.

Chris Lunt – Mr. Lunt is the chief technology officer for the All of Us Research Program. Chris has more than 25 years of experience designing web services and other data platforms and has spent the last 20 years working as a technology executive. Previously, he worked as an HHS entrepreneur, helping with the rollout of the Affordable Care Act. Mr. Lunt will describe his work in All of Us considering how to incorporate the balance of equitable access along
with scientific utility, security, and privacy. He will define activities in All of Us that are working towards making privacy preserving technologies more accessible (e.g., homomorphic encryption), as well as how risk assessment and tradeoff decisions are made for the All of Us research infrastructure.

Shelley Rusincovitch – Ms. Rusincovitch is the Associate Director of Informatics for Duke AI Health and Duke Forge at Duke University and is a member of the leadership team for Duke+DataScience (+DS). In this role, she supports infrastructure development in translational research and program development including hands-on experiential learning opportunities for students and other learners to work with data. Ms. Rusincovitch will detail the importance of operational practices in addressing security and privacy risks, and balancing access to real-world data with goals of inclusiveness and fostering diversity of team contributions to research projects.

Statement

I, Luke Rasmussen, confirm that all panelists listed in this proposal have agreed to participate in this panel. Panelists are aware that they must register for the conference and that there are no travel or registration funds available.

References


Utility of the All of Us Researcher Workbench in Educational and Research Settings

Adrienne Roman, PhD¹, Hiral Master, PT, MPH, PhD¹, Jessica Ancker, MPH, PhD¹, Evan Brittain, MD¹, Patrick Wu, PhD²
¹Vanderbilt University Medical Center, Nashville, TN, USA; ²Vanderbilt University School of Medicine, Nashville, TN, USA

Abstract
The NIH’s All of Us Research Program is creating one of the largest, most diverse biomedical datasets accessible through the All of Us Researcher Workbench, a secure Google cloud-based platform where researchers can currently access a wide variety of data (e.g., EHR, survey, wearables, and genomics) from almost 330,000 program participants. The goal of this panel is to showcase the All of Us dataset and Researcher Workbench and provide methods and real-world use cases for leveraging this diverse dataset within educational and research settings.

Learning Objectives
• Become familiarized with the All of Us Research Program and Researcher Workbench
• Learn strategies for incorporating the Researcher Workbench (big data and advanced analytics) into educational courses
• Learn approaches for how to use the Researcher Workbench with trainees to answer clinical research questions
• Gain an awareness of potential challenges and benefits to using the Researcher Workbench, from a student’s perspective.

A general description of the panel and the issue(s) that will be examined and a brief description of each panelist’s presentation
• This panel will explore what the All of Us Research Program and Researcher Workbench is and how it can be applied within educational and research settings to teach techniques using big data and promote advancements in precision medicine.
• Adrienne Roman will provide an overview of the All of Us Research Program and Researcher Workbench.
• Jessica Ancker will provide first-hand experience incorporating the Researcher Workbench into an introductory biomedical informatics course at Vanderbilt University School of Medicine.
• Evan Brittain will discuss different techniques and approaches for introducing and using the Researcher Workbench within a research setting to investigate clinical research questions.
• Patrick Wu will discuss his experience as a student using the Researcher Workbench to answer a research question.

An explanation why the topic of this panel is timely, urgent, needed, or attention grabbing is required with a discussion of anticipated audience
Big data is a growing area in the healthcare field, since it has the potential to answer intricate and clinically relevant questions that can advance the field. The panelists will be talking about ways to incorporate one of the most diverse datasets into educational and research settings. These insights are not only useful to researchers and educators, but also to trainees and students as they will learn about different applications of big data. It will also aid the audience identify new opportunities to advance their research and/or educational training.

A list of discussion questions to enhance audience participation:
1. Are the data free to use within educational settings?
2. Are there any additional resources available to instructors or students who want to use this data within a specific course or for an independent study?
3. Are there step-by-step directions such as you just demonstrated to help someone get started in building the dataset and analyses on your site?
4. Can instructors download data from the platform for use within the classroom?
5. Are there any prerequisites (e.g., coding knowledge) that an instructor be made aware of when they are planning to use this data within a course?
6. Can you collaborate with others from different institutions?

A statement from the panel organizer that all participants have agreed to take part on the panel:
I, Adrienne Roman, as panel organizer can attest that all participants have agreed to take part in this panel.
Machine Learning, Data, Algorithms, and Equity in Biomedicine and Healthcare

Anthony Solomonides, PhD¹, Charisse Madlock Brown, PhD², Naoko Muramatsu, PhD³, Dina Paltoo, PhD⁴, Jim Phuong, PhD⁵, Arash Shaban-Nejad, PhD³, Vignesh Subbian, PhD⁶

¹NorthShore University HealthSystem, Evanston, IL; ²The University of Tennessee Health Science Center, Memphis, TN; ³University of Illinois, Chicago, IL; ⁴National Heart, Lung, and Blood Institute, Bethesda, MD; ⁵UW Medicine IT Services, Seattle, WA; ⁶University of Arizona, Tucson, AZ.

Abstract

AI has made rapid progress in medicine and healthcare but has also come under intense scrutiny. Current debates in AI explore bias and fairness in the commissioning, development and deployment of AI systems, and these questions are particularly acute in biomedicine and healthcare delivery. Many publications have documented the problem, while funding initiatives have addressed the issue directly, focusing attention on ways to mitigate bias in data sets and to introduce methods to promote fairness. Biases have been shown to exist that disadvantage minority racial and ethnic groups, women, sexual identity minorities, and the disabled, sometimes even in the face of explicit attempts to counteract biases. The panel will map aspects of bias mediated by social determinants of health and human cognitive biases, and will explore failures in the preparation of data sets, in algorithm development, and in implementation. Methods to address these problems will be introduced and briefly discussed.

Aims

After participating in this panel, members will be able to,

- identify examples of successes and failures of AI and machine learning (ML) in healthcare;
- describe how biases in data sets can impact the quality and fairness of a ML algorithm;
- analyze a hypothetical scenario for an AI application in healthcare and identify examples of issues that may arise in the development, deployment, and maintenance stages; and
- describe methods to mitigate pre-existing and other biases in AI systems and promote equity in healthcare.

Introduction

The recent successes of Artificial Intelligence (AI)—successes not only in terms of technical results but also in gaining acceptance in practice—have been accompanied by persistent questions about the underlying science, about the well-documented imbalances in terms of racial equity and other forms of bias, and about the willingness of the industry to look at itself and examine its practices in any depth. Obermeyer et al⁷ have shown how using healthcare expenditure as a proxy for health status systematically biases outcomes in favor of the better off: planning based on expenditure means those with more to spend on health get priority in services—surely not the intended effect, especially as its most egregious result is to underplay the needs of traditionally minoritized groups. Even where the explicit goal is fairness, the result may backfire: Vyas et al⁸ show how “race corrections” may work against those they were designed to benefit. Inequitable access to services also biases historical data. When those data are absorbed into a machine learning process without corrections, the effect may be replicated and so appear to be the rational “decision” of an algorithm—an artifact that has no reason to be prejudiced.

Data may be biased historically, as noted, or through poor practices in the process of their acquisition and adaptation for machine learning. Healthcare data often throw up quality issues: they may partially reflect the reason why they were collected in the first place; they may have gaps and be missing certain values, possibly at random or in some systematic way; there are methods to impute missing values, but they are far from infallible; the data may be too large to be transformed uniformly, but they must be presented to a learning algorithm in a consistent format; transformations may also generate errors. These problems arise in what is termed the “pre-processing” phase. The translation of the purpose of the proposed system into a feasible design, including a judgment on the suitability of the data for that purpose, and the development and testing of the desired solution make up the “in-processing” (or design/development) phase and provide another possible source of error and bias. Does the algorithm perform as well on subgroups of the population as it does on the whole, and with a similar degree of confidence? What can be done to address under-representation in existing data sets? Who should be involved in the feature engineering process and in what ways?
What methods are available to challenge the equipoise and stability of the algorithm? Implementation, deployment and “algorithmogivility”—the term for post-deployment monitoring—are part of the third phase, “post-processing.” A different set of issues must be addressed here: Is the system well adapted to its context, e.g., how does the system impact the behavior of its human operators and users? A danger is that a poor fit between humans and machines could reinforce cognitive biases—notably, attribution, availability, or anchoring biases—on the part of the human user, thus reinforcing the machine’s skewed decision-making. Finally, does the system continue to behave in a stable and predictable manner? If it is adaptive, does it exhibit any drift in behavior, and if so, is this evidence of some fundamental instability? The vision in healthcare has often been summarized as the “Learning Health System.” If the system is learning “from experience,” it is inevitably adapting and evolving, so its dependability under changing conditions must be ascertained.

The panel will address these issues from various points of view: bias and fairness, equitable design and engineering, niche categories of AI (e.g., tools for independent living for the elderly or cognitively impaired), principles for fair data and fair development of ML systems, dependable and responsible systems, explanation and transparency in AI, and governance of AI. This panel is sponsored by the Ethical, Legal, and Social Issues Working Group.

Program

The moderator, Dr. Anthony Solomonides, will introduce panel members and the motivation for the panel.

Dr. Dina Paltoo will lead off by introducing the problem of AI and equity as has recently been recognized in multiple publications and is being addressed through a major funding initiative by NIH, the Artificial Intelligence/ Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD). This initiative aims to establish mutually beneficial and trusted partnerships to increase the participation and representation of currently underrepresented researchers and communities in the development of AI/ML models. AIM-AHEAD hopes to build capacity and capability in these communities through coordinated infrastructure, training, and access to data and resources, and aims also to support research questions that can use Electronic Health Record (EHR), Social Determinants of Health (SDoH) and other types of data to redress health disparities and advance health equity.

Dr. Charisse Madlock-Brown will discuss her joint work with Dr. Jim Phuong in ingesting, evaluating, and using (SDoH) data at patient and community levels in the context of the COVID pandemic. Additionally, they will discuss efforts to integrate SDoH data within the National COVID Cohort Collaborative (N3C) EHR warehouse.

Dr. Arash Shaban-Nejad will describe his work on employing tools from Explainable AI, knowledge representation, and semantic web to integrate community-level social determinants of health (socimarkers) with individual-level data (e.g., biomarkers and EHR data) to investigate pathways from socioeconomic and environmental risk factors to health and medical disadvantages. He will demonstrate how semantic explainability supports more trustable inference to answer key questions in population health and medicine, reduce biases and enable more transparent ethical decision-making.

Dr. Vignesh Subbian will discuss methods for the mitigation of biases in data sets and methods for the development of fair algorithms for biomedicine. He will also describe efforts at establishing standards for organizational governance of AI and share lessons from and implications of recent, rapid partnerships to respond to federal initiatives focused on AI and health disparities.

Dr. Naoko Muramatsu will discuss opportunities and challenges in AI applications in healthcare, with a special focus on home-based care for people aging with or into disabilities and their care team. She will bring her methodological and substantive expertise in conducting research on home and community-based services in the US and abroad.

Contributors

The panel was designed to be diverse, both demographically and intellectually.

Charisse Madlock-Brown, PhD, MLS (cmadlock@uthsc.edu), an associate professor in Health Informatics and Information Management at UTHSC, is co-lead for the N3C Social Determinants of Health (SDoH) team. The team’s goal is to identify questions that either validate current research or answer new questions for local policy around COVID-19, the impact of groups experiencing resource challenges, and the impact of the pandemic on inequalities. She has a broad background in health informatics, focusing on data-driven analysis of multimorbidity trends and social determinants of health (SDoH).

Dr. Naoko Muramatsu, PhD, MHSA (naoko@uic.edu) is a professor of Community Health Sciences at the School of Public Health and a Fellow at Institute for Health Research and Policy at University of Illinois Chicago (UIC).
Trained in health services research and sociology, she is passionate about improving the health of diverse aging populations and the quality of long-term care systems through research, interventions, and technology. As PI and Co-I on several universities-, foundation-, NIOSH-, and NIH-funded grants, she has investigated how resources available at personal, community, and policy levels interact with each other to affect the health and well-being of persons aging with or into disabilities, with a special focus on social support and healthcare resources. She is the PI of an NIH-funded randomized controlled trial to test the effectiveness of a gentle physical activity program, led by home care aides for frail older adults in a real-world context of Medicaid-funded home care program.

Dina N. Paltoo, PhD, MPH (dina.paltoo@nih.gov) is the Assistant Director, Scientific Strategy and Innovation in the Immediate Office of the Director of the National Heart, Lung, and Blood Institute (NHLBI), where she serves as a senior advisor to the NHLBI Director and provides leadership and strategic direction to complex scientific initiatives and programs related to the NHLBI mission. Dr. Paltoo moved to NHLBI from the Office of the Director, National Library of Medicine (NLM), where she served as the Assistant Director for Policy Development and led NLM’s policy and legislative activities that promoted responsible stewardship and access to scientific and clinical data and information, and for health IT. She was the Director of the Division of Scientific Data Sharing Policy and the Director of the Genetics, Health, and Society Program within the NIH Director’s Office of Science Policy (OSP). Dr. Paltoo received her BS in Microbiology and PhD in Physiology and Biophysics from Howard University and her MPH from the Johns Hopkins Bloomberg School of Public Health.

Jim Phuong, MPH, PhD (jphuong@uw.edu) is a Research Data Scientist in Biomedical and Health Informatics at the Harborview Injury Prevention and Research Center, located in UW Medicine IT Services. He has collaborated on COVID-19 data engineering, quality improvement, and biomedical data science projects, co-led the N3C Social Determinants of Health (SDOH) Domain Team research efforts and engineered SDOH data extracted, transformed and loaded from electronic health records to expand OMOP Common Data Model’s features and utility. Jim provides informatics collaborative support for the National Trauma research prioritization.

Arash Shaban-Nejad, PhD (ashabann@uthsc.edu) is Director of Population and Precision Health and an Associate Professor in the UTHSC-OAK-Ridge National Lab (ORNL) Center for Biomedical Informatics, and the Department of Pediatrics at the University of Tennessee Health Science Center (UTHSC). Before moving to UTHSC, he was a Postdoctoral Fellow of the McGill Clinical and Health Informatics Group at McGill University. He received his Ph.D. and MSc in Computer Science from Concordia University, Montreal and Master of Public Health (MPH) from the University of California, Berkeley. Additional training was received at the Harvard School of Public Health. His primary research interest is Population Health Intelligence, Clinical and Epidemiologic Surveillance, Explainable AI (XAI) and Big-Data Analytics using tools and techniques from AI, Knowledge Representation, and Semantic Web.

Anthony Solomonides, PhD (tony.solomonides@gmail.com) is Program Director for Outcomes Research and Biomedical Informatics at NorthShore University HealthSystem’s Research Institute. He serves as Site PI for the Institute for Translational Medicine, the UChicago-Rush CTSA. He was Associate Director of the NorthShore Patient Safety and Quality Fellowship. In the wake of the recent pandemic, he has been active in the National COVID Cohort Collaborative (N3C) cultivating methodological issues, social determinants of health, and “long covid.”

Vignesh Subbian, PhD (vsubbian@arizona.edu) is an Assistant Professor of Biomedical Engineering, Systems and Industrial Engineering and Director of Health Data Science and Informatics at the University of Arizona (UArizona). Dr. Subbian leads the Computational Medicine and INformatics (COM-IN) Collaboratory, an engineering-driven, cross-disciplinary biomedical research and training hub at the UArizona. Funded through the NSF and AHRQ, research efforts in the COM-IN Collaboratory leverage systems engineering and computational methods including machine learning for clinical and healthcare applications. Dr. Subbian is the PI on an NSF Smart and Connected Health award to develop advanced computational models and informatics tools for critical care medicine, and also the co-lead for the N3C Critical Care Domain Team.

References
De-Identified Record Linkage: the power of combining PPRL and a Linkage Honest Broker

Umberto Tachinardi, MD, MS¹, Shaun Grannis, MD, PhD¹, Abel Kho, MD, MS², Jasmin Phua³, Ken Gersing, MD⁴

¹Indiana University School of Medicine and Regenstrief Institute – Indianapolis, Indiana; ²Departments of Medicine and Preventive Medicine - Northwestern University Feinberg School of Medicine - Chicago, Illinois; ³Datavant Inc - San Francisco, CA; ⁴National Center for Advancement of Translational Sciences (NCATS) - Bethesda, MD

Abstract

The panel will first introduce the applications of data linkage, the challenges of linking de-identified data and present PPRL (Privacy-Preserving Record Linkage) as a solution to match individuals using tokens generated by a PPRL software. It will be explained why the tokens are secure, unidirectional (no re-identification) and compliant with HIPAA’s “expert determination”. The “matching” component will be discussed, once tokens replace true identifiers in datasets they are compared to other tokens, evaluations methods will be addressed. The role of a token clearing house (the Linkage Honest Broker - LHB), responsible to operate matching and produce linkage assets (approved by governance) that will be used by operators of actual data linkage (i.e. data enclaves). Finally, the panel will use the NIH use-case to show how PPRL can help large research datasets to be linkable using PPRL and an LHB, using the example of the National COVID-19 Cohort Collaborative (N3C).

Introduction

In the context of clinical and translational research, data linkage is a powerful tool to allow researchers to analyze more variables that can help identify factors that influence diseases and health. Merging different datasets through linkage requires appropriate technical methods, governance, and compliance. The advancement of data-intensive analytical methods, like Machine Learning and Artificial Intelligence, depends on the ability to generate or capture novel data (such as traditional Clinical Trials) or re-use and augmentation of existing data. One method to re-use and link existing data while respecting privacy is by linking records using only de-identified data. Privacy-Preserving Records Linkage (PPRL) generates de-identified tokens to replace true identifiers, and then uses these tokens as matching elements.

While tokens do not expose identity per se, the combination of tokens with the data associated with them can produce patterns that can potentially lead to re-identification. This is an example where the addition of an honest broker structure (the Linkage Honest Broker - LHB) plays a role in reducing those risks. In this model, PPRL tokens will only be shared between sites that produced the tokens and the LHB. The data connected to those tokens are labeled using pseudo-identifiers with no relationships with true identifiers or PPRL tokens. The LHB receives the pairs with both tokens and pseudo-ids. The tokens are matched against other tokens residing in the LHB platform. Once linkage assets (crosswalks between pseudo-ids for more than one dataset) are requested, and governance clears the request, the linkage platform (the infrastructure that will operate the physical data linkage) receives the assets from the LHB to perform the actual linkage. The LHB never receives or process actual data, and the linkage platform never receives or process tokens.

The NIH (NCATS) decided to pilot this model (PPRL/LHB), as a way to provide critical linkages (i.e. mortality, imaging) to the National COVID-19 Cohort Collaborative (N3C). The pilot will instruct the NIH of PPRL/LHB advantages and risks and develop know-how on development, testing, evaluating, and implementing these technologies to expand to other datasets.

PPRL (Privacy-Preserving Record Linkage): Real-world examples and lessons from other networks

Dr. Abel Kho is an internationally recognized expert in privacy-preserving record linkage, having published the first large-scale real-world application of this method¹. He will share with Jasmin Phua, from Datavant (a commercial provider of PPRL solutions) the presentation of an outline of the PPRL resources and subsequent applications in various settings. They will also describe some key design options in response to local considerations and challenges.
Specifically, explain how PPRL has been used to link records across States and sectors (e.g., health research and social services) and the impact of shifting patient identity features and data quality and standardization on linkage quality.

**The Honest Linkage Broker Platform**

Dr. Tachinardi, the leading architect and Director of the Honest Linkage Broker platform at the Regenstrief Institute, will present an overview of the LHB modules: ingestion, matching, and PPRL. The ingestion process is where the “transit” tokens (produced by the PPRL software using keys known only to both parties) paired with the respective pseudo-identifiers (pseudo-IDs) are sent by the data contributing sites to the LHB. The LHB will then harmonize the tokens (rehash using an LHB key) and match them with all existing tokens that share the same PPRL/LHB agreement. On the matching process, if duplicates are found (same individuals), a list with those duplicates will be sent to the Collaborative Data Repository (usually an enclave) to process de-duplication. Once PPRL requests are received (based on use-case governance), the LHB produces the crosswalk assets (map tables) containing the matched pseudo-IDs. The asset will then be sent to the linkage platform, where the actual data merger and necessary adjudication will occur. Figure 1 presents a simplified overview of the PPRL/LHB architecture and workflows.

![Figure 1. Overview of the PPRL/LHB model architecture and workflows.](image)

**De-Identified matching methods and evaluation**

Dr. Grannis, a well-known research authority in patient matching for electronic health records systems and Dr. Kho, will present the work they did to define the optimal tokens for N3C as a use-case to demonstrate how tokens are defined and certified. Both presenters will also discuss the many questions they often hear about PPRL matching accuracy, efficiency, and the real impact of those methods over undesired biases.

**The PPRL/LHB pilot for NCATS**

Dr. Gersing, Director of Informatics at NCATS, will describe the N3C governance process being implemented across the 88 institutions supplying data to N3C and the 250 institutions and more than 1600 investigators accessing the 7.2 billion rows of COVID data in N3C. This will include a discussion comprising community guiding principles, data contributor PPRL control, and investigator data access. He will also provide an overview of innovative features of the N3C, like linkage using the PPRL/LHB platform to integrate radiology images from the National Cancer Institute's (NCI) Cancer Imaging Archive (TCIA), NCI Digital Pathology Repository, and mortality data to N3C. He will also
describe early tests of linking de-identified data between N3C, and other NIH-sponsored programs such as the All-of-Us Research Program and National Heart Lung Blood Institute (NHLBI) BioData Catalyst.

References

Panel: Lessons Learned from a National Data to Health Coordinating Center

Adam Wilcox PhD, David Eichmann, PhD, Christopher Chute, MD DrPH, Philip R.O. Payne, PhD, Melissa Haendel, PhD

1Washington University in St. Louis, School of Medicine, St. Louis, MO; 2University of Iowa, Iowa City, IA; 3Johns Hopkins University, Section of Biomedical Informatics and Data Science, Baltimore, MD; 4University of Colorado Anschutz, Denver, CO

Abstract
This panel describes activities of the National Center for Data to Health (CD2), which is currently in its fifth and final year of funding. CD2H was intended to support the informatics community in the Clinical and Translational Science Awards (CTSA) program, in supporting best practices in research informatics, adoption and use of standards for data sharing, collaborative informatics innovations, and informatics and data science education. The panel is divided into the different organizational “cores” of CD2H, where the primary leader of each core can describe the goals, approach, results, and lessons learned. Most importantly, the panelists will reflect on how changes in healthcare, research and informatics over the past 5 years have affected the strategies and approaches of the cores, and CD2H generally. This provides an opportunity with the audience to discuss and share lessons learned in the areas and recommendations for future approaches.

General Description
In 2017, the National Center for Advancing Translational Sciences funded a Data to Health Coordinating Center (CD2H-CC) for the CTSA program. With the original funding lasting 5 years, the intended goal for the center was to “support the activities of the CTSA Program in using data to translate discoveries to health benefit.”(1) NCATS recognized the opportunities created by advances in informatics and analytics to accelerate the use of research and electronic health record (EHR) data for improving health. At the time, specific activities anticipated for the coordinating center were related to supporting a collaborative informatics community in the CTSA program, defining and promoting best practices for data and standards for interoperability, fostering collaborative innovation in informatics, and advancing biomedical research informatics and data science education. The funded National Center for Data to Health (CD2H) has addressed these areas of activity by structuring around core focus areas. These core areas have been in resource discovery, informatics maturity and best practices, next generation data sharing, and tools & cloud infrastructure. The resource discovery core focused on creating and enhancing tools and information resources to support their use among CTSA organizations; the informatics maturity core worked to determine, define, and disseminate best practices in data use and informatics; the next generation data sharing core worked to support open-science approaches while linking CTSA community data sharing with other efforts; and the tools and cloud infrastructure core focused on creating a computing architecture to provide CTSA hubs to support affordable, easy to use, and scalable deployment of tools.

CD2H is now in its final year, which creates an important opportunity to identify challenges and successes, consider lessons learned from coordination activities, reflect on external factors that have influenced the goals of the program, and recommend next steps. This panel includes leaders from the CD2H program of these individual cores, who will describe different efforts to achieve the program goals. Important changes have drastically affected the translational data use landscape during these years; for some core activities these changes have required substantially different approaches, while for other cores the changes have demonstrated some permanence of principles and methods. These differences will be considered among the panelists. Discussions with the audience will provide clarity of the principles and lessons discovered, but also consider what are the current critical needs in using data for translational science and how those needs could be addressed with informatics leaders and other coordinating activities. The topic is timely due to the transition in the CD2H program, as well as critical lessons learned from the COVID pandemic and the CD2H response with the National COVID Cohort Collaborative (N3C).

Panelists
David Eichmann, PhD, is Associate Professor Emeritus at the University of Iowa. Prior to his retirement in the summer of 2021 he served as Director of the School of Library and Information Science at the University of Iowa. He is presently a CD2H PI and co-director of the CD2H Resource Discovery Core, as well as a co-lead of the Portals and Dashboards team in the National COVID-19 Cohort Collaborative (N3C). His research interests involve semantic information extraction, integration and visualization and use of these technologies in support of team science. He will describe the work of the Resource Discovery core in the federation of multiple sources of information relevant to clinical and translational science, the integration of those resources into a common discovery framework, and the implementation of a discovery portal enabling access by the clinical and translational research community.

Adam Wilcox, PhD is Professor in the Informatics Institute (I2) at Washington University in St. Louis, Director of the Center for Applied Clinical Informatics in I2, and a member of the Patient Centered Outcomes Research Institute’s Methodology Committee. His research in informatics has focused on applications of informatics principles to improve care, from developing decision support applications, to designing architectural strategies for health information exchange, to creating population health research databases to study issues of health disparities, to methods of improving the use of data and analytics in healthcare. He leads the Informatics Maturity and Best Practices Core for CD2H, where he focuses on using maturity models to advance applications of research informatics best practices. These activities include improving literacy about maturity models, developing and aggregating models and indices, and establishing processes for leveraging models. He will describe these activities over the course of the grant, from the initial development of the Research Informatics and Open Science Maturity Model (RIOSM), identifying factors related to improved accessibility of maturity models, and helping the CTSA informatics community develop and implement maturity models in multiple domains (social determinants of health, data sharing, data analytics, natural language processing, data quality). He will also describe the current state of maturity model use and development for research informatics, along with suggestions for next steps in their implementation.

Christopher Chute, MD DrPH, is the Bloomberg Distinguished Professor of Health Informatics, Professor of Medicine, Public Health, and Nursing at Johns Hopkins University, Chief Research Information Officer for Johns Hopkins Medicine, and Head of the Section of Biomedical Informatics and Data Science. He has had a deep interest in semantic consistency, harmonized information models, and ontology. His current research focuses on translating basic science information to clinical practice, and how we classify dysfunctional phenotypes (disease). He is presently PI on a spectrum of high-profile informatics grants from NIH spanning translational science, including coPI of the Center for Data to Health (CD2H). He leads the Next Generation Data Sharing workstream in CD2H, and has fostered data harmonization and integration as demonstrated in the National COVID Cohort Collaborative, which he co-leads. His presentation will focus on common data models, data ingestion, quality control, syntactic and semantic mappings. He will emphasize the strong synergies that exist between locally controlled, federated data repositories across organizations with centralized integration models such as N3C.

Philip R.O. Payne, PhD is the Janet and Bernard Becker Professor and Director of the Institute for Informatics (I2) at Washington University in St. Louis. He also serves as the Associate Dean for Health Information and Data Science and Chief Data Scientist for the Washington University School of Medicine. He is the author of over 200 publications focusing on the intersection of biomedical informatics and the clinical and translational science domains. Dr. Payne’s research group currently focuses on: 1) machine learning and cognitive computing approaches to the discovery and analysis of bio-molecular and clinical phenotypes; 2) interventional approaches to the use of electronic health records and clinical decision support systems; and 3) the design and evaluation of open-science platforms that enable collaborative and cumulative approaches to discovery science. He co-leads the Tool and Cloud Infrastructure Community Core of the Center for Data 2 Health (CD2H), where his efforts have focused on the design and deployment of scalable and elastic cloud-computing environments that can accelerate the deployment of software meant to support clinical and translational research, especially where those platforms can facilitate collaboration and data sharing across traditional organizational boundaries. In addition, he has engaged in projects that seek to create inventories of
reusable tools that can be used to assess the quality and fitness for analysis of data sets being collected and shared in the context of multi-site research programs. He will describe the activities of CD2H related to the implementation of such shared cloud computing architectures, collaborative software development and quality assurance capabilities, and cloud-based “sandboxes” intended to provide a shared environment for the comparative evaluation of novel informatics tools and methods.

Melissa Haendel is the Chief Research Informatics Officer and Professor of Biochemistry and Molecular Genetics at the University of Colorado Anschutz Medical Campus; and is the Director and Principal Investigator of CD2H. Her background is in both wet and dry lab translational science, with a focus over the past decade on the development of ontologies, semantic engineering technologies, and open science infrastructure programs. Dr. Haendel’s vision is to weave together healthcare systems, basic science research, and patient generated data through development of data integration technologies and innovative data capture strategies. Dr. Haendel co-leads the Monarch Initiative, an international consortium dedicated to utilizing model organism genotype-phenotype data, deep phenotyping, and graph-based integration techniques to improve rare disease diagnosis. She also co-leads the NCATS Data Translator, which aims to integrate hundreds of data resources for mechanism and drug discovery. Dr. Haendel is the co-lead for the GA4GH Clinical and Phenotypic workstream, where she supports cross-disciplinary international teams, development of standards for clinical genetics in rare disease and cancer, and improving access to data across the world. As the PI and Director of CD2H, Dr. Haendel will moderate the session and provide broad perspective of the overall goals of CD2H and how the cores support this broad vision. As a leader in multiple translational science initiatives, she can also provide perspective on trends in related projects and topics that affect and are affected by data coordinating activities.

Discussion Questions

- Resource Discovery: How do we support the broad spectrum of clinical and translational researchers in their information seeking?
- Informatics Maturity and Best Practices: In what ways do maturity models help organizations improve practices? What do you see as important features of models that improve their utility?
- Next Generation Data Sharing: How can the multiple clinical research common data models co-evolve to mutually enhance data quality, comparability, and consistency?
- Tools & Cloud Infrastructure: How can cloud computing platforms enable rapid-cycle and agile software development processes, especially where such projects involve collaboration between multiple organizations, informaticians, data scientists, software engineers, and end-user communities?
- Overall: In what ways have broad changes to the use and sharing of clinical data for research changed the needs for coordination? What other changes do we expect that would need to be addressed differently with data, which we can help coordinate now?

All participants have agreed to participate on the panel at the AMIA 2022 Joint Summits meeting.

References:

Enabling Re-Use of Computational Phenotyping Algorithms

Laura K. Wiley, PhD1, Luke V. Rasmussen, MS2, Rachel L. Richesson, PhD3
1University of Colorado School of Medicine, Aurora, CO; 2Northwestern University Feinberg School of Medicine, Chicago, IL; 3University of Michigan Medical School, Ann Arbor, MI

Abstract
Identifying patients from the electronic health record (EHR-based computational phenotyping) is notoriously inefficient and costly. Despite these challenges, there are a large number of phenotype algorithms for the same traits that have been published in the literature. This introduces heterogeneity into the research enterprise as these algorithms may be identifying subtly different patient populations. Despite the fact that reusing existing algorithms would improve research efficiency and reproducibility, algorithm re-use remains uncommon. This panel will discuss three critical barriers to algorithm reuse: findability, implementability, and generalizability. Panelists will provide an introduction to each barrier and synthesize active research characterizing these challenges. They will also engage audience members to share their own expertise and experience, catalyzing the development of solutions and best practices to address these barriers.

Organizer & Moderator: Laura K. Wiley, PhD

Background and General Description
A common pain point in translational research is the difficulty in identifying patient cohorts from electronic health records (EHRs). Computational phenotyping algorithms typically rely on using multiple types of EHR data to infer the phenotype of a patient (e.g., through diagnosis or use of medications or procedures for the condition of interest) and can employ any combination of rule-based logic, natural language processing, and/or machine learning methods. Despite the cost and complexity of algorithm development, re-use of phenotyping algorithms has been limited, leading to a proliferation of algorithms developed for common clinical conditions. For example, reviews have identified 66 published algorithms for asthma,1,2, 30 for heart failure,3 and 17 for acute myocardial infarction.4 In addition to the sheer volume of algorithms, it is not established that these algorithms identify clinically equivalent populations despite nominally being developed for the same phenotype.

There are legitimate scientific and practical reasons supporting the use of multiple algorithms for the same trait including identifying specific subpopulations or requiring particular performance characteristics. For example, clinical trials contacting all identified patients need high positive predictive value, whereas those recruiting after record review may prefer a higher sensitivity approach. However, the continued focus on algorithm development vs re-use comes at substantial cost to the research enterprise. In addition to the time and financial impact of these duplicative efforts, it also introduces unnecessary heterogeneity into the scientific process as studies use populations that are not necessarily equivalent or comparable. Increased algorithm re-use would increase the comparability of research as studies use more unified populations.

While there are structural incentives that discourage algorithm re-use that are difficult to change (i.e., the pressure to create and publish novel methods), there are also a number of barriers that can (and should) be solved. In this panel we will address three of these barriers through didactic presentations by panelists:

1) **Algorithm Findability**: Researchers choose phenotyping algorithms based on a variety of factors: the patient population it identifies, performance characteristics, types of data and/or technical resources required for implementation, etc. The heterogeneity in reporting practices and a lack of standardized metadata make it challenging to find and evaluate existing algorithms to determine if they are appropriate for a new use case.5

2) **Algorithm Implementability**: To accurately re-use an existing algorithm, researchers must not only marshal the correct data and technical resources, but they must also faithfully implement the original algorithm logic. This can be challenging due to widespread underspecification and ambiguity of existing algorithm descriptions.6

3) **Algorithm Generalizability**: Another critical factor for researchers to consider is the setting of the original algorithm development and whether the algorithm is likely to have similar performance in their system. Phenotype algorithms developed using skewed data sources (for example, Veterans
Affairs data that skew heavily male) may introduce significant sampling bias that may make it suboptimal for implementation in a system with more equal gender balance. Through open discussion with attendees, this panel will leverage the experiences of informaticians from the past two decades of phenotype development to begin to catalyze solutions to these barriers and develop best practices to guide the broader research community.

**Target Audience**

Building on the popularity of phenotyping-focused panels at previous Informatics Summits, this panel is targeted to a broad range of attendees including:
- Researchers dealing with the real-world challenge of trying to decide which algorithm to use
- Developers of phenotype algorithms
- Experts developing new technologies and standards for phenotype development and sharing
- AMIA members wanting to learn about the latest innovations and challenges in the phenotyping field

**Structure and Audience Participation**

Given the evolving nature of phenotype quality evaluation and re-use best practices, we want to ensure sufficient time for discussion amongst panelists and audience members. Thus we have specifically chosen to include only three panelists. The panel will be structured as follows:
- Introduction to the Topic (5min; Laura Wiley)
- Panel Presentations (20min each; 60min total)
- Panel/Audience Discussion (25min)

In addition to audience questions we propose the following moderator driven discussion questions:
1. What barriers and facilitators to re-use of computable phenotype algorithms have you observed, and what differences in these barriers exist between institution-specific use cases versus larger network research?
2. What drives your decision on how you publish a phenotype algorithm (e.g., narrative description, pseudocode, executable code)?
3. What strategies have you implemented, or are considering to implement, to address biases in computable phenotype algorithms?
4. What innovations do you feel are lacking to promote re-use, adoption, and generalizability of computable phenotype algorithms?

**Learning Objectives**

By the end of this panel attendees will be able to:
1. Identify current technical, social, and cultural barriers to re-use of computable phenotype algorithms.
2. Describe a framework for evaluating the fitness for purpose of a computable phenotype algorithm.
3. Explain the challenges related to phenotype representation and portability.
4. Assess the potential generalizability of a computable phenotype algorithm.

**Panel Participants and Topics for Presentation**

**Algorithm Findability - Rachel Richesson**

Algorithm re-use fundamentally relies on the ability of researchers to 1) find existing algorithms, and 2) determine whether they would be appropriate for their use case. Historically algorithms have been difficult to find as they are only available in the primary literature and often buried in supplemental details. Phenotype libraries have been posed as one solution to the findability challenge, bringing existing algorithms and their implementation details into a single, searchable location. However, libraries only solve the first challenge (finding existing algorithms). Without appropriate metadata to help researchers determine the applicability of an algorithm to their use case, the utility of phenotype libraries will ultimately be limited.

In this presentation Dr. Richesson will discuss: 1) how algorithms can be determined to be fit for purpose; 2) a framework to support phenotype sharing including metadata to describe algorithm purpose, search tools (e.g., phenotype libraries), and socio-technical solutions.
Algorithm Implementability - Luke Rasmussen

Algorithm re-use somewhat obviously relies on a faithful re-implementation of the existing algorithm in a new system. However, this can be difficult as health systems have different data structures, terminologies, and utilization patterns that have to be accounted for during the implementation process. Due to these challenges, the most flexible representations of phenotype algorithms has been with narrative description of the data elements and logic rules. However, when these descriptions are unclear or incomplete researchers must make decisions about how to proceed. Without an understanding of the author’s original intent, these decisions may lead to unintended divergence and identification of a different patient population than desired.

In this presentation Mr. Rasmussen will discuss: 1) common challenges with algorithm portability (i.e., implementing an algorithm at a new site); 2) the types of under-specification, vagueness, and ambiguity observed among algorithms released through the Phenotype KnowledgeBase; 3) alternative methods for phenotype representation and communication of author intent.

Algorithm Generalizability - Laura Wiley

Appropriate re-use of algorithms requires that they are generalizable (i.e., identify the same patient population with similar performance characteristics to the original implementation) in a new context and healthcare setting. There are many components of the original implementation context to consider when assessing generalizability including the data source (was it filtered to particular population), health system (what types of patients and services), population demographics (gender, race/ethnicity, urban/rural, etc). Sampling bias, where certain patient populations are not included in the training data, can lead to poor performance in more diverse populations. Because cohort identification is the first step in any EHR-based analysis it is critical that we ensure equivalent performance for all patients to avoid creating or exacerbating health disparities.

In this presentation, Dr. Wiley will discuss: 1) domains to consider when evaluating the potential generalizability of algorithms; 2) the frequency of published phenotype algorithms that provide sufficient details about the implementation context to evaluate generalizability; 3) examples of biased algorithm performance that could perpetuate or exacerbate health disparities.

Statement

I, Laura Wiley, confirm that all panelists listed in this proposal have agreed to participate in this panel. Panelists are aware that they must register for the conference and that there are no travel or registration funds available.

Works Cited

Applying Natural Language Processing Neural Network Architectures to Augment Appointment Request Review of Self-Referral Patients to an Academic Medical Center

Christopher A. Aakre, MD, MSc
1Mayo Clinic, Rochester, MN

Abstract

Selecting appropriate consultations for self-referred patients to tertiary medical centers is a time and resource intensive task. Deep learning with natural language processing can potentially augment this task and reduce clinician workload. Appointment request forms for 8168 patients self-referred to General Internal Medicine were reviewed and recommended downstream appointments from manual triage were tabulated. This paper describes the development and performance of thirty-nine deep learning algorithms for multi-label text classification: including convolutional neural networks, recurrent neural networks, and pretrained language models with transformer and reformer architectures implemented using Pytorch and trained on a single graphic processing unit. A model with multiple convolutional neural networks with various kernel sizes (1-7 words) and 300 dimensional FastText word embeddings performed best (AUC 0.949, MCC 0.734, F1 0.775). Generally, models with convolutional networks were highest performers. Highly performing models may be candidates for implementation to augment clinician workflow.

Introduction

Patients are frequently referred to academic medical centers for management of complex medical conditions, to get second opinions, and to be evaluated for undiagnosed conditions. Patients can also self-refer to many academic centers. Reviewing self-referred patient appointment requests to coordinate appropriate specialty referrals is a time intensive process involving physician review. Incorrect specialty referral selection at the time of initial review may increase patient costs by necessitating multiple return visits or delay diagnosis and treatment initiation. Correct specialty referral selection at the time of appointment request review can also optimize patient scheduling logistics. Recent advances in natural language processing (NLP) may provide an avenue to automate this time-intensive task.

Text classification is a fundamental task for natural language processing. Patient self-referrals are unstructured text documents that can provide insight into potential diagnoses and symptom severity. Historically, several NLP methods have been utilized for automated text classification in medicine. Many rule-based classification systems have been designed; however, rule development by domain experts is time-consuming and may not generalize well across text input sources.1 Term frequency-inverse document frequency and Latent Dirichlet Allocation are two powerful unsupervised text classification methods which utilize bag-of-words unigram text representations and do not require expert involvement for rule creation. However, bag-of-words approaches are limited in that they require large vocabularies, do not handle misspellings, and assumes word independence; preprocessing can improve performance.4 Word embeddings, such as Word2Vec and GloVe, have accelerated the field by encoding semantic similarity in word vector representations – enabling higher level language features such as analogy.5 However, use of Word2Vec and GloVe are limited with medical domain specific words as they are frequently out-of-vocabulary. FastText word embeddings, by incorporating subword information into word vector representations, can represent out-of-vocabulary words while preserving semantic similarities. Novel state-of-the-art transformer-based language models, such as BERT and GPT-2, incorporate word position, context, and the attention mechanism to develop vector representations of text. These vector representations can then be used similarly to word embeddings in text classification tasks. Developing these state-of-the-art language models require resource-intensive training on large datasets; organizations such as HuggingFace have made these state-of-the-art pre-trained language models available for free – streamlining development of domain-specific natural language classification models by further training with annotated domain-specific data. The goal of this quality improvement project was to evaluate the accuracy of several neural network-based NLP models for augmenting or automating the time-intensive task of specialty referral selection for self-referred patients at our destination academic medical center.

Methods
Appointment request forms (n=8168) for patients seeking medical evaluation at Mayo Clinic through General Internal Medicine from 5/4/2018 - 10/4/2019 were analyzed; start date correlated with the implementation of a new electronic medical record system. Institutional Review Board review (IRB#20-005901) was performed for this quality improvement project. Appointment requests consist of patient-reported symptom narratives and reasons for seeking evaluation. Specialty referral recommendations were collected for all 8168 approved patients and manually reviewed to ensure accuracy of (L=28) possible specialty referrals (including “no consults recommended”). Labels were updated to reflect actual referrals completed after face-to-face visit; overall, about 5-10% of labels were updated. This manual review served as the “gold standard” for this supervised machine learning application. This dataset was split into training (n=6534, 80%), development (n=817, 10%), and test (n=817, 10%) sets. Dataset splits were stratified to keep proportions of specialty labels similar between sets (Table 1).

<table>
<thead>
<tr>
<th>Specialty</th>
<th>Training set (n=6534)</th>
<th>Development set (n=817)</th>
<th>Test set (n=817)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allergy</td>
<td>429 (6.6%)</td>
<td>62 (7.6%)</td>
<td>45 (5.5%)</td>
</tr>
<tr>
<td>Breast</td>
<td>100 (1.5%)</td>
<td>11 (1.3%)</td>
<td>12 (1.5%)</td>
</tr>
<tr>
<td>Cardiology</td>
<td>1483 (22.7%)</td>
<td>168 (20.6%)</td>
<td>176 (21.5%)</td>
</tr>
<tr>
<td>Dermatology</td>
<td>1067 (16.3%)</td>
<td>125 (15.3%)</td>
<td>117 (14.3%)</td>
</tr>
<tr>
<td>Endocrinology</td>
<td>1039 (15.9%)</td>
<td>120 (14.7%)</td>
<td>143 (17.5%)</td>
</tr>
<tr>
<td>Fibromyalgia Clinic</td>
<td>1058 (16.2%)</td>
<td>128 (15.7%)</td>
<td>124 (15.2%)</td>
</tr>
<tr>
<td>Gastroenterology</td>
<td>1743 (26.7%)</td>
<td>230 (28.2%)</td>
<td>226 (27.7%)</td>
</tr>
<tr>
<td>Gynecology</td>
<td>175 (2.7%)</td>
<td>18 (2.2%)</td>
<td>24 (2.9%)</td>
</tr>
<tr>
<td>Hematology</td>
<td>293 (4.5%)</td>
<td>41 (5.0%)</td>
<td>35 (4.3%)</td>
</tr>
<tr>
<td>Infectious Diseases</td>
<td>181 (2.8%)</td>
<td>25 (3.1%)</td>
<td>16 (2.0%)</td>
</tr>
<tr>
<td>Integrative Medicine</td>
<td>49 (0.7%)</td>
<td>6 (0.7%)</td>
<td>1 (0.1%)</td>
</tr>
<tr>
<td>Nephrology</td>
<td>487 (7.5%)</td>
<td>73 (8.9%)</td>
<td>76 (9.3%)</td>
</tr>
<tr>
<td>Neurology</td>
<td>2252 (34.5%)</td>
<td>277 (33.9%)</td>
<td>255 (31.2%)</td>
</tr>
<tr>
<td>Oncology</td>
<td>22 (0.3%)</td>
<td>2 (0.2%)</td>
<td>3 (0.4%)</td>
</tr>
<tr>
<td>Ophthalmology</td>
<td>382 (5.8%)</td>
<td>48 (5.9%)</td>
<td>41 (5.0%)</td>
</tr>
<tr>
<td>Otolaryngology</td>
<td>754 (11.5%)</td>
<td>91 (11.1%)</td>
<td>81 (9.9%)</td>
</tr>
<tr>
<td>Pain Medicine</td>
<td>66 (1.0%)</td>
<td>15 (1.8%)</td>
<td>14 (1.7%)</td>
</tr>
<tr>
<td>Pharmacy Consult</td>
<td>57 (0.9%)</td>
<td>8 (1.0%)</td>
<td>7 (0.9%)</td>
</tr>
<tr>
<td>Physical Medicine</td>
<td>2056 (31.5%)</td>
<td>251 (30.7%)</td>
<td>238 (29.1%)</td>
</tr>
<tr>
<td>POTS Clinic</td>
<td>39 (0.6%)</td>
<td>9 (1.1%)</td>
<td>10 (1.2%)</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>607 (9.3%)</td>
<td>66 (8.1%)</td>
<td>85 (10.4%)</td>
</tr>
<tr>
<td>Pulmonology</td>
<td>828 (12.7%)</td>
<td>92 (11.3%)</td>
<td>104 (12.7%)</td>
</tr>
<tr>
<td>Rheumatology</td>
<td>835 (12.8%)</td>
<td>90 (11.0%)</td>
<td>103 (12.6%)</td>
</tr>
<tr>
<td>Sleep</td>
<td>633 (9.7%)</td>
<td>79 (9.7%)</td>
<td>90 (11.0%)</td>
</tr>
<tr>
<td>Urology</td>
<td>615 (9.4%)</td>
<td>85 (10.4%)</td>
<td>64 (7.8%)</td>
</tr>
<tr>
<td>Vascular</td>
<td>303 (4.6%)</td>
<td>34 (4.2%)</td>
<td>42 (5.1%)</td>
</tr>
<tr>
<td>Women’s Health</td>
<td>99 (1.5%)</td>
<td>9 (1.1%)</td>
<td>11 (1.3%)</td>
</tr>
<tr>
<td>No consults</td>
<td>311 (4.8%)</td>
<td>50 (6.1%)</td>
<td>58 (7.1%)</td>
</tr>
</tbody>
</table>

A maximum of 384 words from each request was allowed for analysis – about 93% of all samples from each dataset split were shorter than this limit. A second similar, but unlabeled dataset from the legacy electronic health record was used for semi-supervised training; this dataset contained 227,672 patient-reported symptom narratives with no diagnosis or referral data.

Thirty-nine deep learning models with different architectures and hyperparameters were built as Pytorch torch.nn.Module subclasses (Table 2). Base model architectures were built from scratch based on the respective whitepaper or constructed from available pre-trained models. The final layers of all models utilized two fully connected layers (sizes=[16*L, L]) separated sequentially by dropout (p=0.3)\(^2\), layer normalization\(^3\), and mish activation.\(^4\) Sigmoid activation was applied to the final layer outputs to obtain the objective: multi-label binary classification. Facebook’s FastText\(^\text{®}\) subword embeddings were used in non-transformer models to account for frequent misspellings encountered in the patient entered forms; the original 300 dimensional word embeddings and a smaller 64 dimensional word embedding generated from principal component analysis were tested. Kim and Zhang model architectures described in Table 2 are one-dimensional convolutional neural network (CNN) models.
The Kim CNN model uses a range of kernel sizes (1 to 7) and number of convolutional filters based on kernel size and multiplier value, and then concatenates the globally max pooled output of each convolutional layer. The Zhang CNN model uses 6 stacked convolutional + pooling layers with kernel sizes of [7,7,3,3,3,3]. Permutations of these models were tested separately and combined with hyperparameters noted in Table 2. Recurrent neural network (RNN) performance was tested using bidirectional Gated Recurrent Units (GRU), or QuasiRNN (RNN + CNN) alone, as dual-layer ELMo architectures, or followed by EXAM layers. The following pre-trained transformer-based models available from HuggingFace’s Transformers’ python module were used: BERT, DistilBERT, ALBERT, GPT2, DistilGPT2, RoBERTa, and DistilRoBERTa.Text was pre-processed in accordance with the respective transformer model requirements using the HuggingFace AutoTokenizer module. Lastly, we implemented the reformer, a newer model architecture seeking to optimize weaknesses of transformer models with longer text and memory use.

The AdaBelief optimizer with sharpness-aware minimization and stochastic weight averaging was used for parameter weight training ($lr=1e-3$). Sharpness-aware minimization is a newer technique for smoothing the loss landscape and has been shown to improve model generalizability. Learning rate was found by experimentation in the range [1e-5, 1e-2]. Adam, Adadelta and stochastic gradient decent optimizers were also tested during hyperparameter search and found to have worse performance on the development dataset. Gradients were accumulated for every 8 samples before gradient updates. Binary cross-entropy, averaged across labels, with label smoothing ($\epsilon=0.1$) was used as the objective function. Label smoothing is a regularization method that prevents overconfidence in label selection and can be useful when ground truth annotations may be imprecise.

Pytorch-lightning 1.4.2 was utilized for abstraction of the model training process. Model training was performed in three steps. First, all models were trained on the supervised training set for up to 1024 epochs, with early stopping callback monitoring development set hamming loss ($\text{patience}=8, \Delta_{\text{min}}=1e-4$). Second, all models were used to create an ensemble model and trained on the training set. The ensemble model was constructed by concatenating each individual model’s predictions from the training dataset as model input, followed directly by a classification head in the same configuration described above. The trained ensemble model was then used to generate pseudo-labeled predictions for the unlabeled dataset. Label predictions from the ensemble model were converted to binary labels; the cutoff threshold for a positive label was 0.5. Validation was performed regarding the “no consultations” label. Before the final training step, all model parameter weights were re-initialized. Final models were trained for up to 1024 epochs on the full manually annotated training set (n=6534) concatenated with a random sample (n=16384) of the pseudo-labeled dataset. The development dataset hamming loss was monitored with early stopping ($\text{patience}=8, \Delta_{\text{min}}=1e-4$). Each model’s best parameter weights on the training dataset were used on the evaluation dataset to calculate the following performance metrics: area-under-the-curve (AUC), hamming loss, precision, recall, F1 score and Matthew’s correlation coefficient. AUC, precision, recall, and F1 score were calculated with macro-averaging. All model training and testing were performed on a single Nvidia GeForce GTX 1080 Ti with 11 GB RAM, python 3.9.6 and Pytorch 1.9.0.

Results
Thirty-nine neural network models were evaluated for this multi-label text classification task of predicting specialty consultation needs from patient generated appointment requests (Table 2). A text example and predictions are shown below:

**Example:** I have inflammation in my wrists, fingers, shoulders, chest, knees, and feet. I am being treated for Rheumatoid arthritis, which I am told I have a very aggressive type, lupus, mixed connective tissue disorder, Sjogren’s, Hashimoto’s – all of which can cause inflammation. I’m not sure what is flaring when, but I have felt a lot of inflammation all over my body in the last two years. I am also incredibly exhausted much of the time due to these conditions.

**Labels:** Fibromyalgia, Rheumatology

Overall, the best performing model used 300-dimensional FastText word embeddings with parallel CNNs of various kernel sizes (AUC 0.949, MCC 0.734, F1 0.775). Generally, the CNN and QRNN models with EXAM layers performed better than the larger and newer transformer-based models for this text classification task. Models using the 300-dimensional FastText embeddings were more accurate than the lower-dimensional word representations.
Table 2. Model architectures and hyperparameters with evaluation metrics from holdout set.

<table>
<thead>
<tr>
<th>Model Arch</th>
<th>Trainable parameters (M)</th>
<th>Filters/layer x 6</th>
<th>Max filter #</th>
<th>Multiplier</th>
<th>Biirectional recurrent units</th>
<th>EXAM units</th>
<th>AUC</th>
<th>MCC</th>
<th>Hamming loss</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>FastText 64D embeddings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhang</td>
<td>99.9</td>
<td>2048</td>
<td></td>
<td></td>
<td>0.704</td>
<td>0.410</td>
<td>0.119</td>
<td>0.484</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kim</td>
<td>5.7</td>
<td>1024</td>
<td>256</td>
<td></td>
<td>0.903</td>
<td>0.648</td>
<td>0.066</td>
<td>0.700</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kim</td>
<td>8.5</td>
<td>1536</td>
<td>384</td>
<td></td>
<td>0.907</td>
<td>0.659</td>
<td>0.058</td>
<td>0.710</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kim+Zhang</td>
<td>15.7</td>
<td>512</td>
<td>1024</td>
<td>256</td>
<td>0.913</td>
<td>0.649</td>
<td>0.060</td>
<td>0.701</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiGRU</td>
<td>2.2</td>
<td>512</td>
<td></td>
<td></td>
<td>0.890</td>
<td>0.625</td>
<td>0.073</td>
<td>0.681</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiGRU-ELMO</td>
<td>7.6</td>
<td>1024</td>
<td>256</td>
<td></td>
<td>0.887</td>
<td>0.644</td>
<td>0.075</td>
<td>0.696</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiGRU-EXAM</td>
<td>2.6</td>
<td>512</td>
<td>512</td>
<td></td>
<td>0.908</td>
<td>0.657</td>
<td>0.072</td>
<td>0.706</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiGRU-EXAM</td>
<td>8.1</td>
<td>1024</td>
<td>512</td>
<td></td>
<td>0.896</td>
<td>0.668</td>
<td>0.069</td>
<td>0.715</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiQRNN</td>
<td>0.7</td>
<td>512</td>
<td></td>
<td></td>
<td>0.519</td>
<td>0.226</td>
<td>0.296</td>
<td>0.300</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiQRNN-ELMO</td>
<td>1.3</td>
<td>1024</td>
<td>256</td>
<td></td>
<td>0.524</td>
<td>0.239</td>
<td>0.270</td>
<td>0.314</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiQRNN-EXAM</td>
<td>1.1</td>
<td>512</td>
<td>512</td>
<td></td>
<td>0.518</td>
<td>0.226</td>
<td>0.299</td>
<td>0.300</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiQRNN-EXAM</td>
<td>1.8</td>
<td>1024</td>
<td>512</td>
<td></td>
<td>0.877</td>
<td>0.600</td>
<td>0.102</td>
<td>0.644</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FastText 300D embeddings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhang</td>
<td>103.3</td>
<td>2048</td>
<td></td>
<td></td>
<td>0.864</td>
<td>0.673</td>
<td>0.067</td>
<td>0.722</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kim</td>
<td>11.8</td>
<td>1024</td>
<td>256</td>
<td></td>
<td>0.942</td>
<td>0.719</td>
<td>0.049</td>
<td>0.760</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kim</td>
<td>17.7</td>
<td>1536</td>
<td>384</td>
<td></td>
<td>0.949</td>
<td>0.734</td>
<td>0.048</td>
<td>0.775</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kim+Zhang</td>
<td>22.7</td>
<td>512</td>
<td>1024</td>
<td>256</td>
<td>0.935</td>
<td>0.735</td>
<td>0.047</td>
<td>0.776</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiGRU</td>
<td>3.0</td>
<td>512</td>
<td></td>
<td></td>
<td>0.909</td>
<td>0.705</td>
<td>0.058</td>
<td>0.749</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiGRU-ELMO</td>
<td>9.1</td>
<td>1024</td>
<td></td>
<td></td>
<td>0.732</td>
<td>0.305</td>
<td>0.270</td>
<td>0.353</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiGRU-EXAM</td>
<td>2.3</td>
<td>256</td>
<td></td>
<td></td>
<td>0.910</td>
<td>0.702</td>
<td>0.060</td>
<td>0.747</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiQRNN-ELMO</td>
<td>3.3</td>
<td>512</td>
<td>512</td>
<td></td>
<td>0.929</td>
<td>0.726</td>
<td>0.052</td>
<td>0.767</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiQRNN-EXAM</td>
<td>9.5</td>
<td>1024</td>
<td>512</td>
<td></td>
<td>0.931</td>
<td>0.725</td>
<td>0.057</td>
<td>0.766</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiQRNN</td>
<td>1.4</td>
<td>512</td>
<td></td>
<td></td>
<td>0.525</td>
<td>0.234</td>
<td>0.268</td>
<td>0.310</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiQRNN-ELMO</td>
<td>1.5</td>
<td>256</td>
<td></td>
<td></td>
<td>0.521</td>
<td>0.220</td>
<td>0.324</td>
<td>0.292</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiQRNN-EXAM</td>
<td>1.8</td>
<td>512</td>
<td>512</td>
<td></td>
<td>0.907</td>
<td>0.692</td>
<td>0.075</td>
<td>0.730</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiQRNN-EXAM</td>
<td>3.2</td>
<td>1024</td>
<td>512</td>
<td></td>
<td>0.910</td>
<td>0.712</td>
<td>0.055</td>
<td>0.754</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transformers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT-base</td>
<td>110.5</td>
<td></td>
<td></td>
<td></td>
<td>0.500</td>
<td>0.012</td>
<td>0.127</td>
<td>0.071</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT-large</td>
<td>336.5</td>
<td></td>
<td></td>
<td></td>
<td>0.504</td>
<td>0.012</td>
<td>0.127</td>
<td>0.071</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DistilBERT</td>
<td>67.4</td>
<td></td>
<td></td>
<td></td>
<td>0.514</td>
<td>0.012</td>
<td>0.127</td>
<td>0.071</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALBERT-base</td>
<td>12.7</td>
<td></td>
<td></td>
<td></td>
<td>0.499</td>
<td>0.226</td>
<td>0.318</td>
<td>0.298</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALBERT-large</td>
<td>19.1</td>
<td></td>
<td></td>
<td></td>
<td>0.499</td>
<td>0.231</td>
<td>0.316</td>
<td>0.300</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPT2-base</td>
<td>125.5</td>
<td></td>
<td></td>
<td></td>
<td>0.720</td>
<td>0.312</td>
<td>0.171</td>
<td>0.391</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DistilGPT2</td>
<td>83.0</td>
<td></td>
<td></td>
<td></td>
<td>0.883</td>
<td>0.621</td>
<td>0.092</td>
<td>0.665</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RoBERTa-base</td>
<td>125.7</td>
<td></td>
<td></td>
<td></td>
<td>0.546</td>
<td>0.012</td>
<td>0.127</td>
<td>0.071</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DistilRoBERTa</td>
<td>83.2</td>
<td></td>
<td></td>
<td></td>
<td>0.527</td>
<td>0.012</td>
<td>0.127</td>
<td>0.071</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biomed-RoBERTa</td>
<td>125.7</td>
<td></td>
<td></td>
<td></td>
<td>0.545</td>
<td>0.012</td>
<td>0.127</td>
<td>0.071</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reformer</td>
<td>4.6</td>
<td></td>
<td></td>
<td></td>
<td>0.577</td>
<td>0.214</td>
<td>0.336</td>
<td>0.285</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Discussion
This paper describes the development and evaluation of multiple deep learning models which suggest specialty consultations from patient-generated symptom descriptions for chronic medical conditions. The goal of this quality improvement project was to create a deep learning algorithm that could augment or automate the time-intensive task of chart and appointment request review currently being performed by General Internal Medicine clinicians. To develop these algorithms, we utilized standardized documentation after implementation of a new electronic medical record to construct a manually labeled dataset from semi-structured data captured during normal clinical workflows. The application of deep NLP algorithms to triage patient-generated text is not completely novel, however previous applications using this patient-generated text have been limited to acute care conditions or patient portal messages.30-34 These prior applications have used rule-based models, bag-of-words...
models, RNNs, CNNs, and transformers with few comparisons of accuracy among model architectures, as performed in this study.

The types of models evaluated in this study were deliberately chosen to assess different types of neural network architectures for this classification task. FastText word embeddings were used as the base word-vector representation for many of the models in this study because of frequent misspellings and medical domain-specific words found in initial review of patient-generated text. SentencePiece and byte pair encodings, utilized by various transformer models, also employ subword representations. 35,36 The size and complexity of the models evaluated varied significantly; for example, some transformer models contained over 100x more parameters than simpler models. Practically, technical debt associated with deploying multiple large models for separate classification tasks can quickly become limiting and may favor smaller models employing pre-trained word embedding inputs; the best performing model had only 17.7M parameters.

Generally, we found that among non-transformer models, CNNs outperformed RNNs, and embedding dimensionality was important for model accuracy. Transformers did not provide significant advantages in this specific text classification task. One explanation for this may be that conceptually, the clinicians performing triage may be recognizing n-gram word combinations similar to the filter and kernel size combinations used in the Kim CNN model architecture. It is not clear why transformer models, which produce state-of-the-art performance on NLP benchmarking tasks, underperformed on this specific task; this finding highlights the importance of trying multiple model architectures to find the best fit for the task.

Significant class imbalance was present in this dataset, with positive labels only representing 9.7% of all labels. Initial training utilizing binary cross-entropy loss resulted in poor discrimination on the development set. Other loss functions with better performance on datasets with significant class imbalance, including focal loss and class-balanced focal loss, were evaluated, but did not generate improved performance (data not shown). 37,38 Another loss function based on Matthew’s correlation coefficient was recently proposed to guide segmentation of images with significant class imbalance; this was also tested and did not improve accuracy (data not shown). 39

In this work, we evaluated the ability of multiple neural network model architectures to accurately predict specialty consultation needs from patient-generated text compared to the current gold-standard human review. The best performing model had an AUC of 0.949 and is a candidate for implementation with the appointment triage workflow to augment clinician effort. Future work includes optimizing prediction accuracy by incorporating data after patient arrival and considering application to predict appointment needs for general outpatient appointment requests.

Conclusion
Augmenting human triage with machine-learning based natural language processing and categorization may improve efficiencies and reduce clinician workload from manual chart review. High-level machine learning frameworks and state-of-the-art pre-trained models enable rapid prototyping of multiple model architectures and facilitate hyperparameter tuning to find the best models for specific machine learning tasks, such as text classification. Although the newer transformer architectures produce state of the art performance in many natural language processing tasks, CNN model architectures had significantly better performance in this specific task.

References
doi:10.1108/eb026526
doi:10.1177/journal.pone.0232525
doi:10.1101/12179-2-40.0006-9
doi:10.3115/v1/d14-1162
7. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of Tricks for Efficient Text Classification. 15th Conf Eur Chapter
Biomed Inform
Natural Language Processing for Enterprise-scale De-identification of Protected Health Information in Clinical Notes

Noor Abu-el-rub, PhD1, Jay Urbain, PhD2, George Kowalski, BS2, Kristen Osinski2, Robert Spaniol, PhD3, Mei Liu, PhD1, Bradley Taylor2, Lemuel R. Waitman, PhD4
1Division of Medical Informatics, Department of Internal Medicine, University of Kansas Medical Center, Kansas City, Kansas; 2Medical College of Wisconsin, Milwaukee, Wisconsin; 3The University of Kansas Hospital, Kansas City, Kansas; 4Division of Health Management and Informatics, University of Missouri, Columbia, Missouri

Abstract
Patient privacy is a major concern when allowing data sharing and the flow of health information. Hence, de-identification and anonymization techniques are used to ensure the protection of patient health information while supporting the secondary uses of data to advance the healthcare system and improve patient outcomes. Several de-identification tools have been developed for free-text, however, this research focuses on developing tools de-identification and adjudication framework that has been tested for i2b2 searches. The aim is to facilitate clinical notes research without an additional HIPAA approval process or consent by a clinician or patient especially for narrative free-text notes such as physician and nursing notes. In this paper, we build a scalable, accurate, and maintainable pipeline for notes de-identification utilizing the natural language processing and REDCap database as a method of adjudication verification. The system is deployed at an enterprise-scale where researchers can search and visualize over 45 million de-identified notes hosted in an i2b2 instance.

Introduction
Electronic health record (EHR) systems provide a promising resource of data to accelerate data-driven solutions and research. In general, there are two forms of data in healthcare systems: structured and unstructured data. Structured data refers to a pre-defined data model with associated value sets which are often stored in a database such as diagnosis codes and patient vital signs (e.g. heart rate). Unstructured data, on the other hand, does not follow a pre-defined format of specific values, it can be found in wide-ranging clinical notes and reports such as physician progress notes, nursing free text patient assessments, procedure and operative reports, and radiology and pathology reports. Structure data’s consistency makes it easier to integrate for research purposes whereas unstructured data requires more preprocessing, normalization and transformation before it can be utilized by researchers. As a result, most of the unstructured data is viewed at point of care by clinicians but often gets unmanaged as an enterprise asset for supporting healthcare system analysis and clinical research.

Clinical notes represent interactions between patients and the healthcare system arising from episodes of patient care in which healthcare providers record the observations, impressions, care or treatment plans, and other activities. Therefore, they contain observational data, family history, and physician interpretations which are key to help better understand patients’ health conditions and predict early diagnosis and treatment. Text notes may include elements recorded in discrete structured format or non-discrete narrative format using different documentation methods ranging from standardized templates that integrate discrete phrases or elements from the EHR to hand-written or dictated notes that later get scanned or transcribed into the EHR. Free-text narrative notes are the traditional method for healthcare professionals to record their practice without limitations. Nursing specialties, such as psychiatric nursing, rely heavily on narrative notes as they allow nurses to pull together events and information in a meaningful way within a subjectively experienced environment, as well as to document time-oriented events. While clinicians generally value flexibility and efficiency, those reusing data often value structure and standardization.

Clinical notes are stored as a free-text format and often contain identifiable or confidential information that poses challenges when attempting to de-identify to the Safe Harbor criteria. While de-identification tools have been developed for free text, not many have been developed or tested to support integration with self-service query tools that predominantly integrate structured data organized by standard ontologies. The University of Kansas Medical Center (KUMC), the Medical College of Wisconsin (MCW), and many other medical centers provide researchers access to de-identified, structured patient data through i2b2 data repository platform. While structured data is de-identified by removing required fields and obscuring dates by date shifting often or reducing temporal resolution, it lacks cohesive representation of patient documentation; specifically, observations and decisions recorded in free text

92
notes. Having these notes de-identified allows researchers to define more robust computable phenotypes versus using structured data alone. Additionally, providing datasets from i2b2 that contain de-identified free text allows us to disclose the minimum information necessary to advance research while protecting patient privacy.

Background and Significance

Information extraction methods were a focus of development mostly between 1987 and 1998 and initially sponsored by the federal agencies outside of the biomedical domain. Information extraction supporting clinical research has been increasingly catalyzed since the National Institutes of Health started the Clinical and Transactional Science Award (CTSA) program in 2007. To extract clinical documentation or clinical text, de-identification step is required to ensure the removal of the 18 HIPAA identifiers described below.

There have been several challenges within the research community that focuses on clinical notes de-identification such as the 2006 i2b2 de-identification challenge, the 2014 i2b2/UTHHealth Shared Task, and the CEGS N-GRID shared task. This led researchers to develop various de-identification tools over the past few years. Most of the de-identification tools reportedly achieve high accuracy, however, their performance significantly drops when applied to real-world datasets. In addition, they often do not meet the scalability requirement for Clinical Integrated Data Repositories (CIDR), and fail to address the following crucial facets: “How can we gain trust in the de-identification process?”, “How can we ensure patient privacy and confidentiality when sharing de-identified clinical notes?”, and “What are the required precautions and procedures an institution must follow to ensure privacy?”. Hence, the goal of this study is to build a reliable framework at an enterprise-level that allows data sharing for de-identified clinical notes across multiple medical centers. This paper describes our approach to clinical notes de-identification across two medical centers to support researchers via self-service i2b2 queries and augmenting research dataset requests with de-identified notes.

Materials and Methodology

**HIPAA Privacy Rule Definition**

The Privacy Rule of The Health Insurance Portability and Accountability Act (HIPAA) requires covered entities’ health information be protected while allowing data sharing and the usage for health information. The HIPAA Privacy Rule was established in 1996 to ensure the proper protection of patient privacy when permitting data sharing and the usage for health information for research purposes. It defines the rules for data disclosure and usages without obtaining patient consent.

According to HIPAA there are two de-identification approaches, the Safe Harbor method and the statistical method, which both involve significant human resources to manually examine EHR content for de-identification. The Safe Harbor method of the HIPAA Privacy Rule can be applied automatically on clinical narrative text using NLP methods, allowing for faster and cheaper de-identification of clinical text. In the Safe Harbor method, there are 18 protected health identifiers (PHI) that should be removed in a document to be considered de-identified (shown in Table 1). The removal of these identifiers may result in information loss; however, it is necessary for supporting data reuse and clinical text mining. The de-identification task is time-consuming and labor-intensive; hence, many tools were developed to automatically remove the PHI instead of relying on humans.

**Table 1. HIPAA Privacy Rule identifiers.**

<table>
<thead>
<tr>
<th>HIPAA identifiers</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Names</td>
<td>10 Account numbers</td>
</tr>
<tr>
<td>2 All geographical address elements smaller than state</td>
<td>11 Certificate numbers</td>
</tr>
<tr>
<td>3 All Dates elements related to the individual (except year)</td>
<td>12 Vehicle serial numbers and identifiers</td>
</tr>
<tr>
<td>4 Phone numbers</td>
<td>13 Device serial numbers and identifiers</td>
</tr>
<tr>
<td>5 Fax numbers</td>
<td>14 Web resource locators (URLs) and links</td>
</tr>
<tr>
<td>6 Email addresses</td>
<td>15 IP addresses</td>
</tr>
<tr>
<td>7 Social security numbers</td>
<td>16 Biometric identifiers (e.g. fingerprint)</td>
</tr>
<tr>
<td>8 Medical Record numbers</td>
<td>17 Full face photographic images</td>
</tr>
<tr>
<td>9 Health plan beneficiary numbers</td>
<td>18 Any unique identifying number, code, or characteristic</td>
</tr>
</tbody>
</table>

**HERON i2b2-based Data Repository**

The Kansas University Hospital uses the Epic EHR where research data is derived mainly from Epic’s Clarity relational database that contains more 7,000 tables with over 60,000 columns. Data from Clarity are extracted,
transformed and loaded (ETL) using Structured Query Language (SQL) and Python programming language into an i2b2 compatible star schema, de-identified, and transferred to separate server to be assessed by researchers using the i2b2 application. Since 2010, the KUMC HERON i2b2-based research repository has had over 70 releases that increase the richness of clinical, sociodemographic, and administrative information such as nursing flowsheets, tumor/cystic fibrosis/trauma/cardiac catheterization registries, patient self-reported findings via the patient portal, and social history data in addition to basic patient demographics, diagnosis, laboratory results, and medication data. Incorporating free text findings from multi-disciplinary flowsheet documentation, clinician notes, and reports remained a major gap in our goal of providing comprehensive clinical data.

The Medical College of Wisconsin (MCW) as well uses Epic for their Electronic Medical Record System and extracts from the Clarity reporting system the necessary tables to support the Clinical Research Data Warehouse, which has a translation layer that supports the (ETL) of de-identified data into an i2b2 star schema. In addition to Epic, MCW also integrates a number of historical single-use systems to augment and provide a richer longitudinal context to their patient population going back to 2004, including the NAACCR tumor registry, IntelliDose for chemo, etc. MCW continues to enrich with other clinical systems on their campus to enrich specific domains such as EKG for cardiovascular, genomics for precision medicine.

**Enterprise-scale de-identification framework**

Figure 1 shows the flow of the de-identification and adjudication process for the purpose of loading data to i2b2 de-identified server for researchers to search and visualize. There are two steps for the system: notes de-identification and notes adjudication. The notes de-identification step removes all identifiable information such as names of patients, locations, identification numbers, dates, and age above 90 in order to meet HIPAA Privacy Rule criteria, it includes text preprocessing, regular expression processing, named entity recognition, date shifting, PHI seeding, blacklisting de-identification, and whitelisting identified information to be retained. During notes adjudication, notes are audited to ensure PHI removal and the note de-identification performance was evaluated before releasing the notes.

![Figure 1. Overview of the enterprise-scale de-identification framework](image)

**Notes De-identification**

1. **Text Preprocessing**

Text in the EHR usually comes from different resources such as dictation software, direct keyboard entry, and templated forms that integrate EHR structure data. Since text is authored in different formats, text pre-processing is
crucial for the de-identification process and can significantly increase data consistency. During preprocessing we normalized character encodings so all notes share the same encode, we corrected broken sentences due to the Epic record boundaries limitations, where one clinical note, as a result, could be stored across multiple database records, therefore, we concatenated broken records into notes and restored missing line breaks. In addition, we cleaned text and removed its natural complexity by stripping some punctuations that could potentially prevent identification of certain PHI (e.g. backslash, underline), and we fixed camel-case text by adding spaces between words concatenated into compound words.

2. **Regular Expression Rules**

A set of regular expressions was developed to de-identify several identifiers including MRN, addresses, emails, phone numbers, and zip codes. We defined 24 different regular expressions to search for patterns where these identifiers are possibly mentioned. Table 2 shows some examples of regular expressions that were used, typically when regular expressions find PHI, it masks the identifier and replaces it based on the identifier type. For example, 07328644 is replaced with [xx-xx-xx-xx-xx] and Justin@gmail.com is replaced with xxx@xxx.xxx.

<table>
<thead>
<tr>
<th>Regular Expression</th>
<th>Example</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ([0-9][1,2])([0-9)[4]) [0-9][4] [0-9][4]</td>
<td>Month/Year</td>
<td>Date</td>
</tr>
<tr>
<td>2 [0-9a-zA-Z+][0-9a-zA-Z+]</td>
<td><a href="mailto:Justin@gmail.com">Justin@gmail.com</a></td>
<td>Email</td>
</tr>
<tr>
<td>3 ([0-9][1,2])([0-9][1,2])</td>
<td>01-23-45-67</td>
<td>MRN</td>
</tr>
<tr>
<td>4 ([0-9][1,2])([0-9][1,2]) ([0-9][1,2])([0-9][1,2])</td>
<td>505-779-4055</td>
<td>Phone number</td>
</tr>
<tr>
<td>5 [d] ([\d\d])</td>
<td>877123</td>
<td>Zip code</td>
</tr>
</tbody>
</table>

3. **Named Entity Recognition**

Named Entity Recognition (NER), also known as entity extraction, is used to identify entities in the text and classify them into predefined classes such as person name, location, and organization. There are many well-known NER tools like NLTK\(^6\), Stanford CoreNLP\(^7\), and SpaCy\(^8\). Most of these tools are pre-trained models that used machine learning algorithms. We evaluated named entity recognition software and several training models from the Stanford NLP group\(^7\), Apache OpenNLP\(^9\), and the MITRE Identification Scrubber Toolkit MIST. A combination of three named entity recognition models developed by Stanford consistently outperformed OpenNLP and MIST\(^20\). The named entity recognition software and models from the Stanford NLP group achieved “out of the box” named entity performance of 92.6% on a test set of patient records.

When using named entity recognition in our NLP pipeline, named entity recognition increased the de-identification performance. This shows the importance of using NER to detect named entities that regular expressions failed to identify. Figure 2 illustrates an example of using the named entity recognition module on notes.

![Figure 2](image1.png)

**Figure 2.** Example of notes after applying NER.

4. **Date Shifting**

Dates related to patient record events such as surgery date and birth date are shifted for anonymization, but the relation between patient record dates is preserved by assigning a random offset to each patient. Date shifting is an important feature for notes usability, since it preserves the chronological order of events, allowing users to better understand
patient history and timeline. Therefore, when regular expressions component finds a date pattern, instead of masking it with its type (e.g. [DATE]), the algorithm adds a random offset to all dates within the same note as shown in Figure 3 with an example offset set to 10.

![Figure 3. Example of notes after applying date shifting.](image)

5. **PHI Seeding**

Typically, EHR data contains clinical notes, patient demographics, encounter information which are stored in the same database, since the de-identification process removes patient information from clinical notes, this information can be easily obtained from the database, therefore we utilize the EHR data and feed patient information to the tool to improve the performance. In addition to using existing PHI, we also defined new PHI’s, for example, the age of the patient at the time of the visit is considered PHI if the age is above 89. While this information might not be available in some EHR data, it could be calculated using existing PHI like patient birth date and visit date. A list of PHI and new PHI was created for each patient, then we used regular expressions to remove any occurrences of these PHIs in the clinical text notes. This component can be enabled and disabled depending on whether there is access to the EHR data.

6. **Blacklist and Whitelist Matching**

A Blacklist and a whitelist are used to handle automatic de-identification and identification of difficult terms that regular expressions and NER tool are unable to identify. Based on domain experts, we define blacklist terms that the algorithm failed to de-identify, these terms are names that are mislabeled by the NER algorithm due to the lack of such terms in the training data of the model (e.g. Wauwatosa, Shankar), or geographic locations such as county and city name that NER is unable to remove. To prevent the tool from over-scrubbing non-PHI terms, we create a whitelist to prevent the de-identification of terms that can be easily mistaken for PHI. For example, ED, ADVAIR, and Sarcoidosis are medical terms that are added the whitelist since they are mislabeled as entity names by the NER tool.

**Notes Adjudication and Evaluation**

**Initial Evaluation**

In general, there are multiple methods to verify the performance of clinical notes de-identification. First, standard NLP metrics can be used to measure performance such as recall (sensitivity), precision (positive predictive value) and F value which is the harmonic mean of recall and precision\(^26\). Second, evaluation can be done to compare the performance of the automated systems with that of humans. For example, a study conducted by the Children's Hospital Medical Center, Cincinnati, Ohio, USA compared the performance of the automated systems with humans using two native English speakers\(^5\). Third, evaluation can be achieved using one de-identification system against another standard system such as the performance of the MITRE Identification Scrubber Toolkit (MIST)\(^31\). For this work, we reported the recall and precision metrics to evaluate de-identification performance on the record-level; if at least one identifier is not completely de-identified then the record is not properly de-identified, and if at least one token is mistakenly de-identified then the record is considered over-scrubbed.

Our de-identification method was initially tested at MCW, by creating a stratified 48,000 patient record test set containing 22 types of patient records drawn randomly from 48M patient records within the CRDW. This full 48,000 record set was used for performance evaluation. We randomly sampled 1,000 records from the 48,000-record test set to report the accuracy. On a single PC (Intel Core 2 Quad CPU @3GHz and 8G Memory) the software de-identified 110 records/sec. We identified 27 errors in a second pass evaluation after correcting errors in the first pass of 1,000 patient records, most of the errors were software failure to remove some parts of patient names (first, last, middle, and initials) and some of the regular expressions generated false positives. Therefore, the system was improved by adding the PHI seeding component that allowed identification of patient names and other related health information stored in the Epic system, where these identifiers get scrubbed if they appear in the notes. In addition, another challenge we...
faced was over-scrubbing some terms (e.g. Drug names and procedures names), we addressed the issue by expanding the whitelist terms and creating several JUnit tests to run regression validation for the regular expressions.

After this initial testing, several validation audits on different de-identified note types were performed to test if patient names were successfully scrubbed. We found 55 patient name leaks in a sample of 11,367 Discharge Summaries (0.48%), 45 PHI leaks in a sample of 2,000 Echo Notes (2.25%), and 87 PHI leaks in a sample of 5,000 Therapy Notes (1.74%), in which most of these identified PHI are patient nicknames. This audit prompted us to implement a new leak mitigation strategy, which blacklists the patient preferred names using Epic Patient tables since patients could be referred by their nicknames instead of their actual names.

**Enterprise-scale Notes Adjudication**

To be trusted, the automatic de-identification tool provided “acceptable” accuracy, but determining the acceptable performance can vary depending on many factors such as the final purpose of the de-identified documents, the legal agreements that could be imposed to avoid re-identification, and the fact that some PHI categories are more sensitive than others. The KUMC Medical Informatics team worked with the KU health system privacy team to determine standards on an acceptable number of notes to review and follow up with another review process for false negatives. They also worked on setting up guidelines for reviewers and annotators to follow to eliminate any bias and standardize the adjudication and annotation process, and provided the reviewers with proper training.

![Diagram](image)

**Figure 4. De-identification adjudication process.**

Figure 4 describes the process flow of the adjudication process for the purpose of loading de-identified notes to i2b2 de-identified server for researchers to search. Once notes are de-identified, 30 randomly selected notes per note type or measure is selected for review. If a note or a measure is found with patient identifiers (i.e. false negative), it requires modification of the de-identification tool to address these PHI and additional analysis of 30 randomly selected notes or measures before approval and discussion with the Data Request Oversight Committee (DROC) formed by representatives from the KU health system and university. Once reviewers from both entities indicate approval for release within the REDCap project, the de-identified text and corresponding facts will automatically be incorporated into the KUMC clinical data warehouse, HERON, for the next release. If the research impact is insignificant then the note type is blacklisted (e.g. a nursing free text field to record a nurse practitioner’s paper).

REDCap is used to validate and approve the fitness of the de-identification on the various types of unstructured data that integrate de-identified notes and personal health identifiers from i2b2 with auditing tools to support adjudication in REDCap. First, we utilized de-identification software using NLP methods to de-identify clinical free text. Second, we used REDCap as a method to store de-identified data and hold the audit files to track the review process. We created two complementary REDCap projects with the first project for holding all audit files and the second for
managing the review process for dual adjudication and commentary in which it documents the review process and all related information including the number of notes reviewed, the number of notes not properly de-identified (false negatives), the number of notes accurately de-identified (all positives), the number of notes falsely de-identified (false positives), risk level assessment, the decision of release approval and any additional comments by the reviewers. As a result, notes types approved by both reviewers will be loaded to i2b2 de-identified server for researchers to search. Figure 5 shows the REDCap de-identification review and adjudication tool for note-by-note type.

![Figure 5. De-identification review and adjudication tool for note-by-note type hosted in REDCap.](image)

**Final Evaluation**

We first evaluated nursing flowsheets using the proposed adjudication process. We evaluated 4010 different flowsheet types and randomly sampled 30 notes per type. After running the de-identification tool on this dataset, two annotators reviewed a total of 48,377 notes and found the reported average precision was 74.8% and the average recall was 84.5% at record-level. The Inter-rater agreement between the two specialists was K=90.8% which considered almost perfect agreement. As a result, 3,254 flowsheets types were approved and released in HERON, while 756 types were rejected due to performance results and their insignificant impact on the research community.

Throughout the flowsheet reviewing process, we used the reported errors to enhance the system, thus, we updated the regular expressions to handle other types of date formats (e.g. MM/DD) and phone numbers that are written without dashes. Also, we observed that both backlist and whitelist could be expanded based on domain knowledge of the institution. For example, adding local counties and cities to the blacklist increased the overall performance. Therefore, we added more terms to the blacklist of KUMC which resulted in an expanded list of 1857 terms. We used Junit for automated regression testing to ensure our tool successfully remove previously tested PHI after incorporating new changes.

Consequently, we gained more confidence in the de-identification process after several rounds of flowsheet audits and tool modifications that improved the overall de-identification performance. We then reviewed physician notes where we audited 100 clinical notes (240 records; notes in Epic are stored in multiple records due to 4000-character column size restriction) from 44 different note types including progress notes, discharge summary, and H&P notes. 4 errors were identified in 240 records achieving 98.1% recall and 57% precision at record-level. To investigate the low precision, we computed precision and recall at instance-level (tokens are classified to PHI or non-PHI) and found the tool achieved 99.7% recall and 94% precision at instance-level. Reporting precision and recall metrics at instance-level is more accurate and interpretable than record-level since record-level evaluated record instead of tokens; a
record is labeled as false negative if at least one PHI token was not de-identified and labeled as false positive if at least one token is mistakenly de-identified. However, for enterprise-evaluation we used record-level since evaluating using instance-level is inefficient and time-consuming when reviewing thousands of notes.

At enterprise scale across all note types at KUMC, our pipeline de-identified 90,881,400 records in 4.8 days with a processing rate of 218 records/second, and we were able to release millions of de-identified notes which included: 112 types of physician notes (34 million), 4010 types of flowsheets (6 million), pathology reports (380 thousand), radiology exams (2.2 million), and radiology impressions (2.2 million) into HERON for i2b2 free text search and de-identified data delivery. This experiment was conducted on 32 threads using a JVM limited to 4 GB on LINUX for Intel Xeon CPU ES-4617 2.9 GHz cores. Initially, performance degraded with extensive whitelist and blacklist additions, but the method was enhanced by utilizing hashing techniques when searching backlist and whitelist terms for efficient search. Also, we modified the tool to run in a multithreading manner to allow multiple notes to be processed simultaneously to enhance the tool scalability.

**Discussion**

In this paper, we introduced an enterprise-scale framework for clinical note de-identification and adjudication. Our note de-identification process is a hybrid approach of NLP techniques, rule-based taggers, dictionaries, and domain knowledge. It follows a modular design system where modules are independently created and can be easily modified and replaced to facilitate software adoption and reusability, for example, the named entity module can be replaced, whitelist and blacklist can be easily updated based on the institution needs, and additional regular expressions can be added. We complemented technical advances with an adjudication method using REDCap for tracking and auditing the de-identified notes review process with health system privacy officials. These combined advances allow the release of notes for i2b2 searches and allowing clinical researchers to visualize notes in the i2b2 timeline plug-in. Figure 6 shows an example of an i2b2 query incorporating free text search for “AD-PKD” (Autosomal dominant polycystic kidney disease) in the discharge summaries; allowing clinical researchers to review the richer context provided by narrative.

The de-identification system has been deployed at two medical institutions (KUMC and MCW) after an intensive evaluation using different datasets of various note types and it is made public ([https://bitbucket.org/MCW_BMI/notes-deidentification](https://bitbucket.org/MCW_BMI/notes-deidentification)) to further collaborations.

Similar to continuous quality improvement cycles, several experiments were performed to evaluate the tool and enhance system performance. At first, an initial test was conducted on 1000 notes stratified from 48,000 patient records of 22 types, 27 errors were identified. In addition, we evaluated the performance of patient name removal by running the tool on a dataset that consisted of 11,367 Discharge Summaries, 2,000 Echo Notes, and 5,000 Therapy Notes (1.74%) and found 187 patients that were called by their preferred names (i.e. nickname) instead of their actual names which the tool failed to identify. This audit led to an improved new leak mitigation strategy leveraging the patient preferred names from one of the EHR Patient tables to customize blacklisting.

Notes adjudication is an important component to the overall de-identification process, it promotes trust, transparency, and collaboration with the privacy office, and allows institution to formulate an acceptable threshold for errors and prioritize improvements in the de-identification process. Simultaneously, it is very time consuming. Using REDCap in our work helped tracking the review process and generated reports of the review outcomes such as the approved note types, comments of rejected note types, and disagreement between the two reviewers.

Following the notes adjudication process, we conducted an evaluation test on a large dataset of nursing flowsheets, we considered 4010 flowsheet types where we applied the adjudication process for each type. Two annotators reviewed 48,377 random notes drawn from various note types. The system achieved an average precision of 74.8% and a recall of 84.5% (record-level). Throughout this process, we used the reported errors to enhance the system by adding more regular expressions and expanding the backlist and whitelist terms based on the domain knowledge of the institution. Also, we utilized hashing techniques when searching backlist and whitelist terms as well as modifying the tool to run in a multithreading manner to enable better scalability. As a result, 3,254 flowsheets types were approved and released in HERON, while 756 types were rejected due to performance results and their insignificant impact on the research community (e.g. a text field used to store the pager number of clinical team members).

The several rounds of flowsheet audits and system modifications increased confidence in the overall de-identification process and engagement with the health system privacy office which was critical prior to the release of de-identified free text data and incorporate other types of notes as well. A key safeguard is the structure of our data sharing and data use agreements. While not requiring institutional review board approval, data requests are reviewed, approved,
require the recipient of the de-identified data safeguard the received data as if it was a limited dataset, and notify the medical informatics team if any PHI is detected. After flowsheets, we reviewed physician notes and audited 100 clinical notes from 44 different note types. A total of 4 errors were identified in the 240 records and achieved 98.1% recall (record-level). The result shows that adjudicating physician notes after nursing flowsheets helped achieve high precision and recall since several improvements were incorporated in the de-identification tool during the review process. However, our initial focus on performance assessment was on de-identification and future work, and continuous improvements are needed to refine and expand whitelists where our method is erroneously scrubbing medical terminology.

There were several limitations encountered during this study. First, a single note in the Epic EHR may be stored across multiple database records due to a field size limitation of 4000 characters. This affected note visualization and readability when using i2b2 timeline, so methods were altered to store merged records into a single Character Large Object (CLOB) field instead of varchar. Second, the default Epic notes terminology is defined using local or inherited non-standard terminologies; with nearly 50% of clinical notes assigned to a generic “progress note” category. We plan to deploy a standard ontology such as LOINC to improve the i2b2 text search time and facilitate clinical notes aggregation and exchange across research consortia.

Figure 6. i2b2 query using text search on de-identified notes.

Conclusion

In this paper, we introduced a framework for note de-identification that can accurately de-identify free-text EHR notes while preserving clinical content and described our notes adjudication process using REDCap project. Our experimental results demonstrated high performance and scalability of the proposed note de-identification framework. The system was tested at two medical institutions and currently deployed for i2b2 text search and data requests. We are continuously releasing new note types by prioritizing note types based on their significance to the research community.

Acknowledgment

This work was supported by a CTSA grant from NCATS awarded to the University of Kansas for Frontiers: University of Kansas Clinical and Translational Science Institute (UL1TR002366) and the Medical College of Wisconsin Clinical and Translational Science Institute (UL1TR001436). The contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH or NCATS.
References

Framework for Optimizing Air Ambulance Locations

Sabin Ahmed, Ph.D.1, Robert D. Lieberthal, Ph.D.2, David M. Hechtman1, B.A., Lauren A. Rayson, M.S.1, Douglas R. Amirault2,3, Susan Haas, M.D., M.Sc.2
1The MITRE Corporation, McLean, VA; 2The MITRE Corporation, Bedford, MA; 3Bates College, Lewiston, ME

Abstract

Air ambulances can provide more rapid access to medical care than ground ambulances for rural, underserved, and hard to reach populations. However, the existing allocation of ambulance bases across metropolitan and rural areas is driven primarily by individual operator decisions rather than a health outcomes-based approach. This paper describes a framework for optimizing air ambulance services delivery based on healthcare demand and locational constraints to other modes of transportation. In particular, the paper highlights the need for combining data and how data can be used to identify locations where air ambulance services could be located based on impact. We utilize an information systems approach, applying linear programming models to identify the optimal base locations at the state and regional level. Two data driven use cases for the state of Virginia and New England demonstrate the application of our approach and underscore the importance of data interoperability in health transportation planning.

Introduction

Air ambulances serve over half a million patients a year and have gained greater importance in the healthcare landscape as rural hospitals continue to close.1,2 Additionally, the current Covid-19 pandemic has highlighted the need for rapid transportation of patients to medical care. In some rural areas, air ambulances are the only means of getting patients to trauma centers and other healthcare facilities within the “Golden Hour” after an event.3 Helicopters, which account for over 70 percent of air ambulances, provide emergency scene response and interfacility transfers while fixed wing aircraft typically provide longer distance airport-to-airport transport.4

Concerns about equitable access to and high costs of air ambulance services have plagued emergency medical transportation providers for decades. Research shows that the existing allocation of ambulance bases across both metropolitan and rural areas is driven primarily by industry strategic decision-making and profit motive rather than maximizing the performance of the health-transportation system. Lack of competition is a particular problem, as two private equity firms operating helicopter air ambulances accounted for 64% of the Medicare market in 2017.4 In addition, this misalignment of incentives—contributing to suboptimal health outcomes, sky-rocketing costs for services, surprise medical bills, and documented disparities in access across urban and rural locations—motivates the need for a comprehensive approach to how this service is allocated.

The Government Accountability Office (GAO) found that, in 2016, Medicare spent around $0.5 billion dollars on air ambulance payments.5 In 2017, the median price charged nationally by air ambulance providers was $36,400 for helicopter transports and higher for other aircraft.6 There are also reports of disparities in access, and of underserved areas.7 Policy makers may find that disparities in access to air medical services are closely tied to other issues that make it difficult to address health outcomes directly. Provider costs are high and the increase in air ambulance fees in the National Fee Schedule is reported to contribute to high patient costs.8 Moreover, the Airline Deregulation Act of 1978 limits federal and state legislation from regulating services, fees, and routes.9 Until the recent passage of the 2020 COVID-19 Economic Relief Bill, there had been very limited regulatory or legislative action to address the cost of air medical services. While the recent legislation protects consumers from surprise bills when they seek emergency care and are transported by an air ambulance, it does not address continued issues such as misallocation of bases, redundant supply, high overall cost of care, and lack of coverage for non-emergency care.9

Given this regulatory gridlock and complexity of the air ambulance industry, part of the solution to improving air ambulance related health outcomes could be to incentivize ambulance providers to establish bases near areas where there is a medical need for rapid transport and constraints, or large distances required for accessing ground ambulances. However, there is not currently a data-driven, health-outcome-based framework for identifying the optimal air ambulance allocation in the U.S. Such a framework could help drive a more effective conversation around local, state, and national policy mechanisms to achieve the desired allocation, and provide decision-makers with tools to inform policy decisions and incentives to improve air ambulance access, health outcomes, and lower cost.
In this paper we demonstrate a framework for optimizing the delivery of air ambulance services based on the healthcare demand and locational constraints to other modes of transportation. Identification of the optimal location of air ambulance bases is based on a wide variety of healthcare and transportation data, transportation demand modeling, geospatial analysis, and linear programming optimization models. Specifically, we use detailed event-level emergency medical services (EMS) data from the National Emergency Medical Services Information System (NEMSIS) to identify the health and demographic drivers of air ambulance usage. The information on demand drivers is combined with ZIP code-level data on chronic disease risk factors, health outcomes, and clinical preventive services from the Centers for Disease Control and Prevention (CDC) to identify locations that are at relatively higher risk of requiring emergency transportation to reach healthcare facilities. This information is used to parameterize two linear programming models to optimize allocation of air ambulance bases.

Methodology

Base Locations

We identify the optimal location of air ambulance bases using an analysis of candidate air ambulance patients, as well as potential “supply nodes,” i.e., candidate locations for air ambulance bases. Candidate patients come from event-level emergency medical services (EMS) data to explore the key healthcare and demographic drivers of air ambulance usage, including but not limited to heart attacks, strokes, and various levels of trauma. Subsequently, we use disease incidence and prevalence data to identify the populations that are at risk of these health events. Candidate air ambulance bases are based on the 3,436 hospital service areas (HSAs) identified by the Dartmouth Atlas Project. Each hospital service area is a collection of ZIP codes containing one or more hospitals—with its exact boundaries defined by Medicare patient data.

We use the Maximal Covering Location Problem (MCLP) and the Maximal Expected Covering Location Problem (MEXCLP) to maximize the number of covered patients for a given number of air ambulance facilities. These models locate a limited number of bases in a configuration that maximizes the number of total covered patients. Within each local market, air ambulance providers are often flexible about exact base locations—in many cases opening, closing, and relocating bases over the long run. Proximity to patients is one consideration, and some providers open satellite bases as exurban areas grow—but cost (e.g., potential co-location or consolidation with other medical, public safety, local government, or aviation service facilities), local real estate conditions, and other operational needs (additional hangar space or better sleeping quarters) may dictate exact base locations. MCLP and MEXCLP allow for the application of an iterative modeling approach that can incorporate additional data and features to increase the sophistication, and fit, of the model to actual transportation practice.

The modeling approach includes assumptions about the service area and availability of air ambulances at a rotor-wing base, to parameterize two linear programming models to optimize allocation of air ambulance bases. The MCLP model assumes that there are always aircraft and crew available to respond to potential calls. The MEXCLP relaxes this assumption, allowing for redundant coverage in the event of busy aircraft and crew, mechanical failure, illness, or inadequate staffing. Within each hospital service area, we make the simplifying assumption that the geographic centroid of the HSA will be the only location considered for a potential air ambulance base. We make this assumption because it would be impractical to consider all existing, planned, or potential airports and medical facilities. We believe this simplifying assumption is realistic in most cases, as air ambulance bases are typically located near hospitals.

Conditions of Interest

In this paper we focus on heart attacks and strokes as key individual-level drivers of air ambulance usage. Subsequently, we identify population-level characteristics that contribute to high prevalence rates and incidence of these individual-level health risks. The Centers of Disease Control and Prevention (CDC) PLACES Survey (formerly known as the 500 Cities Project) provides detailed, zip-code- and population-level data on health outcomes, prevention, and unhealthy behaviors. Specifically, we use coronary heart disease among adults aged 18 years and above and stroke among adults aged 18 years and above as the key population-level determinants of air ambulance demand. This information is used to identify the number of potential candidates for air ambulance transports in each zip-code that are at risk of heart attacks and strokes. The fact that heart attack and stroke are incidence variables while cardiovascular disease (CVD) is a prevalence variable allowed us to use two types of burden of disease variables as example applications in our model. The use of these two conditions was used to implement versions of our model with the publicly available versions of the NEMSIS data we had available (see the Conclusion section for additional information about the limitations and generalizability of this approach).
Moreover, we identify utilization of air ambulance services through two additional data sets: CMS Medicare Part B Public Use Files (PUF) and an integrated electronic health records (EHR) / claims data set from Clarivate (previously Decision Resources Group a.k.a. DRG). The Medicare Part B PUF files show all utilization of air ambulance services provided for traditional Medicare beneficiaries under Medicare Part B. This data is aggregated at the provider (air ambulance company) level for 2012 – 2018. Air ambulance data includes the location of the service provider, which we used to determine where bases were located. The DRG data contains claims and EHR data for the state of MA for the years 2016-2019. This data includes claim level (patient charges) for 4 current procedure terminology (CPT) codes related to air ambulance utilization. These codes relate to all conditions, and were not limited to the two conditions, stroke and heart attack, used for the NEMSIS analysis.

**Locales**

In this paper, we show results for two data driven use cases at the state and regional level for Virginia and New England, respectively. We model three “busy fractions” using MEXCLP, p=0.0 (a special case that is equivalent to MCLP), 0.2, and 0.4. Assumptions of MEXCLP include a consistent busy fraction across bases, which is a simplification but certainly a more realistic one than the MCLP assumption that the air ambulances are always available. The higher the busy fraction, the more redundant coverage is prioritized by the MEXCLP approach.

Both the MCLP and MEXCLP models utilize Allagash\(^\text{13}\), an open-source spatial optimization library written in Python. The most recent version of Allagash only supports the Location Set Covering Problem (LSCP) and MCLP models. As such, we extend Allagash to support MEXCLP and test our implementation using a small-scale example from the original MEXCLP paper. Allagash then uses the linear programming package PuLP\(^\text{14}\) as an interface to the COIN-OR Branch-and-Cut (CBC) solver\(^\text{15}\). We retrieve the tile maps from the internet using the contextily package\(^\text{16}\) and plot the actual and modeled air ambulance base locations on these maps using GeoPandas GeoDataFrames (also known as geopandas.GeoDataFrame)\(^\text{17}\).

**Data**

Our technical approach to understand demand, costs, utilization, and other aspects of the ambulance market required understanding the overall air ambulance landscape within the U.S. We detail the multiple data sources needed to perform our analysis in this section. Table 1 summarizes the data sources utilized in this analysis.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Data Source</th>
<th>Data Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRG2 MA data</td>
<td>Decision Resources Group Agency for Healthcare Research and Quality</td>
<td>Clinical and claims data for all of MA</td>
</tr>
<tr>
<td>Healthcare Cost and Utilization Project (HCUP)</td>
<td></td>
<td>All payer hospital discharge data</td>
</tr>
<tr>
<td>National Emergency Medical Services Information System (NEMSIS) database</td>
<td>University of Utah and state EMS agencies</td>
<td>All reported EMS events for 47 states</td>
</tr>
<tr>
<td>National Provider Identifier (NPI) registry</td>
<td>Centers for Medicare and Medicaid Services</td>
<td>Air ambulance provider data (office and/or base location, state(s) of licensure)</td>
</tr>
<tr>
<td>Homeland Infrastructure Foundation-Level Data (HIFLD Open Data)</td>
<td>Department of Homeland Security</td>
<td>Hospital and trauma center locations</td>
</tr>
<tr>
<td>The Atlas &amp; Database of Air Medical Services</td>
<td>Association of Air Medical Services and CUBRC</td>
<td>Air ambulance base location data</td>
</tr>
<tr>
<td>Medicare Part B public use file (PUF)</td>
<td>Centers for Medicare and Medicaid Services</td>
<td>Aggregated claims for Medicare beneficiaries</td>
</tr>
</tbody>
</table>

**Current access to air ambulance services**

The Atlas & Database of Air Medical Services (ADAMS) provided the most comprehensive and systematic publicly available data on rotor- and fixed-wing air ambulance services, bases, and fleet volumes at the time of our initial analysis\(^\text{18}\). Due to coverage gaps in the current data, the database is currently offline. However, ADAMS remains the best and most comprehensive historical dataset on the air ambulance industry and has been incorporated into the Department of Homeland Security’s internal geographic information system (GIS) portal. According to 2019 ADAMS
data, only about a third of the U.S. geographic area is within 15-20 minutes of a helicopter response. On the other hand, 86.2 percent of the U.S. population is within 15-20 minutes of response. While this is most of the U.S. population, this still leaves over 42 million people residing outside of response area18. The analysis of coverage based on ADAMS data, however, is based on geographic area and population density and does not account for the healthcare demand for air ambulance services.

To better understand the current state of air ambulance access and access disparities, we combined the ADAMS dataset with CMS’s National Provider Identifier (NPI) registry data to fill in gaps on air ambulance providers. In the NPI data, “healthcare providers acquire their unique 10-digit NPIs to identify themselves in a standard way throughout their industry.”19. We used this data as a starting point for initial geospatial analysis of access and availability of air ambulance services to understand the current state of access across the U.S. To get a more accurate picture of ambulance allocation and access we are currently analyzing data on additional factors, including fleet volumes across bases, incidence of disease, and other determinants of air ambulance usage and need beyond population density measures.

**Demand for air ambulance services**

We identify demand by using event-based data from the National Emergency Medical Services Information System (NEMSIS) database and formally explore the drivers of utilization of air ambulance services1. The 2019 NEMSIS Public-Release Research Dataset includes 34,203,087 EMS activations collected from 10,062 agencies located in 47 states and territories. We identify individual-level determinants of air ambulance utilization using NEMSIS data on transport mode, medical conditions, type of destination facility, and patient demographics.

The majority of air ambulance trips are interfacility transports, as shown in Figure 1.

![Figure 1. Types of Air Ambulance Transport from the 2019 NEMSIS Public-Release Research Dataset](image)

Figure 2 shows the distribution of the complaints reported by dispatch to the responding EMS unit. Close to one quarter of all complaints being responded to are for transport from one facility to another. Twenty nine percent of all complaints being reported do not provide additional medical information, shown here with “No Other Appropriate Choice” and “Sick Person”. The large portion of all calls that these categories represent demonstrates how little information EMS responders sometimes have when responding to a call.
Results

The framework demonstrates a visualization showing the map of the service area, superimposed with one set of circles denoting the actual demand of such services, and another set of circles denoting the output of a model based on demand. We show those results for our two use cases, the state of Virginia and the New England region, below.

State of Virginia

The two maps below show the coverage of current versus modeled air base locations in the state of Virginia. Modeled air base locations are depicted as shaded 20-nautical-mile flight circles, while actual base locations are unshaded (i.e., transparent circles bounded by a thin black outline).

Air base locations were modeled using stroke (Figure 3) and coronary heart disease (Figure 6) survey data from PLACES. In the two examples below, the stateside distribution of respondents with a history of stroke or heart disease was similar enough that the two model runs yielded an identical network of air ambulance bases.

Figure 2. Complaint Reported by Dispatch to Responding Air Ambulance Unit from the 2019 NEMSIS Public-Release Research Dataset (due to rounding, percentages do not add up to 100%)

Figure 3. Present-day fixed-wing base locations in Virginia versus modeled locations (MCLP model; 18 bases; 20 nautical mile coverage radius; previously diagnosed stroke survivors from CDC PLACES used as candidate population; hospital service area centroids used as candidate air ambulance base locations)
Regional model for New England

The two large maps below (Figure 5 and Figure 7) show MCLP model results for the New England region. Modeled base locations differ slightly—in the Keane, New Hampshire, region, for instance—due to difference in the geographic distribution of stroke and coronary heart disease respondents. To allow for backup coverage—optionally siting multiple bases in close proximity—we also modeled New England using the MEXCLP model, for both stroke (Figure 6) and coronary heart disease populations (Figure 8). Each of the MEXCLP figures depicts three potential scenarios for backup coverage—from left to right, scenarios with a busy fraction of $p = 0.0$, $p = 0.2$, and $p = 0.4$ are shown.

As the busy fraction increases, it is necessary to locate additional bases throughout New England—especially in the more highly populated coastal regions. The number of bases is held constant at 22, which is the current number of bases available currently in this region, in these model runs. As a result, less populated areas are not allocated a base at the highest busy fractions—and additional bases would be necessary to provide coverage to all regions that have air ambulance service at present). Adding bases to increase coverage or removing bases to decrease costs are applications of our approach that are beyond the scope of this analysis.
**Figure 5.** Present-day fixed-wing base locations in New England versus modeled locations (MCLP model; 22 bases; 20 nautical mile coverage radius; previously diagnosed stroke survivors from CDC PLACES used as candidate population; hospital service area centroids used as candidate air ambulance base locations)

**Figure 6.** Present-day fixed-wing base locations in New England versus modeled locations (MEXCLP model; 22 bases; 20 nautical mile coverage radius; previously diagnosed stroke survivors from CDC PLACES used as candidate population; hospital service area centroids used as candidate air ambulance base locations)
Figure 7. Present-day fixed-wing base locations in New England versus modeled locations (MCLP model; 22 bases; 20 nautical mile coverage radius; CDC PLACES respondents diagnosed with angina or coronary heart disease used as candidate population; hospital service area centroids used as candidate air ambulance base locations)

Figure 8. Present-day fixed-wing base locations in New England versus modeled locations (MEXCLP model; 22 bases; 20 nautical mile coverage radius; CDC PLACES respondents diagnosed with angina or coronary heart disease used as candidate population; hospital service area centroids used as candidate air ambulance base locations)

Conclusions
In this paper, we demonstrated a simplified framework for optimizing air ambulance allocation based on a health outcomes focused approach. We expect this framework to serve as a tool to inform policy decisions that help achieve healthcare performance objectives. For example, the tool can be used by policy makers to identify high supply or duplication of air ambulance services and close access gaps across geographical regions based on location
specific characteristics. Moreover, the tool can be used to install new air ambulance bases in underserved areas; close or consolidate existing bases that lead to redundant supply; reduce transport times to hospitals, improve survival; and reduce the magnitude of air ambulance costs. Moreover, state and local licensing and regulatory agencies can use the optimal location for air ambulance bases if they wish to ensure adequate coverage based on healthcare needs. This methodology can help determine the provision of licensing based on whether operators are meeting a community need.

The main limitations of our approach are the use of two specific clinical conditions, as well as two geographies (Virginia and New England) that may not be generalizable. The use of the two specific conditions may not be applicable to the wider range of conditions for which air ambulances may be used, particularly those with lower prevalence for which less data is available as well as patients who lack a defined diagnosis at the time of pick up. However, use of the detailed NEMSIS data set likely will allow for representing most of these patients in a future application. The use of Virginia and New England may not be generalizable to other regions or the country as a whole, especially for regions that have a lower count of bases or lack air ambulance bases all together. That may make this approach applicable only to larger regions or those that have sufficient population and healthcare capacity to make use of at least one air base. In general, a user requirement analysis and a user evaluation of the system are critical to design systems that are actually useful, usable, and adopted by end users. We have not implemented either as part of the current analysis, so both may be crucial next steps that are needed to adopt our tool in practice.

One other limitation of the approach is the limited validation of the modeling to date. We have validated the code (software) used to implement the model through independent assessments by two team members as well as a peer-review of our code by an expert in epidemiology and software development who is not a member of the research team. We also have validated the approach through application to the state of Massachusetts (results not shown and available upon request). We plan additional validation in the future both through the use of more detailed data and with a comparison of our results with a qualitative assessment of the air ambulance research literature that currently is in progress.

Detailed, data-driven analyses of these elements of the air ambulance industry are not yet available in the U.S. Our framework would help allocate air ambulance base locations based on healthcare demand and individual operator decisions (supply). This will require a comprehensive collection of data across transportation and healthcare domains. Analysis of this data would allow for the extension of our approach to the entire population within a given state or region, as well as future national extensions. Our framework can be used as a tool for decisionmakers when evaluating both the locations of existing base locations and identifying areas for new bases. This tool will provide additional information about who will be impacted by decisions like this, helping determine and forecast the effectiveness of future policies and supplier decisions. Additional information like this would be useful at the state- and local-level, to identify where base locations can best-service the populations with the greatest need, and for base operators, to better understand the expected number of patients and the clinical needs of those patients. Feasibility of adding base locations or removing bases ultimately would need to be assessed at the local level, using our tool as well as clinical and financial analyses of whether changes would adequately address patient need and cover the costs of operators (the authors performed a separate analysis of the financial feasibility of a new base not shown here and available upon request).

The existing study is based on public data on emergency events that does not provide the exact location of each incident. Future work arising from both this project and users of the tool we have produced may include specific analyses of the potential impacts of the changes implied by our model on policy and people (patients). Inclusion of data on location-specific EMS events and traffic and road network data to account for lack of access to trauma centers via ground ambulance would improve fidelity of the geospatial air ambulance demand analysis and allow for applications of the tool such as prioritizing underserved areas when optimizing air ambulance allocations. Furthermore, ADAMS air ambulance base location data was extracted from annual reports, in PDF format, that the Association of Air Medical Services (AAMS) has discontinued. While AAMS will potentially replace ADAMS with a similar product, there is currently no successor to ADAMS, and any future updates will need to be collected using open-source intelligence, via sources such as CMS, insurance providers, the FAA, and crowdsourced aviation and helicopter websites.

References


6. Pulver A. Allagash 0.3.0 [Internet]. GitHub [cited 2022 Jan 4]. Available from: https://apulverizer.github.io/allagash/.


Clinicians’ Perspectives in Using Patient-Generated Health Data to Improve Ischemic Heart Disease Management

Khuder Alaboud, MS¹,²; Maniza Shahreen, MS³; Humayera Islam, MS¹,²; Tanmoy Paul, MS³,⁴; Md Kamruz Zaman Rana, MS⁵; Anissa Morrison, BS²,⁶; Arun Kumar, MD⁷; Abu Saleh Mohammad Mosa, Ph.D., FAMIA¹,²,³,⁴*

¹Institute for Data Science and Informatics (IDSI), ²NextGen Biomedical Informatics Center, ³Department of Health Management and Informatics (HMI), ⁴Department of Electrical Engineering and Computer Science, ⁵Division of Cardiovascular Medicine, University of Missouri, Columbia, MO, USA; ⁶University of California, Santa Barbara, CA, USA; ⁷Correspondent Author.

Abstract
Patients suffering from ischemic heart disease (IHD) should be monitored closely after being discharged. With recent advances in digital health tools, collecting, using, and sharing patient-generated health data (PGHD) has become more achievable. PGHD can complement the existing clinical data and provide a comprehensive picture of the patient's health status. Despite the potential value of PGHD in healthcare, its implementation currently remains limited due to the clinicians' concern with the reliability and accuracy of the gathered data to support decision-making and concerns regarding the added workload that PGHD might cause to clinical workflow. The main objective of the study was to investigate the clinicians’ perspectives towards the use of PGHD for IHD management, focusing on data sharing, interpretation, and efficiency in decision-making. The study consists of semi-structured interviews with seven clinicians. Study results identified four main themes: data generation, data integration, data presentation, data interpretation and utilization in clinical decision-making.

Introduction
Ischemic heart disease (IHD), a type of chronic heart condition affecting about 16 million adults in the USA, is responsible for almost one-third of deaths in both developing and developed countries [1,2][3–5]. IHD, also called coronary artery disease (CAD), is responsible for acute events like myocardial infarction, also known as heart attack [6,7]. The treatments may include medication, procedures, and/or bypass surgery, which can be complex and costly. Patients with chronic conditions such as IHD often require continuous monitoring and care management rather than single interventions. Data collected outside the clinics can inform ongoing care management and provide essential insights into a patient’s health and well-being [1].

Patient-generated health data (PGHD) are health-related data created, recorded, or gathered to provide more profound insights into a patient’s condition. The patient is primarily responsible for creating or gathering PGHD to help address health concerns and usually includes health history, treatment history, biometric data, symptoms, and lifestyle choices [8]. It can better reflect a patient’s true health status and potentially provide value in healthcare delivery. Unlike traditional medical settings, where the healthcare providers collect and manage data within their offices, PGHD is collected outside clinical settings. PGHD allows the providers to expand their knowledge about the patients outside of clinical encounters, creating a more holistic view of patients outside the clinical setting that can inform better medical decision-making. Additionally, patients who use PGHD are more actively engaged in their care and better understand their health status [1]. Moreover, PGHD allows providers to track patients’ clinical information in real-time using applications and tools that record continuous data from the patient. Such applications help providers to monitor patient conditions from home and help the care team with real-time notification in case of medical emergencies [9,10]. Therefore, the health data shared by the patients can be utilized for real-time monitoring of ischemic heart patients, a valuable feature that can lead to more informed decisions and improve health outcomes [11–13]. Previous studies have examined the use of PGHD for monitoring specific health conditions, its potential role in improving patient engagement in care, and its use in clinical decision-making [11–15]. These studies emphasize that integrating PGHD into clinical care can facilitate an accurate and timely diagnosis and improve health care management. Our study examined the potential efficiencies of PGHD in IHD clinical workflow, capturing various aspects of respondents’ perceptions of using PGHD and organizing responses into common themes and outcomes.

PGHD is a valuable source of information for IHD and can help in improving patient-provider communication, patient engagement, and support care decisions [10]. In addition, PGHD can play a vital role in collecting the patient's health data in real-time and can be linked to an intelligent system that can analyze the data to identify the risk factors to trigger an alert when necessary. Thus, PGHD in IHD can lead to early diagnosis, enhanced care, and reduced inefficiencies in the health care ecosystem [22]. As the healthcare industry moves towards patient-centered and value-
based care, patient involvement and the incorporation of PGHD into practice will become imperative [19]. Despite the potential benefits of integrating PGHD into clinical workflow and decision-making, certain challenges may arise [11]. Healthcare providers are concerned about whether the patient-shared information is accurate and reliable in providing the healthcare team with efficient health information that supports clinicians’ decisions [17]. There are additional concerns that integrating PGHD into care delivery might create an additional workload for clinicians.

To address the aforementioned challenges, this study investigated clinicians’ perspectives and attitudes toward using PGHD in IHD management, focusing on data sharing, interpretation, and efficiency in decision-making for the cardiac health care team. The main reason for choosing IHD was that cardiac care is driven by data and requires continuous monitoring and care management. How easily clinicians can arrive at confident diagnoses depends on the quality and completeness of these data and how it is presented for their interpretation. Clinicians often rely on patient narratives, observations, and patient-reported outcomes during clinic visits to assess the efficacy of heart treatment and reach confident decisions [20]. Data collected between visits can better inform ongoing IHD care management and provide essential insights into a patient’s health and well-being [21]. Moreover, since the collection and use of PGHD can help patients be engaged in their own self-management as well as allow clinicians to monitor their symptoms of IHD remotely, cardiologists may be more likely to adopt and use PGHD as it supports ongoing, real-time monitoring and patient self-management. To understand the clinicians’ perspectives in using PGHD to improve IHD management, this study conducted semi-structured interviews to answer the following research questions: (1) what type of patient data would clinicians like to track for ischemic heart patients, (2) how do we integrate these data into a clinical workflow, (3) what are the clinicians’ preferences on the methods for data gathering, accessing, and visualizing, and (4) how useful and effective those data are in clinical decision making for the health care team?

**Methods**

For this study, qualitative data in the form of semi-structured interviews were collected from the University of Missouri (MU) Cardiology Clinic, Family Medicine, and Internal Medicine. The study was reviewed and approved by the MU Institutional Review Board (IRB #2010369 HS). The semi-structured interviews were designed with open-ended questions for guiding the discussion with a health care team to understand their perspectives about PGHD for IHD. In addition, the open-ended questions allowed further follow-up questions among the participants, which provided them with more opportunities to express their opinions fully. Moreover, interview questions were designed to elicit information regarding what type of data the health care professionals need to track for the ischemic heart patients, how would they like to access those health data, and what is the preferred methods for gathering, accessing, and visualizing the data? Overall, the interviews allowed the participants to evaluate the usefulness and effectiveness of those data in clinical decision-making for the health care team. A total of seven participants were recruited, including physicians (n=4), nurses (n=1), and clinical researchers (n=2). (Table 1) shows the participants’ information, including average professional experience, current roles, and specialty areas.

**Table 1.** Participant summary for the study

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional Experience</td>
<td>Mean = 23.64 years</td>
</tr>
<tr>
<td></td>
<td>Maximum = 40 years</td>
</tr>
<tr>
<td></td>
<td>Minimum = 8.5 years</td>
</tr>
<tr>
<td>Specialty</td>
<td>Family Medicine = 3 persons</td>
</tr>
<tr>
<td></td>
<td>Clinical Research = 2 persons</td>
</tr>
<tr>
<td></td>
<td>Cardiology = 2 persons</td>
</tr>
<tr>
<td>Current Roles</td>
<td>Physician = 4 persons</td>
</tr>
<tr>
<td></td>
<td>Registered Nurse = 1 person</td>
</tr>
<tr>
<td></td>
<td>Clinical Researcher = 2 persons</td>
</tr>
</tbody>
</table>

Each interview was scheduled on a convenient date and time discussed prior with the respondent. The interviews took place at the University of Missouri Health Care, Columbia, MO. The duration of each interview was approximately 20 minutes, and the audio recordings of the interview were collected for the appropriate data transcription afterward. Thematic analysis (TA) was used to analyze the data and understand the clinicians’ attitudes towards PGHD sharing.
and interpretation. TA is the process of identifying patterns or themes within qualitative data. In addition, this method allows coding and theme development directed from the content of the dataset. It offers flexibility to the researchers to incorporate multiple theories applicable to research questions beyond individual experiences. TA consists of a six-step process (Figure 1) to identify, analyze, and report qualitative data seeking to provide insights into patterns and common themes across the qualitative data [22]. Subsequently, we assigned specific codes to identify the features of the data for each interview. These codes provide the context of the interviews with all the relevant data patterns. The codes are then combined to develop a potential theme of the data sets.

![Figure 1. Thematic Analysis phases](image)

**Results**

The outcomes derived from the interview questions were organized into four main themes, including (1) data generation and data collection, (2) data integration, (3) data presentation (4) data interpretation and utilization in clinical decision-making. Figure 2 shows the flow of patient-generated health data in each of the four themes.

![Figure 2. PGHD model](image)

1. **Data Generation and Data Collection Tools**

This theme focuses on seeking the participants' opinions based on their experience on what type of healthcare data are more important to track and record for the patients diagnosed with IHD, as shown in Table 3. Blood pressure, weight, physical activity, and blood sugar were frequently mentioned as important data types to collect at home continually. All the participants agreed that blood pressure is the most important data to be tracked because hypertension is the most common cause of IHD and has been poorly controlled by the patients. Moreover, one participant pointed out that home blood pressure monitoring gives a more accurate reflection of one's true blood than clinical blood pressure measurement.

*It is important to measure blood pressure at home because clinical readings might not reflect a person's true blood pressure.*

The second important feature, according to the participants, was daily weight tracking at home. Participants noted that tracking daily weight is important because it is very common for people with heart failure to experience rapid changes in their weight. Weight gain could be a sign of fluid buildup in the body that might lead to heart failure because the heart is not pumping properly. Patients with an ischemic heart condition are recommended physical activity and exercises; the participants elicited the usefulness of developing a mobile application, smartwatch, or patient portal feature to allow these patients to track their physical activity. Participants also recommended providing educational videos on how to perform certain exercises and allowing patients to add daily tasks, goals, and progress using those applications. In addition, for diet recording, the applications should allow the patients to monitor weight, calories consumed, the number of carbohydrates they take, and the amount of water they drink, allowing monitoring of the nutritional value of daily food consumption. Patients might also have the option to connect to smart scales for automatic weight monitoring or the option to add weight data manually. When a discrepancy is detected regarding the nutritional value norms or the patient has too high a glycemic index in any meals, the physicians want to receive notifications and send recommendations on reading educational articles relevant to their conditions. Another health
data that was highly recommended by the participants to be tracked was blood glucose, especially if the patient has diabetes or is taking insulin. Participants noted that self-monitoring of blood glucose data could help clinicians to consider signs of prediabetes when prescribing medications.

Table 3. Types of health to track at home

<table>
<thead>
<tr>
<th>Health Data</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood Pressure</td>
<td>7</td>
<td>100</td>
</tr>
<tr>
<td>Weights</td>
<td>6</td>
<td>85.71</td>
</tr>
<tr>
<td>Physical Activity</td>
<td>3</td>
<td>42.85</td>
</tr>
<tr>
<td>Blood Sugar</td>
<td>3</td>
<td>42.85</td>
</tr>
<tr>
<td>Pulse</td>
<td>2</td>
<td>28.57</td>
</tr>
<tr>
<td>Angina Episodes</td>
<td>2</td>
<td>28.57</td>
</tr>
<tr>
<td>Medication Adherence</td>
<td>2</td>
<td>28.57</td>
</tr>
<tr>
<td>Diet</td>
<td>1</td>
<td>14.28</td>
</tr>
</tbody>
</table>

The participants stated that currently, there is no standardized way to collect and manage PGHD. However, with all the existing data collection methods, it could be useful to develop an application to collect and analyze real-time health data in a standardized manner to inform medical decisions effectively.

"Having the patient go home and do their blood pressure recording, potentially using an app would help us. Currently, we do it in the written format during patient visits using survey scales while patients are waiting in the waiting room."

2. PGHD Integration and Accessibility

One important theme discussed with the participants was the integration of PGHD into care delivery and data accessibility at the point of care. Although the care teams believe that the access to PGHD collected from digital health applications, wearables, and in-home medical devices can lead to better remote patient management, the challenge lies in integrating PGHD without creating additional workload to the frontline clinicians. Therefore, it is crucial to provide effective remote patient management access while supporting more active clinical intervention. The health care professionals reported that they want the integration of PGHD in the EHR to make it easier to use and more accessible at the point of care. One clinician described her experience:

"I would like that data to go right into our EMR to allow us to view the patient's note and create a note for their visit. It's a lot easier for me not to have to open a whole separate application. It would be nice to have that fall right into the rest of the data, such as labs, vital signs, diagnostic tests."

Moreover, participants emphasized that integrating technologies into care delivery without additional workload can help the clinical teams make the PGHD more acceptable in their workflow. Significant trends of the health-related data should be notified electronically to the health care team. It can be beneficial to take necessary actions for the patients.

"As a provider, I may have hundreds of patients. So that's what I mean is at some point, there have to be triggers or an alert system to notify me when I need to access the data. I wouldn't access it until receiving a notification that a patient needs clinical intervention. It would be time-consuming to look at the data every day to try to figure out who's in trouble."

The health care team identified data accessibility at the point of care, privacy, and confidentiality concerns of using PGHD as essential areas to consider when implementing health IT tools to assist patients with collecting PGHD. The health care professionals' preference to have PGHD integrated with the EHR was mitigated by legal concerns related to patient privacy and data storage. When they share this data, the patients want to know that it is appropriately received and used by their providers, kept private, and secure in their EHR.

3. PGHD Presentation and Visualization

Meaningful display of the health data generated at home is essential, according to the participants. The data summary should be presented for meaningful display to ensure a complete understanding of the data. PGHD provided the health care professionals with data dashboards through Web-based platforms that collect, summarize, and visualize PGHD. The participants reported that the ability to sort or summarize data in a descriptive manner or graph it in different ways helped health care professionals see patterns more quickly in the data and extrapolate meaningful value from these data. Clinicians are not expected to study various PGHD records and make conclusions about the patients' health.
status. Instead, technology is here to assist and notify clinicians in predefined cases, such as IHD. The system should constantly monitor patient health status to detect abnormalities. It should identify patterns and changes at a specific time interval or depend on other health factors for patients with an ischemic heart condition. In addition, it is essential to make clear data presentations that highlight abnormal values in different colors to help users identify any deterioration of the patient's status within seconds as one participant expressed her opinion.

Clinicians don’t want to see daily weights; they want to see if the weight has been trending up or down. Something that takes all of those inputs and summarizes them into something that physician can quickly review. For example, a patient’s daily weight recording information goes into the provider’s health data analytics system. Suppose the trend indicates that the patient gains weight too quickly. The patient’s care team receives an automated notification about this change.

Most participants preferred graphical representation of the health data to help provide a more accurate picture of the patient's health status and identify the upward or downward trends, including charts/graphs like a bar chart, pie chart, etc. As participants expressed their opinions:

Graphically represented data is easiest to understand and discuss with the patient to make medication changes or therapeutic changes. Graphical visualization can reduce the cognitive load and can help identify spikes and trends and tell you if things are falling either above or below a specific parameter. Graphical visualization can also make it easier to look at large pieces of data at a glance, unlike raw data, which can be hard to analyze and sort through. Graphs can represent fluctuation points and relevant trends and quickly identify the peak points or the low points and help you pinpoint the exact time or exact data you want to see.

4. Data Interpretation and Utilization in Clinical Decision Making

This section focuses on the usefulness of PGHD in clinical decision-making for the health care team. The essence of PGHD is to provide an insight to support clinicians’ decision-making process in developing a treatment plan and track post-discharge treatment progress. Participants reported that PGHD is aggregated and analyzed by the organization's health data analytics system to process the results and compare them to previous measurements. Suppose the analysis uncovers negative trends in the patient’s health status, for example, a sudden or steady gain in daily weight, increased heart rate, systolic or diastolic blood pressure. In that case, it will automatically notify the care team about possible health risks.

Many participants consider engaging patients with IHD as vital in the decision-making to develop a treatment plan and track post-discharge treatment. Patients can track and share health data such as heart rate, rhythm, sleep, temperature, blood glucose, and blood pressure from home. Since clinicians mainly rely on the information presented during visits to the clinics, having access to PGHD data would lead to a better understanding of the health status of patients with IHD, especially when it comes to highly variable factors such as blood pressure readings.

We may be less available to see our patients as frequently as we would like as cardiologists. So, if you see the patient less frequently, then knowing what’s going on between the visits potentially, you can even instruct the patient to contact the office to make an earlier visit. If you notice if your blood pressure is constantly above a certain number, if your sugars are above a certain number constantly, if your weight goes up by three pounds, we would like to see you before the next scheduled visit. So, it may reduce symptoms, readmissions, and events.

Many participants identified monitoring patients more closely as a tool that can help reduce clinical events for post-discharge IHD patients. In addition, having customizable clinical metrics based on the users’ specific needs to reduce information overload can help in an emergent situation to provide the healthcare team with immediate guidance regarding the point of care.

Discussion

Utilization of PGHD in Clinical Workflow

Integrating PGHD into IHD workflow brings efficiencies that are essential for comprehensive cardiac care. PGHD can be generated by using different methods utilized in a patient’s day-to-day life outside of clinics. From the literature, digital health tools for collecting PGHD can be organized into three main methods. First, wearable devices allow people to track and collect different biometric data such as heart rate, pulse, blood pressure, and temperature. Second, mobile applications allow the patients to manually enter their information and track their lifestyle measurements such as physical activity, diet, and medications. Third, registered medical devices can track data, such as heart rate, blood glucose, and other biometric data, often through remote monitoring [10,13,14,23–29]. A
technology that integrates as many of these methods as possible in a standardized way could help create a much fuller view of the patient’s condition for providers. However, implementing PGHD into the clinical workflow will bring challenges that need to be considered for its success. These include the volume of PGHD generated from the continuous monitoring can be significantly high, creating further information overload [17]. Also, the requirement of regular reviewing of the information will generate additional workload. Thus, this will bring the need for an incentive model for the providers [30]. In addition, the organizations need to consider appropriate levels of training for the clinical staff. Another challenge will be recruiting and enrolling patients into the PGHD program. Various approaches must be adopted to create enthusiasm, including face-to-face conversations during office visits and online interventions. Also, the objectives and expectations must be clearly communicated with the patients.

Efficiency of PGHD in Decision-Making
The quality and completeness of data are essential for clinical decision-making. From the providers' perspective, as reported in our study findings, it is highly desired that the PGHD is integrated into the clinical workflow through the electronic health record (EHR) system. From the EHR development and maintenance perspective, maintaining a high volume of data in the EHR could be concerning. As a result, a middle-tier application might be utilized to reduce the data overload to the EHR system, but the clinical decision-making components such as alerting, summarizing, and visualization are integrated into the clinical workflow through the EHR system [31]. It is essential to recognize that the efficiency of PGHD in decision-making is highly reliable on the successful implementation of meaningful altering, summarization, and visualization. Our study findings pointed out that meaningful utilization can help early intervention and help control symptoms such as high blood pressure and reduce further health complications. For simplifying the interpretation and analysis of PGHD, providing a set of tools such as clinicians facing dashboard for sorting a panel of patients with higher risks needing immediate attention, methods for identifying significant values, or tools for visualizing data over time can help significantly boost the use of PGHD in the clinical decision-making process.

Remote Monitoring and Clinical Intervention
Health data analytics is key for monitoring a patient’s post-discharge recovery and preventing acute events or progression of chronic conditions such as IHD. PGHD enables clinicians to access health data and continuously monitor patients after being discharged from the hospital. For example, hypertension is closely linked to increased cardiovascular risk, and patients could have a condition known as white-coat hypertension, in which a patient's blood pressure readings are higher when taken at the doctor's office compared to other settings, such as at home. This may be attributed to factors such as the patients' experience during medical appointments, tiredness from traveling, concern regarding medical costs, etc. [32,33]. EHR contains vitals, lab results, medical images, medications, procedures, and other data. The clinician can rely on these data to create the basis for a treatment plan. However, EHR is only the starting point. To reduce the risk of developing post-discharge acute events, providers can enable PGHD to complement the existing clinical data in the EHR. The health status of a patient can be tracked against a pre-established health goal between in-person clinical visits, thus enabling the real-time monitoring of the changes in the health status from day to day. Thus, real-time data can be used to identify the risk factors to trigger an alert and notify providers for early intervention. The results confirm that clinicians are interested in using technology that shows PGHD in the form of summarized data, identifies patterns and trends over time and notifies clinicians immediately to react proactively to adverse health events. Clinicians also believe that visualization and summary representation of the dataset would be beneficial for making better and more efficient clinical decisions. Effective monitoring of the health status based on the PGHD can ensure the timely identification of high-risk patients. It creates an opportunity for early intervention reducing complications, hospitalizations, and readmissions. However, more work is required for the efficacious implementation of PGHD-based remote monitoring, which relies on correctly identifying the characteristics of health systems, patient populations, and intervention designs [19].

Supporting Patient Engagement and Self-management
Integrating PGHD in cardiac care can substantially help patients to engage in their care planning and clinical decision-making. A study investigating the patient perspective on PGHD [34] reported the patient motivations for PGHD engagement. Some patients may set specific health goals and track their progress based on the health data. Research indicates PGHD is considered an effective tool to help increase patient health literacy, improve health consciousness and health behavior, which in turn enhances patients’ knowledge on health conditions and risks [34]. Aware of the health conditions and risks, patients and their family members can gain more control over patients’ health and recovery process. Additionally, the patients are more willing to provide the clinicians with any information that could be useful for effective diagnosis and care management. These data enable clinicians to adjust the treatment plan and help patients
achieve better health outcomes [35]. One of the important challenges that health care providers need to consider is interrupted measurements and incomplete data. In addition, the patient engagement level is subject to the objectives for PGHD collection and may change with time. For example, if PGHD is collected for acute health care, a patient may be reluctant to share the health data after regaining the normal function [35]. Therefore, the changes in the level of patient engagement need to be assessed for the continuity of the participation.

Conclusion

Based on the study results, four main themes were identified: data generation and collection, data integration and accessibility, data presentation and visualization, data interpretation and utilization in clinical decision-making. Various aspects of PGHD were discussed in this study to highlight the efficiency of PGHD in IHD clinical decision-making. Also, the study discussed some challenges that might hinder PGHD implementation. The growth of consumer technologies, including mobile apps and wearable devices, has allowed patients to collect their health-related data outside of healthcare encounters. It can also lead to a significant increase in health-generated data. From a health informatics perspective, implementing digital health solutions can facilitate data collection, sharing and integration with healthcare system. This study emphasizes the importance of PGHD and the usefulness of the data in clinical decision-making. Incorporating PGHD into IHD clinical workflow can help clinicians develop a treatment plan and track post-discharge treatment progress. PGHD can also influence how patients with IHD report and monitor health conditions. Patients who contribute to PGHD are more actively engaged in their care and may have an increased understanding of their health status and influence how individuals report and monitor symptoms. Data collected outside of healthcare encounters can impact remote monitoring and provide a valuable source of information for ongoing IHD management. These generated data be liked to an intelligent system that can analyze the data to identify health factors. Also, the data should be summarized and graphically visualized to reduce the cognitive workload. Changes in patient’s health measurements and overall trends can be identified with automated push notifications in case of possible health risks. Additional research is needed to understand the patients’ perspective of PGHD. Also, more research is necessary to include clinicians from various departments.

References

10. Lavallee DC, Lee JR, Austin E, et al. mHealth and patient generated health data: stakeholder perspectives on opportunities and barriers for transforming healthcare. mHealth 2020;6:8–8. doi:10.21037/mhealth.2019.09.17


Lavallee DC, Lee JR, Austin E, et al. mHealth and patient generated health data: stakeholder perspectives on opportunities and barriers for transforming healthcare. *mHealth* 2020;6:8–8. doi:10.21037/MHEALTH.2019.09.17


Conceptualizing a Data Infrastructure for the Capture, Use, and Sharing of Patient-Generated Health Data in Care Delivery and Research through 2024 PRACTICAL GUIDE. 2018.


119
BEHRTDAY: Dynamic Mortality Risk Prediction using Time-Variant COVID-19 Patient Specific Trajectories

Alaleh Azhir, MSc¹, Soheila Talebi, MD², Louis-Henri Merino, BSc³, Yikuan Li, MSc¹, Thomas Lukasiewicz, PhD¹, Edgar Argulian, MD, MPH², Jagat Narula, MD, PhD², Borislava Mihaylova, MSc, DPhil¹.⁴

¹University of Oxford, Oxford, UK; ²Icahn School of Medicine at Mount Sinai, New York, NY, USA; ³Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland; ⁴Queen Mary University of London, London, UK

Abstract

Incorporating repeated measurements of vitals and laboratory measurements can improve mortality risk-prediction and identify key risk factors in individualized treatment of COVID-19 hospitalized patients. In this observational study, demographic and laboratory data of all admitted patients to 5 hospitals of Mount Sinai Health System, New York, with COVID-19 positive tests between March 1st and June 8th, 2020, were extracted from electronic medical records and compared between survivors and non-survivors. Next day mortality risk of patients was assessed using a transformer-based model BEHRTDAY fitted to patient time series data of vital signs, blood and other laboratory measurements given the entire patients’ hospital stay. The study population includes 3699 COVID-19 positive (57% male, median age: 67) patients. This model had a very high average precision score (0.96) and area under receiver operator curve (0.92) for next-day mortality prediction given entire patients’ trajectories, and through masking, it learnt each variable’s context.

Introduction

In mid-April 2020, New York had more cases of confirmed coronavirus patients than any country in the world. In many patients admitted to hospital, this disease involves an intricate interplay of multiple biological pathways involving various organs producing interdependent, measurable signals. Although epidemiological and clinical characteristics of patients with COVID-19 in various parts of the world have been reported¹², a dynamically updated assessment of patients’ disease progression over time is lacking. Identifying and modeling the changes of these measurements over time allows for dynamic mortality indicators for COVID-19 patients, which can aid decision making in hospitals.

Throughout COVID-19 patients’ stay in the hospital, various parameter measurements are collected on a daily basis by medical staff as the patient undergoes care in different units. The measured data consists of static measurements, which do not change throughout patients’ stay, and periodic measurements such as lab tests and bedside vital signs. The wealth of information gathered throughout patients’ stay and recent advances in efficient machine learning approaches allow for great progress in developing highly accurate models to be used in inference tasks such as in a dynamically updating mortality risk model.

Approaches in modeling patients’ stays require individuals to be represented as vectors of features, which are usually engineered through consultations with experts. The benefit of deep learning is that through a sequence of layers that transform the input information both linearly and non-linearly, the model itself can learn useful features and a representation of individuals with minimal medical guidance. Deep learning methods are increasingly being applied to answer health care questions. In 2014, it was shown that deep neural networks could potentially outperform support vector machines and decision trees with manual feature engineering in various prediction tasks⁵. Later efforts incorporated the temporal orders of events through the use of recurrent neural networks (RNNs), convolutional neural networks (CNNs), and long short-term memory (LSTM). Deepr (Deep record) was used for estimating the probability of readmission by treating one’s medical history as a sequence of concepts and inserting code words between different consecutive visits to denote the time difference between them⁴. RNN auto-encoder models were augmented with an attention mechanism to allow for attending to various time steps using soft/hard attention⁶. To capture the long-term dependency of events (i.e., hypertension can remain a risk factor throughout one’s life), DeepCare, an LSTM architecture with attention mechanism was introduced⁷.

Recently, transformer models were proposed for natural language processing, which by incorporating self-attention create intermediate representations of the input by accounting for its previous and future time steps. The success of
these models has led to their applications to electronic health (EHR) datasets such as BEHRT and TAPER8,9. Given that BEHRT has outperformed some well-established deep learning risk prediction models, such as Deepr4 and RETRAIN10, we applied this model to daily COVID-19 patients’ hospital stay measurements to predict next day mortality.

Methods

Dataset Generation: We retrospectively retrieved de-identified health records of patients admitted to five hospitals within the Mount Sinai Health System (MSHS) between March 1st and June 8th 2020, all of whom had tested positive for COVID-19. This study was approved by the Institutional Review Board at the Icahn School of Medicine at Mount Sinai. We then constrained our cohort to patients who were not admitted again to the hospital for a second time during our time period, thereby removing any patient with more than one admission. The data included patient demographics, past medical history, daily vitals, and all labs during their hospital stays. Relevant patient events (ICU admission during patients’ stay, and final outcome: discharge to home, discharge to hospice care or death) were recorded.

Dataset Preprocessing: To reduce data complexity, we limited the number of days of patients’ hospital stay to a maximum of 30 days. For patients whose length of stay exceeded 30 days, we only considered the last 30 days of their stay. Furthermore, to remove questionable laboratory measurements and reduce input data complexity, we applied plausible ranges and categorized all of our continuous laboratory measurements into normal, low, and high value categories based on the cutoff values found in the literature. Measurements outside these ranges were set as missing.

For each patient $p \in \{1, 2, ..., P\}$, the hospital stay consists of a sequence of daily measurements; each day contains dynamically changing vital and laboratory measurements and static age, BMI, patient’s history, and other information.

We denote each patient’s hospital day as $D_p = \{d^1_p, d^2_p, d^3_p, ..., d^n_p\}$, where $n_p$ denotes the patient p’s length of stay in the hospital, and $d^j_p$ contains all the static and variable measurements on the jth day shuffled to a random order (i.e., $d^j_p = \{O2 sat normal, BMI overweight, WBC count low, CRP NA, ...\}$). To prepare data for BEHRTDAY, we order the hospital days temporally, and use BEHRT’s format. This adds the term CLS to denote the start of the medical history and the term SEP to denote the space between each day, obtaining the following sequence per patient: $V_p = D_p = \{CLS, d^1_p, SEP, d^2_p, SEP, d^3_p, ..., SEP, d^n_p\}$.

Model Development: We applied a transformer method (BEHRT*) to predict the next day mortality in COVID-19 patients given the patients’ entire stay in the hospital, as a binary classification problem. Various challenges of this dataset include: 1. complex and interdependent interactions between past, present and future concepts, 2. interdependence of our daily measurements (e.g., hypertension affects blood pressure, white blood cell count being related to lymphocyte percentage, etc.), and 3. irregular practices leading to different frequencies at which various labs are measured during patients’ hospital stay. This model depicts daily laboratory measurements or static patient demographics as words, every hospital day as a sentence, and the patient’s entire hospital stay as a document to use multi-head self-attention, positional encoding, and masked language model (MLM8, 11). BEHRT pre-trains deep bidirectional representation of daily medical measurements (static or variable) by jointly conditioning on both left and right contexts in all layers.

BEHRT learns one’s hospital stay through a combination of four embeddings: daily status, position, day, and segment (Figure 1). The daily status embedding informs the model about the static patient information and trajectories of one’s past variable vitals or laboratory measurements. Due to common trajectories that are shared for COVID-19 patients, knowing one’s past laboratory measurements could help improve the accuracy of predicting one’s future laboratory values. As positional embedding, we used BEHRT’s default pre-determined encoding (inspired by15), increasing monotonically by 1 from position 1. The day embedding provides the model with a sense of time and depends on when time-varying laboratory values were collected. The segment embedding includes two representations (referred to as A and B) that change alternatively between different days. It is used to separate information from different days but does not increase monotonically like the day embedding. Therefore, it can be considered as a complementary feature other than day embedding for distinguishing information from each day during one’s hospital stay. For each day, we shuffle the order of static and variable patient information embedded in the daily status ($d^j_p$) to enforce the attention mechanism to investigate intra-day relationships among concepts.

To avoid overfitting, the main hyperparameters were chosen using a three-fold cross-validation on a randomly chosen 80% training set and 20% validation set. To find the optimal hyperparameters, we tested all 80 combinations of the following settings: intermediate size (150, 200, 250, 350, or 450), hidden size (60, 72, 84, or 96), number of hidden layers (3 or 6), and number of attention heads (6 or 12) – for additional details, see the original BERT paper11. The
number of hidden neurons and layers determine the complexity of the model, while dropout randomly drops input neurons from the network during training to prevent overfitting. During the training process, the optimization function (ADAM) determines how to gradually modify the model parameters to lower the prediction error quantified by the loss function.

**Figure 1.** BEHRTDAY Architecture. The input consists of static and variable measurements throughout a patient’s hospital stay. The BEHRT section consists of a daily status embedding that embeds both static and variable measurements in a random order, day and position embeddings that both encode consecutive numbers from day 1 until the day of discharge, and the segment encoding that alternates between each visit to signal a change in each visit day. To predict the day 7 mortality risk, BEHRTDAY’s Transformer-based architecture is shown on the right. Pretraining is first done by the MLM model that can predict masked daily status tokens. To predict the next day mortality, both the weights pretrained by MLM and the weights for binary classification of mortality prediction are fine-tuned.

**Pretraining using Masked Language Model (MLM):** We initialized daily status, day, and segment embeddings randomly. While pretraining the network (similar to the approach in8), we left 86.5% of daily status words (i.e., temperature normal, lymphocyte low) unchanged, 12% were replaced with [mask] and the remaining 1.5% were replaced with a randomly chosen daily status word (temperature normal → O2 Saturation high). Through this injected noise, BEHRT does not know which words were masked, therefore is forced to store a contextual representation of each daily status word. The low rate of change (13.5%, as used in8) allows the model to still understand the real patient information, and overall COVID-19 disease trajectory. The MLM11 classifier maps the daily status words to the masked words and was evaluated using the precision score (the ratio of true positive over the number of predicted positive samples). The average precision is calculated over all daily status words and all patients.

BEHRT’s transformer-based architecture (Figure 1) takes in the daily embeddings as input and passes them through multiple multi-head attention layers. It is first pretrained by the MLM model that predicts the masked daily status words to learn network parameters and the daily status embeddings. When predicting the next day mortality, BEHRT then fine-tunes the networks weights pretrained by the MLM task and further learns the weights of the binary classification task of mortality prediction.

**Model Evaluation and Interpretation:** To evaluate BEHRTDAY, we assess its learning of disease trajectories by predicting mortality. We used Monte Carlo cross validation to evaluate and compare the performance of our model, by performing 20 random splits of observational units into training (80%) and test (20%) sets. The network pools the representation of the patient and passes it along to a single feed-forward classifier layer for the output (next day mortality prediction). We evaluate the model using the area under receiver operator characteristic (AUROC) curve and average precision score (APS). Furthermore, to evaluate the dynamic mortality prediction, after fitting the model on the entire hospital stay of 80% of our patients, we then predicted next day mortality in our test set over time (i.e., given test set patients’ stay up to day x, where x ∈ {1, 2,...,30}). We evaluated the model’s performance both on a
collective level by calculating APS and AUROC over time, and on an individual level, by plotting different patients’ mortality risk calculated over time. Lastly, we use individual covariate (or group) ablation to measure variable (or groups of variables) importance, by eliminating one variable (or group) at a time, refitting the model and measuring the change in APS, loss or AUROC. Furthermore, we fit the ablated models on our time varying test set to observe the effects of variable ablation over time. Lastly, our model was compared to both a random forest survival model fitted only on admission patient information and a random forest model that incorporates time-varying trajectories. 

Results

Study Population: After removal of 579 patients who were re-admitted into the hospital in our time frame, the final data set includes 3699 COVID-19 positive patients, admitted between March 1st and June 8th 2020. Patients’ median age is 67 years, and 57% are male, with median BMI of 27.7. Figure 2a provides the distribution of patients’ hospital length of stay. The median length of stay is 5 days. Laboratory markers were categorized and tracked during patients’ admission.

Model Optimization: To avoid overfitting and optimize the number of epochs (runs through the data), we plot the test and training loss for our next day mortality model when fitted on the entire length of patients’ stay, over the number of epochs in Figure 2b. One can see that the model converges quickly after 4-10 epochs, which is approximately when the average precision score is maximized, therefore, to avoid overfitting, we only use 10 epochs in our next day mortality model.

Figure 2. a) Distribution of patients’ length of stay in the hospital; b) On the left, we have test and training loss with respect to the number of epochs, and on the right, we have the average precision score and area under receiver operator characteristics curves.
**MLM Embedding:** We used grid search to determine the optimal hyperparameters for our MLM and next day mortality models. This resulted in an optimal architecture of intermediate layer size of 250, hidden size of 96, 3 hidden layers, and 6 attention heads. We trained the MLM task with 50 epochs, and the pretrained MLM model’s average precision score given the patients’ entire hospital stay was 0.38. Note that the pretrained MLM model is predicting which words are masked, which then gets used and modified by the next-day mortality risk model.

To evaluate our MLM embedding technique, we reduced the dimensionality of the daily status measurements’ vectors using t-SNE to 2 (Figure 3). Here, we see that comorbidities, blood gas (arterial and venous), full blood counts, kidney, and inflammatory and coagulatory markers group together. Furthermore, we can observe that high, low and normal values also group together. In Plot F, we can see that high respiratory rate and age fall near ICU admission as expected clinically. The reason for dispersion of various liver markers could be due to dimensionality reduction (from 96 to 2 dimensions). This confirms BEHRT’s ability in understanding various daily status markers, and grouping them based on their similar context.

![Figure 3. MLM embedding plots: we can see that most neighborhoods correspond to similar clinical markers of various organs and are aligned with medical knowledge. Specifically, co-morbidities, blood gas measurements, and full blood counts all fall next to each other.](image)

**Discharged Expired or Discharged Home:** In Figure 4, we plot the ROC curve and precision-recall curve, with their corresponding 95% confidence interval, for internal validation of the next day mortality given patients’ entire hospital stay. As expected, the model is highly accurate with APS of 0.96 (95% CI of 0.9495 - 0.9695) and AUROC of 0.92 (95% CI of 0.8989 - 0.9391).
Figure 4. ROC and precision-recall curves: the confidence interval is obtained through 25 Monte Carlo cross-fold validation.

To assess variable importance in predicting outcomes given patients’ entire hospital stay, we utilized ablation, the result of which is shown in Figure 5. On the left, we ablate each variable individually, and on the right, we ablate categories of variables based on what organs/systems they inform. The control (where nothing has been ablated) is shown as a star. We expect that the higher the variable importance, the higher the drop in the APS and AUROC. On the left, we observe that ablating variables such as CK-MB, Age, pO2 Venous, LDH, BMI, base excess, several comorbidities such as Hepatitis, CAD, AFIB, and coagulation marker PTT seem to be associated with higher drops in AUROC and APS compared to the control. When grouped together, we observe that groups of Vitals, Sex, and BMI have the highest importance values, as they are associated with largest drops in AUROC and APS.

Figure 5. Ablation: On the left and right, control (represented by star) is the model that includes all features, and no variable has been ablated. As individual features or groups of features get ablated, the performance of the model changes. We expect that when more important variables are ablated, there would be larger drops in APS and AUROC.
**Dynamic Next Day Mortality**: We then evaluated the performance of our model for the next day mortality over time by limiting our test data set to include up to the first $x$ days of each patient’s hospital days where $x$ ranges between 1 to 30. Since we are considering patients’ hospital stay for up to $x$ days, for the model’s performance measured for up to 10 days in the graph, it includes for example both a patient who was discharged after 4 days and the next day mortality risk for his or her 5th day of stay, as well as a patient who was discharged after 15 days, and his or her first 10 days of hospitalization and next day mortality risk for his or her 11th day of hospital stay. Therefore, this results in cases resulting in death to increase as more days of patients’ hospital stay is considered. Lastly, as the ICU marker was binary and did not contain the information regarding which day the patient was admitted to the ICU, this section’s analysis has ablated this variable (it is not included in any of the models below including "None").

In Figure 6, we evaluate our model’s (trained on patients’ entire hospital stay for the training set) APS for the next day mortality risk prediction given test patients’ $x$ days of hospital stay, when ablating each category of variables. Ablations of variable measurements are displayed on the left, and those of static measurements are displayed on the right. Looking at when all variables are present (besides ICU marker) (the green “None” model), the model has an APS and AUROC of about 0.92 and 0.83, respectively, given only the first day’s data (2nd day’s mortality risk), which then drop to a minimum value of 0.90 for APS and 0.81 for AUROC when given measurements up to patients’ sixth day of hospital stay, and then gradually increases to a value of 0.96 for APS and 0.93 for AUROC when given patients’ up to 30th day hospital stay (i.e., if someone gets discharged after 4 days, then for calculating the APS and AUROC at the 30th day point, we are predicting that person’s next day mortality for day 5). This increase in APS and AUROC is expected, since as we provide more information regarding patients’ hospital stay by including more days, we expect to get more accurate predictions regarding their next day mortality risk. The initial drop and fluctuations when ablating other variables in APS and AUROC for the first week could be possibly random, as there is a lot of uncertainty at the start. The effects of ablating various variables are quite large at the beginning potentially due to having a large uncertainty given less information about patients’ trajectories. A potential reason for the high APS and AUROC for day 1, which gradually decreases throughout the first week, is the fact that if one dies after 1 day of hospital stay, its prediction could be easier than predicting one’s death after multiple days of hospital stay, perhaps since the patient arrives in a very critical stage and there is not enough time for the patient to be managed. This may explain why the accuracy of the model is higher for predicting next day mortality at day 1 compared to day 6 even though it has less information about patients at day 1. This model performs significantly better than a random forest survival model fitted only on the admission patient data (95% APS CI index: 0.77 - 0.80) and a random forest model that incorporates time-varying trajectories\textsuperscript{13} (APS ranging from a minimum of 0.55 at $x = 2$ days to a maximum of 0.92 at $x = 30$ days).

**Figure 6.** Next day mortality risk prediction’s APS for patients in the test set given their information up to day $x$, where $x \in \{1, ..., 30\}$, using BEHRTDAY ablation models. The legend displays which variable has been ablated. On the left, we have variable measurements, and on the right, we have static measurements.
Regarding variable importance, from Figure 6, we can observe that, as expected from Figure 5, ablations of Vitals, Liver and Kidney markers, Race, Age, Sex, and BMI are associated with large drops in the APS and AUROC. We see that predicting next day mortality when given shorter segments of patients’ hospital stay up to day x, where x < 10 days, on some days has higher APS when certain variables are ablated (such as vitals, kidney markers, cardiac, and BMI) which have a high importance when given information regarding more days of patients’ hospital stays. This could also be due to the higher uncertainty and the random process by which the model picks important predictors.

In Figure 7, we chose 2 COVID-19 patients with length of stay 8 days from the test cohort to look at more closely. One is a survivor who initially has a high mortality risk, which decreases, and the other is a non-survivor who starts with low mortality risk, which increases as we provide the main model (not ablating any variable besides ICU) with more days of each patient’s hospital stay (represented as green). We then repeat the mortality risk prediction for these particular patients using the category ablation models when given patient information up to day x, where x is between 1 and 8.

![Figure 7](image_url)

**Figure 7.** The predicted probability of death for a COVID-19 survivor and non-survivor, when given information for up to their x days of hospital stay where x ranges from 1 to 8. Each line corresponds to a different model, the color represents which category of variables has been ablated. The green line ("None") corresponds to the main model where no variable has been ablated.

The non-survivor is a 73-year-old male with a normal BMI who presents with dyspnea and no significant past medical history. On arrival, the vital signs were recorded as blood pressure of 168/79mmHg, heart rate of 69bpm, respiratory rate of 17bpm, and O2 saturation of 95%. The initial lab showed normal WBC of 6.3K/uL, mildly elevated creatinine of 1.45mg/dL, normal liver enzymes, and mild inflammation (CRP 52.18mg/L). These positive factors at the beginning potentially account for the main model’s (green line) low initial mortality risk. Throughout the patient’s hospital stay, his inflammation increases (CRP goes up to 133mg/L, WBC increases to 17.1K/uL), the kidney damage slightly increases as creatinine increases to 1.66mg/dL, and the patient becomes hypoxic (O2 saturation goes down to 87%, and respiratory rate goes up to 37bpm). From Figure 7, we can see that a model that does not consider vitals continues to predict a low mortality risk for the patient, as it misses to account for this patient’s hypoxia. Interestingly, a model that ablates cardiac markers (e.g., troponin) also continues to predict a low mortality risk for this patient. Out of all cardiac markers, this patient only had available measurements for troponin. This patient had his troponin measured two times, both of which came out to be normal (< 0.03). In this patient, perhaps it is not only the troponin measured values but the frequency at which it is measured that serves as the marker for the main model that this patient is critical, since troponin is more often measured in critical patients. A model that ablates race also continues to predict a low mortality risk for this patient, perhaps since certain changes in different race groups are more significant than others (for example an increase in serum creatinine may be associated with a higher increase in mortality risk in Asians than in Whites). Two other models that falsely categorize this patient (mortality risk < 0.5) are models that ablate inflammatory markers or facility. As described, this patient had an increased inflammation throughout his stay.
Furthermore, this patient was treated in the Brooklyn facility which generally has less resources than the Manhattan facilities, therefore missing to account for this could have led to a false prediction for this patient.

The survivor is a 74-year-old male with an obese BMI, who presents with dyspnea and an acute kidney injury on a chronic kidney disease (CKD). On arrival, the vital signs were recorded as blood pressure of 103/71mmHg, heart rate of 105bpm, respiratory rate of 19bpm, and O2 saturation of 93%. The initial lab showed a leukocytosis (WBC: 36K/ul), with lymphopenia, and kidney injury (serum creatinine 2.29mg/dL). This coupled with the past medical history of CKD put the patient at high next day mortality risk. However, throughout patient’s hospital stay, his O2 saturation increased and stabilized at 95-96%, creatinine level decreased to 1.27mg/dL, leading to an improved kidney function, and the inflammation decreased (WBC went down to 8.7K/ulL). These improvements possibly account for the model’s downward mortality risk trend as more days of this patient’s information is provided. For this patient, a model which ablates kidney markers (e.g., creatinine, glucose, and sodium) misses to account for the improvement in this patient’s kidney condition, which is quite important in this patient due to his history of CKD. Therefore, the kidney ablation model predicts the mortality risk to remain high due to missing to account for patient’s kidney condition improvement. A model that ablates sex and therefore does not distinguish between a male and a female also does not have a mortality risk below 0.5, therefore categorizes this patient incorrectly, perhaps due to its inability to distinguish different degrees of improvements in females and males. By these examples, we display how a particular feature (cardiac) that can have a high importance in one patient (non-survivor in this case), can be not as important in another patient (the survivor), and this calls for an algorithm that can decipher which parameters are important in each patient.

Discussion

Although many mortality-risk models have been recently developed for Covid-19 positive patients, they mostly use static data at admission. In this paper, we developed a dynamically updating mortality risk prediction approaches by incorporating time-varying information about the patients. Compared to a random forest survival model that solely uses patient information at admission, incorporating time-varying information significantly improves the model’s performance. Furthermore, compared to a random forest model that incorporates time-varying trajectories fit on the same dataset, the use of the transformer-based BEHRTDAY architecture improved the average precision score.

It offers a transformer-based prediction algorithm providing the next day mortality risk for each patient based on their previous lab measurements and demographics information since admission. Its results suggest that it has high APS and AUROC for next day mortality prediction in COVID-19 patients, and through masking, it understands each variable’s context and associates each category of variables together. It can be potentially used by clinicians to identify patients at immediate risks, and identify factors contributing to their increased risk of next day mortality for implementing better individualized treatments. The improvement in performance comes at a very small necessary fine-tuning cost (given the resources to run a deep learning model), therefore can be implemented to make use of time series data in COVID-19 positive mortality prediction models.

A few limitations of BEHRTDAY’s its complexity and running time. To reduce complexity, we had to categorize each variable into a few categories to avoid having a large dictionary for the model. Even then, due to its sequence size of about 3000 words (about 100 variables for 30 days), the MLM process had an average running time of about 4 hours on JADE (Oxford’s powerful computational system, using multiple cores/GPUs), and the next day mortality prediction has an average running time of about 3 hours per train dataset.

The study dataset has several limitations. Different hospitals in the Mount Sinai System had different protocols when it came to COVID-19 positive patients, therefore leading to in-consistent frequencies at which certain labs (i.e., troponin or D-Dimer) were measured. Furthermore, it should be considered that each hospital has different resources, and when performing modeling, the “Facility” variable is trying to account for such differences. For some measurements (e.g., creatinine kinase), we have a lot of missing values, therefore future work needs to be done to validate our approach. Moreover, the COVID-19 negative dataset includes patients who were admitted to the hospital during the earlier months of the pandemic in New York, thus are likely only representative of hospital admissions during this stage of the pandemic.

Conclusion

In conclusion, we developed an interpretable deep learning model for dynamic COVID-19 mortality prediction from a dataset encompassing 5 hospitals of 3699 COVID-19 patients. The predictive performance improved by including time series data and deep learning in BEHRTDAY. The model’s interpretation showed that input features interact and compensate for one another by being associated with a higher or lower mortality risk in patients based on patients’ other circumstances. By successfully predicting the next day mortality and identifying features with high-importance
values, this modeling may benefit healthcare institutions for improving care decisions and allocating their resources more efficiently for their COVID-19 patients.

Acknowledgement

This work was partially supported by the AXA Research Fund. We also acknowledge the use of the UK EPSRC-funded Tier 2 facility JADE (EP/P020275/1).

References

Using Shapes of COVID-19 Positive Patient-Specific Trajectories for Mortality Prediction

Alaleh Azhir, MSc¹, Soheila Talebi, MD², Louis-Henri Merino, BSc³, Thomas Lukasiewicz, PhD¹, Edgar Argulian, MD, MPH², Jagat Narula, MD, PhD², Borislava Mihaylova, MSc, DPhil²

¹University of Oxford, Oxford, UK; ²Icahn School of Medicine at Mount Sinai, New York, NY, USA; ³Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland; ⁴Queen Mary University of London, London, UK

Abstract

Machine learning can be used to identify relevant trajectory shape features for improved predictive risk modeling, which can help inform decisions for individualized patient management in intensive care during COVID-19 outbreaks. We present explainable random forests to dynamically predict next day mortality risk in COVID-19 positive and negative patients admitted to the Mount Sinai Health System between March 1st and June 8th, 2020 using patient time-series data of vital, blood and other laboratory measurements from the previous 7 days. Three different models were assessed by using time series with: 1) most recent patient measurements, 2) summary statistics of trajectories (min/max/median/first/last/count), and 3) coefficients of fitted cubic splines to trajectories. AUROC and AUPRC with cross-validation were used to compare models. We found that the second and third models performed statistically significantly better than the first model. Model interpretations are provided at patient-specific level to inform resource allocation and patient care.

Introduction

With the rise of the Delta variant of Covid-19 in June 2021 and Omicron in December 2021, the situation in various countries such as Iran, is to a degree reminiscent of the surge of cases back in mid-April 2020 in New York. The allocation of resources to those most likely to survive is an important and difficult decision that healthcare workers face every day in the pandemic. Covid-19 involves an intricate interplay of various interdependent biological pathways and despite epidemiological and clinical characteristics of patients with COVID-19 in various parts of the world having been reported¹,², a dynamically updated assessment of patients’ disease progression over time is lacking. Monitoring the changes of these measurements over time and dynamically assessing mortality risks for Covid-19 patients could aid in a better allocation of resources.

The increase in automating data extraction from electronic health records has spurred recent efforts to use machine learning to predict individual patient’s mortality risk using dynamic time-series data.³,⁴ Although scoring systems for ICU populations, such as the Simplified Acute Physiology Score (SAPS)³, the Acute Physiologic and Chronic Health Evaluation (APACHE)⁵, and the Mortality Prediction Model⁶ for ICU populations can be useful in predicting mortality risk, they are usually calculated based only on data collected at admission and lack precision.³,⁵ Using the information that is captured in patient’s measurements during hospital stay, including changes in patient indicators overtime could improve mortality risk scoring.

Previous studies have suggested the use of summary statistics of trajectories (maximum, minimum, median, etc.)³ or fitting cubic spline and extracting relevant features⁴ as inputs for mortality prediction as possible approaches to address this limitation. In this paper, we demonstrate that, compared to including the most recent measurements, each among the use of summary statistics and the automated extraction of cubic spline coefficients by random forest classifiers statistically improves both the area under the receiver operating characteristic (AUROC) curve and the area under the precision- recall curve (AUPRC) of predicting the mortality risk for patients with COVID-19. Such approaches can be used to improve any risk model that contains time series data. Furthermore, we improve on these previous studies by providing a patient-specific model interpretation to facilitate random forest model adoption in a health care setting.

Methods

Data Preprocessing: We retrospectively retrieved de-identified health records of patients tested for Covid-19 admitted to five hospitals within the Mount Sinai Health System (MSSH) between March 1st and June 8th, 2020; approved by the Institutional Review Board at the Icahn School of Medicine at Mount Sinai. The data included patient demographics, past medical history, daily vitals, and all cardiac, liver panel, blood test, and other labs during their
hospital stays (some displayed in Table 1). Relevant patient events and outcomes (ICU admission during patient stay, COVID-19 positive or negative test, and outcome: discharge to home, discharge to hospice care, or death) were recorded. Patients with multiple hospitalizations were removed from the dataset. To remove questionable laboratory measurements, cut off values for ranges as found in the literature were applied for all continuous measurements, and the measurements outside of these ranges were set as missing.

Dataset Generation: To use time for mortality predictions, for each patient encounter, we generated (overlapping) observational time period units every day, where each unit has a maximum size of 7 days. For example, a patient hospital stay that lasted 9 days generates the following observational units: (0,1] day, (0,2] days, ..., (0,7] days, (1,8], and (2, 9] days. For each observational unit, the outcome of interest is death within the next day of the end of the unit’s time period.

Risk Modeling: We adapted the models presented by Ma et al. to fit our Covid-19 cohorts. Different risk factors in each observational unit can be fed into any classifier to estimate the next day probability of mortality. We used Ranger, a fast implementation of a random forest classifier, due to its superior performance as a nonparametric method. Random forests not only avoid overfitting but can also handle large, including dependent, data with numerous features, as included in this dataset. Our main hyperparameters are number of trees and the number of features to be considered for each tree (mtry). We found that 500 trees provided the lowest out-of-bag error rates while not compromising the computational time, and fine-tuned mtry for each of our models using the R Caret package. Three nested random forest models were applied to each observational unit to assess the use of trajectory data in next-day mortality prediction. They are defined below:

Model 1: This model includes static at admission patient information (e.g., age) and last recorded value of each longitudinally measured risk factor.

Model 2: This model appends the summary characteristics of longitudinal risk factor trajectories (minimum, first, last, median, maximum, and count) for each predictor to the information included in Model 1.

Model 3: This model replaces the summary characteristics in Model 2 with the cubic spline coefficients after fitting penalized smoothing splines using Leave-One-Out cross validation.

Since Model 3 fits a cubic spline, to avoid overfitting, we only fitted and compared Model 3’s performance to those of Models 1 and 2 in patients whose length of stay exceeded 4 days. As the range for non-missing values across all predictors was bounded away from -10, we encoded missing values with this number to instruct our tree-based models to treat them differently.

When compared to Ma et al. paper, in addition to having a different time frame and window (7 days vs. 24 hours), we included observational units that were shorter than our window size of 7 days, whereas they only include patients that had a stay longer than their window size of 24 hours. Furthermore, in spline fitting for our Model 3, the number of the knots that we chose for smoothing splines was 4 instead of 27, since we had fewer daily measurements.

Model Evaluation: ROC curves are commonly used for model evaluation, however, can be misleading when outcomes (death vs. discharge to home) are highly imbalanced, therefore, we also considered the precision-recall curves. We used Monte Carlo cross validation to evaluate and compare the performance of our 3 models, by performing 20 random splits of observational units into training (70%) and validation (30%) sets. For each split, we fit a ranger random forest to the training set, predict the probability of death in the validation set, and evaluate the AUPRC and AUROC. We used the nonparametric Wilcoxon paired signed-rank test to compare AUROC and AUPRC between our three models. Though the baseline value for AUROC is 0.5 (for a random coin toss), the baseline value for AUPRC depends on the fraction of positives in the dataset. For example, in our case, since 3.4% of Covid-19 positive observational units resulted in death, our baseline AUPRC value is 0.034. The R pROC package was used to compute the 95% confidence interval (CI) of the sensitivity at the given specificity points, with 2000 stratified bootstrap replicates. We replicated the above analysis among COVID-19 negative patients.

Model Interpretation: We applied Shapley additive explanations (SHAP) algorithm to our best performing model to obtain explanations of the risk factor features that drive patient-specific predictions. The SHAP value for a feature does not specify its direct isolated effect but its compound effect while interacting with other features in the model.

Results

Our study includes 3699 COVID-19 positive and 5488 COVID-19 negative patients, admitted between March 1st and June 8th 2020. Patients’ median age is 67 years for COVID-19 positive and 59 years for COVID-19 negative patients. 57% of the COVID-19 positive are male and 45% of the COVID-19 negative are male. The overall median BMI is 27.0 (27.7 for COVID-19 positive, 26.5 for COVID-19 negative). The median length of hospital stay is 5 days. Laboratory markers (such as troponin, WBC count, etc.) were categorized and tracked during the first month of patients’ admission. Selected characteristics of the study patients are presented in Table 1.

Table 1: Characteristics of the hospitalized COVID-19 positive and negative patients. Continuous variables are presented as mean (interquartile (IQR) range), and categorical variables are presented as count (% percentage).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Survivors</th>
<th>Non-survivors</th>
<th>Survivors</th>
<th>Non-survivors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Covid-19 Positive</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex, Female</td>
<td>1120 (44 %)</td>
<td>480 (42 %)</td>
<td>2776 (55 %)</td>
<td>241 (52 %)</td>
</tr>
<tr>
<td>Missing</td>
<td>0 (0 %)</td>
<td>0 (0 %)</td>
<td>2 (0.0 %)</td>
<td>0 (0 %)</td>
</tr>
<tr>
<td>Age, years</td>
<td>63 (51 - 73)</td>
<td>75 (65 - 84)</td>
<td>58 (35 - 73)</td>
<td>75 (61 - 85)</td>
</tr>
<tr>
<td>BMI</td>
<td>28 (24 - 33)</td>
<td>27 (24 - 33)</td>
<td>27 (23 - 31)</td>
<td>25 (21 - 31)</td>
</tr>
<tr>
<td>Underweight</td>
<td>80 (3 %)</td>
<td>56 (5 %)</td>
<td>401 (8 %)</td>
<td>43 (9 %)</td>
</tr>
<tr>
<td>Normal</td>
<td>658 (26 %)</td>
<td>299 (26 %)</td>
<td>1516 (30 %)</td>
<td>170 (37 %)</td>
</tr>
<tr>
<td>Overweight</td>
<td>790 (31 %)</td>
<td>324 (28 %)</td>
<td>1596 (32 %)</td>
<td>107 (23 %)</td>
</tr>
<tr>
<td>Obese</td>
<td>883 (35 %)</td>
<td>368 (32 %)</td>
<td>1413 (28 %)</td>
<td>124 (27 %)</td>
</tr>
<tr>
<td>Missing</td>
<td>134 (5.3 %)</td>
<td>107 (9.3 %)</td>
<td>100 (2.0 %)</td>
<td>18 (3.9 %)</td>
</tr>
<tr>
<td><strong>Smoking Status</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>1397 (55 %)</td>
<td>533 (46 %)</td>
<td>2575 (51 %)</td>
<td>207 (45 %)</td>
</tr>
<tr>
<td>Previous or Current Missing</td>
<td>635 (25 %)</td>
<td>327 (28 %)</td>
<td>1712 (34 %)</td>
<td>168 (36 %)</td>
</tr>
<tr>
<td><strong>Race and Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>711 (28 %)</td>
<td>289 (25 %)</td>
<td>1355 (27 %)</td>
<td>113 (24 %)</td>
</tr>
<tr>
<td>American Indian or Alaska Native</td>
<td>2 (0 %)</td>
<td>0 (0 %)</td>
<td>3 (0 %)</td>
<td>0 (0 %)</td>
</tr>
<tr>
<td>Asian</td>
<td>121 (5 %)</td>
<td>52 (5 %)</td>
<td>265 (5 %)</td>
<td>21 (5 %)</td>
</tr>
<tr>
<td>White</td>
<td>543 (21 %)</td>
<td>314 (27 %)</td>
<td>1614 (32 %)</td>
<td>169 (37 %)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>743 (29 %)</td>
<td>295 (26 %)</td>
<td>1076 (21 %)</td>
<td>79 (17 %)</td>
</tr>
<tr>
<td>Native Hawaiian or Pacific Islander</td>
<td>3 (0 %)</td>
<td>1 (0 %)</td>
<td>8 (0 %)</td>
<td>0 (0 %)</td>
</tr>
<tr>
<td>Other</td>
<td>347 (14 %)</td>
<td>158 (14 %)</td>
<td>549 (11 %)</td>
<td>61 (13 %)</td>
</tr>
<tr>
<td>Unknown</td>
<td>75 (3 %)</td>
<td>45 (4 %)</td>
<td>156 (3 %)</td>
<td>19 (4 %)</td>
</tr>
<tr>
<td><strong>Hospital Location</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brooklyn</td>
<td>372 (15 %)</td>
<td>258 (22 %)</td>
<td>475 (9 %)</td>
<td>71 (15 %)</td>
</tr>
<tr>
<td>Queens</td>
<td>302 (12 %)</td>
<td>261 (23 %)</td>
<td>430 (9 %)</td>
<td>56 (12 %)</td>
</tr>
<tr>
<td>Manhattan – St. Luke’s</td>
<td>515 (20 %)</td>
<td>218 (19 %)</td>
<td>879 (17 %)</td>
<td>87 (19 %)</td>
</tr>
<tr>
<td>Manhattan – West</td>
<td>363 (14 %)</td>
<td>93 (8 %)</td>
<td>1004 (20 %)</td>
<td>66 (14 %)</td>
</tr>
<tr>
<td>Manhattan – East</td>
<td>993 (39 %)</td>
<td>324 (28 %)</td>
<td>2238 (45 %)</td>
<td>182 (39 %)</td>
</tr>
<tr>
<td><strong>Comorbidities</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asthma</td>
<td>137 (5 %)</td>
<td>52 (5 %)</td>
<td>352 (7 %)</td>
<td>26 (6 %)</td>
</tr>
</tbody>
</table>
In total, there were 33,864 observational units for the COVID-19 positive patients and 32,589 observational units for the COVID-19 negative patients. 3.4% and 1.4% of all COVID-19 positive and negative observational units resulted in death, respectively.

We then divided the study patient population into two groups, those with shorter length of stay (≤ 4 days), and those with longer length of stay (> 4 days). Since most of repeated risk factor measurements (besides vital signs) are only measured daily, to avoid overfitting, we only compared Models 1 and 2 for patients with length of stay up to 4 days. As expected, among these patients due to shorter trajectories, only a slight improvement in performance for Model 2 compared to Model 1 is observed (Figure 1); both models, however, have a high predictive power.

The ROC curves and precision-recall curves (PRC) of the three next-day mortality risk models for COVID-19 positive patients with length of stay above 4 days, including comparisons of the AUROC and AUPRC for the next day mortality averaged over the 20 random data splits are presented (Figure 2). Models 2 and 3 have a better precision than Model 1 for both COVID-19 positive and negative patients (all p-values < 0.01). The AUROC of Model 2 is larger than that of Model 1 (p-values < 0.0001). The AUROC of Model 3, however, did not differ significantly from that of Model 1. Furthermore, comparing Models 2 and 3, Model 2’s AUROC is significantly larger than that of Model 3 in both COVID-19 positive and negative patients (both p-values < 0.001). However, Model 2’s AUPRC is only statistically significantly higher than that of Model 3 in COVID-19 positive patients. This suggests that Model 2, by using key statistical summaries of trajectories (minimum, maximum, median, first, and last), performs better than fitting cubic splines (Model 3) and conveys additional predictive information that is not captured by the most recent measurements alone (Model 1). A similar analysis confirmed the superiority of Model 2 to Models 1 and 3 in Covid-19 negative patients. All three models, however, performed significantly better than a random forest baseline survival model fitted only on initial patient admission data (95% AUROC CI index: 0.850 - 0.857, 95% AUPRC CI index: 0.380 - 0.400).

<table>
<thead>
<tr>
<th>Factor</th>
<th>COVID-19 Positive</th>
<th>COVID-19 Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>COPD</td>
<td>86 (3 %)</td>
<td>58 (5 %)</td>
</tr>
<tr>
<td>Sleep Apnea</td>
<td>48 (2 %)</td>
<td>29 (3 %)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>487 (19 %)</td>
<td>279 (24 %)</td>
</tr>
<tr>
<td>Chronic Kidney Disease</td>
<td>257 (10 %)</td>
<td>154 (13 %)</td>
</tr>
<tr>
<td>Cancer</td>
<td>188 (7 %)</td>
<td>102 (9 %)</td>
</tr>
<tr>
<td>Coronary Artery Disease</td>
<td>254 (10 %)</td>
<td>168 (15 %)</td>
</tr>
<tr>
<td>Atrial Fibrillation</td>
<td>113 (4 %)</td>
<td>106 (9 %)</td>
</tr>
<tr>
<td>Heart Failure</td>
<td>153 (6 %)</td>
<td>110 (10 %)</td>
</tr>
<tr>
<td>Chronic Viral Hepatitis</td>
<td>29 (1 %)</td>
<td>8 (1 %)</td>
</tr>
<tr>
<td>Liver Disease</td>
<td>57 (2 %)</td>
<td>36 (3 %)</td>
</tr>
<tr>
<td>ARDS</td>
<td>13 (1 %)</td>
<td>11 (1 %)</td>
</tr>
<tr>
<td>Acute Kidney Injury</td>
<td>135 (5 %)</td>
<td>105 (9 %)</td>
</tr>
<tr>
<td>Acute venous thromboembolism</td>
<td>26 (1 %)</td>
<td>7 (1 %)</td>
</tr>
<tr>
<td>Cerebral Infarction</td>
<td>31 (1 %)</td>
<td>8 (1 %)</td>
</tr>
<tr>
<td>Intracerebral hemorrhage</td>
<td>5 (0 %)</td>
<td>2 (0 %)</td>
</tr>
<tr>
<td>Acute MI</td>
<td>32 (1 %)</td>
<td>25 (2 %)</td>
</tr>
<tr>
<td>Length of Stay</td>
<td>6.9 (4.0 - 11)</td>
<td>8.8 (5.0 - 15)</td>
</tr>
<tr>
<td>ICU (yes: anypoint during admission)</td>
<td>316 (12 %)</td>
<td>521 (45 %)</td>
</tr>
</tbody>
</table>
Figure 1. Average receiver operating characteristics and precision (%) versus recall (%) plot of Models 1 (most recent measurements only) and 2 (trajectories’ summary statistics) of next day mortality risk among Covid-19 positive and negative patients with length of stay ≤ 4 days. AUROC and AUPRC are represented as mean (standard error) for Model 1 column and as mean (standard error); p-value for Model 2 column. p-values are obtained a two-sided Wilcoxon paired signed rank test comparing Model 2 to Model 1 for patients with length of stay ≤ 4 days.

To understand which features are most important in the best performing model (Model 2) among COVID-19 positive patients with length of stay above 4 days, both the SHAP values and Gini Importance values of the top features are summarized (Figure 3). This reveals that lower O2 saturation, lower blood pressure, older age, higher anion gap, and higher white blood cell count are associated with a higher probability of death. However, certain features such as higher troponin level, respiratory rate, and heart rate can be associated with a higher or lower probability of survival depending on patients’ other risk factors. The importance values of predictors were also calculated in COVID-19 negative patients, and the top predictors in those patients were vitals, followed by length of stay followed by an intermix of kidney and complete blood count markers.

To illustrate the influence of features of risk factors, Model 2 predictions are presented for 2 COVID-19 patients from our test cohort with the same sex and age who were admitted to the ICU and followed for up to 22 days. Figure 4 displays features that contributed to the mortality risk prediction for these two patients evaluated using Model 2 after 1, 2 and 3 weeks in hospital. The first case is a 66-year-old male with a history of hypertension, systolic heart failure, atrial fibrillation, and end-stage renal disease with BMI 40 presented with dyspnea. On admission, the vital signs were recorded as blood pressure of 150/100 mmHg, heart rate 140bpm, respiratory rate of 38bpm, and O2 saturation of 91%. The initial lab was significant for leukocytosis with lymphopenia, increased coagulation factors, abnormal troponin, and increased liver enzyme (WBC: 32K/ul, neutrophil: 93%, lymphocyte: 0.6%, platelet: 277K/ul, AST: 93U/L, albumin: 2.9g/dl, troponin: 1.19ng/ml, creatinine: 1.84mg/dl, potassium: 5.6mmol/L, calcium: 8.5mg/dl, LDH: 2212U/L, CRP: 419mg/L, ferritin: 393ng/ml, and D-Dimer: 12.9ug/mL). He had a worsening of liver function and increased LDH and ferritin level. After a hospital stay of 21 days, he was discharged to rehab. In this case, the patient’s mortality risk decreases over time from 0.06 to 0.01. Predicting death after 21 days in the hospital, glucose median of 68.5, last recorded creatinine of 3.69, and min respiratory rate of 25 drive the mortality prediction towards...
non-survival, whereas the last recorded systolic blood pressure, O2 saturation, neutrophil, and heart rate are the most important feature pulling the prediction towards survival.

The second case is a 66-year-old male with BMI 29 and no significant past medical history, who presented with dyspnea. On arrival, the vital signs were recorded as blood pressure 130/80mmHg, heart rate 83bpm, respiratory rate 29bpm, and O2 saturation of 91%. The initial lab revealed a leukocytosis with lymphopenia, kidney injury, mild increase of liver enzyme, and inflammatory factors (WBC: 8.1K/uL, neutrophil: 89.75%, lymphocyte: 1.6%, platelet: 191K/uL, AST: 43U/L, albumin: 2.55g/dl, troponin: 0.013ng/ml, creatinine: 1.84mg/dL, potassium: 5mmol/L, calcium: 8.5mg/dL, LDH: 458U/L, CRP: 298mg/L, ferritin: 1063ng/ml and D-Dimer: 0.85ug/mL). During hospital stay, the lowest oxygen saturation was 80% and mild increase of creatinine, D-Dimer, and liver enzyme were observed. He died in ICU 22 days after admission. In this case, we can see that this patient’s predicted mortality risk increases over time from 0 to 0.25. Predicting survival after 21 days in the hospital, maximum lymphocyte, age, and minimum basophil pull the risk down towards survival (blue arrows), whereas last systolic and diastolic blood pressure, PCO2 venous, BUN, and minimum O2 saturation drive mortality prediction towards non-survival (red arrows). Such dynamic risk prediction by visualizing the main features contributing to non-survival at any point in time can aid clinicians determine whether there are actions that can be taken to lower the mortality risk.

**Discussion**

Although many mortality-risk models have been recently developed for Covid-19 patients, they mostly use static data at admission. In this study, we developed a dynamic next-day mortality risk prediction model incorporating time-updated patient information on the past week’s lab measurements. Our results suggest that trajectories’ summary statistics significantly improve prediction performance compared to including most recent measurements or fitted cubic spline coefficients. The model’s high classification power (AUROC = 0.902, AUPRC = 0.450), suggest that it can be used by clinicians to identify patients at immediate risk, and identify factors contributing to their increased risk of next day mortality for implementing better individualized treatments. The method was also beneficial in improving
the next day mortality prediction in COVID-19 negative patients. Lastly, given the evolving COVID-19 landscape, it is important to re-train these models on more recent COVID-19 patient datasets and externally validate them in other hospitals. Due to its small computational cost and public code availability, this method can be easily implemented and continuously updated using new patients’ time series data to improve mortality prediction among hospitalized patients.

Figure 3. Model 2 (with trajectories’ summary statistics): Gini Importance values for top 30 features (left) and SHAP values for the top 20 features (right) for COVID-19 positive patients. The Gini Importance values were calculated by Model 2 random forest model. SHAP values for the top 20 features calculated by sum of SHAP value magnitudes over all Covid-19 positive observations. Higher SHAP values indicates higher probability of death. All patients in the dataset are run through Model 2, and a dot is created for each person on each feature’s line.

Whereas our results found that trajectories’ summary statistics performs best, Ma et al.’s paper found fitting cubic spline coefficients performs best.\(^4\) The differences between our findings can be due to most of our laboratory factors (besides vitals) being measured only once daily, thus not providing as many datapoints for fitting a spline as are present in their dataset. Furthermore, we also include observational units that were shorter than our maximum observation unit size of 7 days, whereas they only looked at observational units of their maximum size to obtain longer trajectories. These differences in model design could have led to different findings.

Our approach also allows for a clinically interpretable understanding of its top features driving mortality for both an individual patient and the entire dataset. Whereas prior studies reported different outcomes associated with high temperature,\(^17\)–\(^19\) our study shows that although the non-survivor group has a statistically significantly higher temperature both at admission and during their stay compared to the survivor group, the average daily temperature for the non-survivor group still falls below the fever cutoff (100.4°F). O2 saturation was identified by Gini and SHAP importance scores to be the most important feature overall in predicting next day mortality, and as suggested by previous studies, is an important indicator for guiding physicians when to require admission to the ICU.\(^20,21\)

Full Blood Count trajectories displayed more frequent leukocytosis, neutrophilia, and lymphopenia in non-survivors. Neutrophilia, as an expression of hyperinflammatory state, may also indicate a superimposed bacterial infection.\(^22,23\) Lymphopenia could occur potentially due to high ACE2 receptor expression on lymphocytes causing increased susceptibility to Covid-19,\(^24\) cytokine storms causing lymphocytes’ apoptosis,\(^25\) or injured alveolar epithelial cells inducing the infiltration of lymphocytes.\(^26,27\)

Elevated troponin was observed in non-survivors and was deemed as one of the top predictors of next day mortality by SHAP and Gini. This could be secondary to cytokine storm, myocarditis, pro-thrombotic state, or demand ischemia, which all may contribute to the observed poor prognosis.\(^28\)–\(^30\)
Figure 4. Force plots obtained using SHAP values for 2 COVID-19 positive patients in a test dataset, one survivor and one non-survivor. SHAP values identify risk factor features associated with higher (red color) or lower (blue color) mortality risk.

Despite observing elevated LDH levels in both survivors and non-survivors (as seen in our two example cases), a significantly higher LDH level was observed on average throughout patients’ admission in non-survivors. The elevation of LDH has been reported as an independent factor of mortality in patients with severe acute respiratory syndrome.\textsuperscript{31,32} Similar to past studies,\textsuperscript{33} on average and Gini Importance plot, we also observed higher AST and ALT levels in our non-survivors on admission and throughout their stay, reflecting liver damage. Though the mechanism remains unclear, various hypothesis suggest direct SARS-CoV-2 infection of liver cells, drug-induced toxicity, immune-mediated inflammation, or hypoxia.\textsuperscript{33–35}

Our study concurs with the growing evidence of Covid-19-induced hypercoagulable states,\textsuperscript{36,37} as we observed increased levels of D-Dimer in non-survivors on average throughout the first two weeks of admission. On the other hand, Fibrinogen despite being high in both survivors and non-survivors, was not significantly different through patients’ admission between these 2 groups. Regarding thrombocytopenia, although platelet counts fell within the normal range, non-survivors had significantly lower platelet counts during their stay.

We observed significantly lower levels of calcium in non-survivors, with a drop in average calcium levels in those whose length of stay exceeded 10 days. Furthermore, calcium level had high Gini and Shap importance in next-day mortality prediction. Previous studies have reported that calcium interacts with fusion peptides on viruses such as SARS-CoV, MERS-CoV, and Ebolavirus, promoting their replication.\textsuperscript{38–40} Furthermore, it has been shown to be an independent risk factor in Covid-19 hospitalization.\textsuperscript{41} A high creatinine and BUN level, and electrolyte abnormality including hyperkalemia, hyperchloremia, acidosis, and hypernatremia were seen more frequently in non-survivors compared to survivors during the first 2 weeks of stay.
Our study has several limitations. Most laboratory factors (besides vitals) were only measured daily, hence not providing a long enough time series for fitting a cubic spline trajectory (Model 3) within a few days. To overcome this limitation, we compared only the use of summary statistics of trajectories to updating the most recent value in patients with hospital length of stay ≤ 4 days and considered comparing fitting a cubic spline to these two methods only in patients with length of stay longer than 4 days. Therefore, our conclusions regarding Model 3 is restricted to patients with longer hospital stays. In addition, we only looked at observational units with a maximum size of 7 days, longer units could be explored in a future study. Furthermore, different hospitals used different protocols in treating COVID-19 positive patients, therefore leading to different frequencies at which labs (i.e., troponin or D-Dimer) were measured. As noted in Table 1, for some measurements, we have a lot of missing values, therefore future work needs to be done to validate the model in further cohorts. Moreover, the Covid-19 negative dataset includes patients who were admitted to the hospital during the severe months of the pandemic in New York, thus are only representative of hospital admissions during the pandemic. Finally, further features of trajectories (curvature, arc length, etc.) may also prove informative for risk assessment and could be assessed in further studies.

**Conclusion**

In conclusion, we developed an explainable random forest model for dynamic next-day mortality prediction in COVID-19 positive patients using a dataset of 3699 COVID-19 positive and 5488 COVID-19 negative patients in 5 hospitals. The model interpretation showed that risk factors interact and compensate for one another by pulling patients towards survival or non-survival based on patients’ other characteristics. By improving the prediction of the next day mortality and identifying features with high-importance values, this new model may help healthcare institutions improve care decisions for COVID-19 positive admitted patients.

**References**

Automated Dental Cavity Detection System Using Deep Learning and Explainable AI

Niharika Bhattacharjee, University of Illinois at Urbana Champaign, Urbana, Illinois

Abstract

Impacting over 3.9 billion people, dental cavities require a trained dentist for diagnosis. Unfortunately, barriers such as dentophobia, limited dentist availability, and lack of dental insurance prevent millions from receiving care. To address this, an Artificial Intelligence system was developed that detects cavity presence on photographs and visually explains the rationale behind each diagnosis. While previous systems only detected cavities on one extracted tooth showing one tooth surface, this study’s system detects cavities on photographs showing multiple teeth and four tooth surfaces. For training, 506 de-identified images from online sources and consenting human participants were collected. Using curriculum learning, a ResNet-27 architecture proved to be the most optimal after achieving 82.8% accuracy and 1.0 in sensitivity. Visual explanations for the system’s diagnoses were also generated using Local Interpretable Model Agnostic Explanation. This system can explain its diagnoses to users in an understandable manner, which is a crucial skill employed by dentists.

Introduction

Dental cavities rank as the most prevalent chronic disease worldwide [1]. While cavity detection requires the services of a trained dentist, 1 in 3 Americans don’t visit a dentist annually due to cost or dentophobia [2]. Additionally, chances of proper cavity diagnoses are lowered in developing nations, as the dentist to patient ratio is disproportionate (i.e., 1: 1,250,000 in Ethiopia [3]). Due to various financial and infrastructural barriers, millions lack proper dental care. As a result of limited dental resources, 2.4 billion people also suffer from untreated tooth decay, which can lead to toothaches and periapical abscesses [4]. Previous studies have tried to automate cavity detection by developing deep learning (DL) classifiers that detect cavities on X-ray images [5-6]. Unfortunately, these systems aren’t accessible to many worldwide as X-ray images can cost over $180 per series and require a visit to the dentist [7]. Furthermore, aside from dental professionals, most people aren’t familiar with X-ray images. As a result, even if X-Ray imaging was made accessible at much lower costs, end-users may also have a difficult time trusting a system’s diagnosis that utilize such images that they can’t interpret themselves. To address the above limitations, we have developed a DL diagnostic system that detects cavity presence on regular photographs that can be captured by mobile devices. By using such photographs, this system can be used worldwide as over 3 billion people have built-in cameras in their smartphones of which 45% are from emerging economies [8]. Aiming to also gain users’ trust, the system utilizes state-of-the-art explainability algorithms to provide users visual explanations that highlight the rationale it took while making each diagnosis.

The most similar and relevant study to our work is the one reported by Berdoues et al. [9]. Using a random forest algorithm, they detected and highlighted cavity presence on photographs showing only a single tooth. In practice, however, users would provide images showing multiple teeth. Berdoues et al.’s system was also trained to detect cavities only on the occlusal tooth surface. Our system is not only capable of detecting cavities on photographs showing multiple teeth, but even goes two steps further by detecting cavities on three more tooth surfaces in addition to providing user-friendly visual explanations.

In the following sections of this paper, we explore previous research, identify related products available, discuss the benefits of explainable AI, and share the proposed methodology. We then conclude with our reported findings and a discussion highlighting future research and system’s application in developing nations.

Related Work

Technological and Financial Limitations of Current Diagnostic Methods. While diagnosing cavities, dentists often utilize x-ray machines or optical diagnostic devices such as the DIAGNodent™, DEXIS CariVu™, and the Canary System™. However, there are many limitations associated with the optical diagnostic devices mentioned above [10]. Instead of detecting Streptococcus mutans (the cavity-causing bacteria which resides at the cavity), the DIAGNodent only detects fluorescence emitted by bacterial porphyrins, a common bacterium found everywhere in the mouth. The use of trans-illumination prevents the CariVu from detecting cavities on smooth surfaces and around
Visual Examination. Typically, a dentist reaches his/her diagnosis during a visual examination. During an exam, a dentist can identify a patient’s cavity, inform the patient about the cavity’s development, and suggest changes to the patient’s oral routine. As a visual examination mainly relies on a dentist’s trained eye, this method is the most cost-efficient from an equipment perspective. However, visual examination still has one requirement that prevents millions of people from receiving a proper cavity diagnosis; the method requires a dentist. Receiving a dentist’s diagnosis, in general, is difficult for many as cost, fear of pain, and lack of dental insurance pose as barriers [12]. Some of these barriers could be addressed with an AI diagnostic system that automatically detects cavity presence and provides a diagnostic explanation at lower cost and without an initial dental visit.

Previous Research in the Use of AI in Cavity Detection. AI has been applied to several problems in medical imaging. From melanoma detection [13] to breast cancer prediction [14], as well as for diabetic retinopathy diagnosis from fundus imaging [15]. However, research on the application of AI in cavity detection is still limited. The study by Berdouses et al. used Random Forest (RF) and instance segmentation to classify and highlight cavities in 103 images with 80% accuracy. The study’s dataset labelling was done using the International Caries Detection and Assessment System (ICDAS) rubric (Figure 1).

Limitation of Previous Research. While the above study by Berdouses et al. highlighted AI’s potential in the dentistry field, the images and algorithm used did not reflect field conditions. To elaborate, the study’s images were not reflective of what an operator of a cavity classifier would receive from future users. Each of their training images contained a single tooth, an assumption that is not realistic in practice as patients’ images would consist of a mouthful of teeth. Additionally, their RF algorithm was only trained with images showing the occlusal tooth surface. As a result, the algorithm was untrained in detecting cavities on labial, buccal, and lingual tooth surfaces. The study’s ML classifier also required a data scientist to physically extract each image feature needed to make a diagnosis. To address the limitations, present in Berdouses et al.’s work, our study aims to develop a more efficient system that uses deep learning to automatically detect cavity presence on photographs showing multiple teeth and several tooth surfaces (occlusal, labial, buccal, and lingual).

Benefits of Using Photographic Color Images over X-Rays. There are many studies that have trained DL classifiers to detect cavities on X-ray images [5-6]. Although X-ray images are informative, they still require a costly visit to the dentist as X-ray machines aren’t portable. For the millions of people who can’t afford a dental visit, any classifier requiring X-rays is difficult to access. In contrast, photographic images of teeth are easier to obtain as over 3 billion people worldwide have built-in cameras in their smartphones. Cheaper options are available as well, such as intraoral cameras for under $35.00. With regards to user accessibility, many aren’t familiar with what their teeth look like in an X-ray image, but everyone is familiar with what their teeth look like in a photo. Thus, users can easily interpret and trust a classifier’s diagnosis on photographic images, with the help of explainable AI techniques such as Local Interpretable Model-Agnostic Explanation (LIME).

Local Interpretable Model-Agnostic Explanation (LIME). Trust in AI is an emerging field. A popular explainer algorithm applied on image classifiers to develop trust with end-users is called LIME [16]. This algorithm promotes trust by presenting a visual explanation for a given model’s prediction. It produces the explanation by approximating the behavior of the complex model with very simple linear models (i.e. hyperplanes). For a given medical image, LIME can overlay a visual explanation on top of the image, highlighting the important regions present in the image that influenced the classification result. Such explanations help the end user understand the rationale behind

Figure 1. International Caries Detection and Assessment System (ICDAS) clinical scoring rubric used by dentists.

Figure 2. Visual shows a general overview of the LIME algorithm (Arteaga, 2019).
the model’s diagnosis (Figure 2). While LIME would be extremely effective, it has not been applied to cavity detection on photographic color images to our knowledge.

Methodology

General Overview. The following methodology was used to develop an AI diagnostic system that detects cavity presence on photographs and visually explains the rationale behind each diagnosis using an algorithm called LIME. The aim of the study was to select an ANN architecture and training approach that was optimal for detecting cavities. The architectures we experimented with included a hand-designed 12-layer Convolutional Neural Network (CNN) and various extensions of pre-trained image classifiers (ResNet-18, ResNet-27). As training data was limited, two-stage curriculum learning was applied to better train the networks. In addition, the capabilities of LIME were explored to provide visual explanations that could be easily interpreted by the end-users.

2019 Web-Searched Dataset. To train the models to detects cavities on photographic color images, 314 de-identified photographic color images showing cavitated or healthy teeth were collected. These images were taken from online sources (dental blogs, dental presentations, and journals). A dentist was then contacted to professionally diagnose/label the images based on the International Caries Detection and Assessment System (ICDAS) rubric. The cavities present in the dataset were representative of all levels of decay. This is evident as the dentist used all ICDAS class values ranging from 0 (no cavity presence) to 6 (extensive lesion) to label the dataset (Figure 3). Images labeled ICDAS class 0 (no cavity presence) were represented with label - 0, and images labeled with ICDAS classes 1-6 (cavity presence) were represented with label - 1. However, all ICDAS classes were used to analyze ANN’s performance. Using the labeling method above, out of the 314 images, 185 images had no cavities and 129 images had at least one cavity. These 314 were then divided into the 251 training images (80%) and 63 testing images (20%).

12-Layered Hand Designed CNN. The first ANN architecture experimented with was a 12-layered Convolutional Neural Network (CNN), implemented on the DL framework, PyTorch. This CNN's architecture was designed from scratch (Figure 4) with random weights. The training method used was simple supervised learning, in which the CNN gradually adjusted its weights using back propagation after calculating its loss. As there are no public datasets available exhibiting cavitated and healthy teeth images, the CNN trained with limited data. Under these circumstances, supervised learning on a CNN starting with random weights was not the right approach as performance was negatively impacted. Hence, we resorted to more advanced DL techniques to enhance performance.

Phase 1 of Training: Transfer Learning. To improve performance while training with limited data, transfer learning was used. It is well known in the deep learning field that low-level features (ex. lines or curves) are generic features that are present and learnable on any large image training dataset (even a dataset of cats and dogs). When dealing with limited data, a model’s performance is maximized if it previously learns general low-level features on a larger dataset independently, retains that knowledge, and then specializes in learning the higher-level features using the limited dataset. As a result, we experimented with two general purpose CNN models (ResNet-18 and ResNet-27) that were previously trained on the ImageNet1K dataset, which has over 1.2 million images [17]. In terms of architecture design, the ResNet-18 model had 18 convolutional layers with residual blocks and 1 fully connected layer. Additionally, the ResNet-27 had 27 convolutional layers with residual blocks and 1 fully connected layer. These two models had already learned low-level features that could be useful for cavity detection. Their knowledge was preserved as the weights in the CNNs’ preliminary layers were kept nearly unchanged while training on the limited dataset. The resulting models then became specialized in cavity detection as their deeper and last layers were permitted to change to learn the high-level features specific to the limited dental dataset, thus improving performance.
2020 Field-Collected Dataset. To conduct phase two of training, we collected images that were more reflective of the field images future users will provide. While the 2019 Web-Searched Dataset images were appropriate for training, the majority of the images were taken in ideal settings with good lighting (studio quality). These conditions resulted in images combined into one group (cavity presence) and high resolution images. In practice, it’s envisioned that the photographs provided by future users will be of low quality and will be taken at home using an inexpensive camera (ex. $30 intra oral camera). As it was important that the models performed well under the listed conditions, we ensured the above conditions were reflected in the 2020 Field-Collected Dataset. Thus, the 2020 Field-Collected Dataset images were taken at the homes of 11 human participants who consented to having their teeth photographed using a sterilized intraoral camera (Figure 5). Of the 11 consenting participants, 9 were female and 2 were male. The participants’ ages ranged from 8-70 years old and were all born and raised in regions of India. All participants signed a dental photography form consenting to having their teeth photographed and utilized in this study. For underage participants, parents were also asked to sign the consent form. Tooth surfaces photographed included participants’ occlusal, labial, buccal, and lingual surfaces. To minimize discomfort, participants were asked to have their mouth open for 40 seconds and then closed for 10 seconds. This process was repeated until all tooth surfaces were photographed. Using this process, 192 images were collected from the participants. The same labelling process, where the contacted dentist diagnosed the images using the ICDAS rubric, was repeated for this dataset. Out of the 192 images, the dentist determined that 56 images had no cavities and 136 images had presence of a cavity. These images were then divided into 157 training images (80%) and 35 testing images (20%).

Phase 2 of Training: Curriculum Learning. Curriculum learning is a specialized type of transfer learning in which a model is trained in multiple stages before being evaluated on a target validation dataset. The curriculum used in this study consisted of the ImageNet 1k dataset, followed by the 2019 Web-Searched Dataset (stage 1), and then the 2020 Field-Collected Dataset (stage 2) (Figure 6).

Local Interpretable Model-Agnostic Explanations (LIME).
The capabilities of LIME in providing explanations for a model’s cavity diagnosis were also explored in a PyTorch environment. LIME was applied on a ResNet-18 model trained on the dental data. This was done since PyTorch offered built-in support for ResNet-18 architectures. The explanations generated by LIME successfully highlighted the image regions that drove the ResNet-18 to its diagnoses. Specifically, when the ResNet-18 predicted cavity presence for a given image, LIME’s explanation highlighted the tooth surfaces that heavily influenced the model’s diagnosis. These explanations can also inform patients about the specific tooth surfaces that require more attention during their oral routine. By applying LIME, the study’s AI diagnostic system now has the ability to explain its cavity diagnosis to future users in an understandable manner.

Results

General Overview. In the following sub-sections, we describe the results of transfer learning and two-stage curriculum learning experiments on web-searched and field-collected data.

2019 Web-Searched Dataset. The 314-image collection acquired from public sources was split into a training and testing dataset in the ratio of 80:20. The training and testing dataset consisted of 251 and 63 images respectively. The datasets consisted of all 7 classes of decay as per the ICDAS rubric in the proportions shown in Figure 7. Since the study’s main objective was to diagnose images based on cavity presence only (not into the specific 7 ICDAS classes), the ICDAS 1-6 images were combined into one group (cavity presence), and ICDAS
Artificial Neural Network (ANN) Models. The performance of 3 ANN architectures in cavity detection was initially compared using the 2019 Web-Searched Dataset. The architectures included a hand-designed CNN, a ResNet-18 model, and a ResNet-27 model. Table 1 shows each model’s architecture in terms of their layers and if the architecture had an ImageNet1K base model available.

Phase 1 of Training: Transfer Learning. Models pre-trained on the ImageNet1K dataset produced higher accuracy compared to their base models. Pre-training allowed for better learning, which is supported by the accuracies presented in Figure 8. Without pre-training, 62% accuracy was achieved by the ResNet-27. However, with the use of pre-trained weights, accuracy increased to 77.7%. Based on these results, it was evident that transfer learning had improved performance.

Initial 12-layered CNN. As the CNN model wasn’t pre-trained on a larger data set, the model’s performance in detecting cavities was similar to random guessing (50%). Thus, it was concluded that the model was inadequate in detecting cavities.

Pre-trained ResNet-18. Both ResNet models were pre-trained on the ImageNet1K dataset. However, the ResNet-18 diagnosed the 63-image test set with an accuracy of 76.1%. Unlike the initial CNN, the ResNet-18’s accuracy values increased as more training epochs were completed (Figure 9). This indicated that pre-training in conjunction with using a ResNet architecture improved a classifier’s diagnostic ability.

Pre-trained ResNet-27 (Model X). Highest accuracy values were obtained by the pre-trained ResNet-27 model on Caffe. This model was named “Model X”. Model X’s hyperparameters included a 64-image batch size, a learning rate of .001, and a step size of 1500. Model X diagnosed the 63-image test set with an accuracy of 77.8% (49/63 images) after training for 600 iterations. To determine the type of diagnostic errors Model X made, a confusion matrix was also generated (Figure 10). Model X earned a sensitivity score of .69, which indicated that the model can correctly identify patients with a cavity 7 out of 10 cases. Model X also obtained a specificity score of .84 (model can correctly identify patients with no cavities 84 out of 100 cases). Finally, the precision score of .75 indicated that for every 100 cases the model diagnoses as having a cavity, 75 of them would actually exhibit cavity presence. To determine which ICDAS classes were the most difficult for the model to detect, further analysis was done (Figure 11). After comparing the performances of the 3 architectures, it was concluded that the ResNet-27 model was most optimal for cavity detection, thus all other experiments including applying curriculum learning were only applied to ResNet models.

**Table 1. Summary of Each ANN Model’s Architecture**

<table>
<thead>
<tr>
<th></th>
<th>Initial CNN</th>
<th>ResNet-18</th>
<th>ResNet-27</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number (#) of Layers</td>
<td>12</td>
<td>18</td>
<td>27</td>
</tr>
<tr>
<td># of Convolutional Layers</td>
<td>2</td>
<td>17</td>
<td>26</td>
</tr>
<tr>
<td># of Fully Connected layers</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ImageNet1K base model available for transfer learning</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Figure 8. Overall performance of the architectures as transfer learning experiments were conducted.**

**Figure 9. Ideal upward accuracy trend was exhibited.**

**Figure 10. Model X’s confusion matrix on the 63-image test set.**

**Figure 11. Model X provided the most accurate diagnoses for no cavity (31/37 images), moderate decay (5/6 images), and extensive decay (7/9 images).**
2020 Field-Collected Dataset. To ensure field conditions were reflected in the 2020 dataset, the 192-image dataset was taken at the homes of the 11 consenting human participants using an intra oral camera. After receiving the dentist’s labels, the 2020 dataset was split into a training and testing dataset in the same 80/20 ratio described previously. The training and testing dataset consisted of 157 and 35 images respectively. Additionally, all 7 levels of decay were represented in the test dataset (Figure 12).

Model Y. Despite training with more complicated field collected dataset images, we still aimed to produce similar accuracy results (~80%) as Model X. Thus, the next model was pre-trained on ImageNet1K, and then trained for 600 iterations on the 2020 Field-Collected dataset. Final accuracy obtained was 77.14% (27/35 test images). This model was named “Model Y”.

Phase 2 of Training: Curriculum Learning. Based on Model Y’s accuracy, it was evident that transfer learning wasn’t sufficient enough to cope with the complex 2020 Field-Collected images; however, performance was enhanced by applying two-stage curriculum learning. The curriculum used consisted of ImageNet 1k followed by the 2019 Web-Searched Dataset (stage 1), and then the 2020 Field-Collected Dataset (stage 2). The model that emerged using this curriculum was name “Model Z”. Using curriculum learning, Model Z diagnosed the 35-image test set with an accuracy of 82.8% on average (29/35 images) after training for 600 iterations (Figure 13). Curriculum learning allowed for better learning on the 2020 Field-Collected Dataset, which was supported by the two accuracies presented in Figure 13. Accuracy obtained by the base ResNet-27 model with random weights was 71.4%. By using pre-trained weights and curriculum learning techniques, accuracy increased by 11.8%. It also became evident that curriculum learning enhanced performance after comparing Model Z’s accuracy (82.8%) to Model Y (77.1%), intelligent guessing/zero method (68.5%), and random guessing (50%) (Figure 14).

Model Z Confusion Matrix Results. To determine the type of diagnostic errors Model Z made on the 2020 Field-Collected Dataset, a confusion matrix was generated (Figure 15). For cavity cases (ICDAS classes 1-6), the model achieved 100% accuracy (24/24 images). This was supported by Model Z’s sensitivity score of 1.0. This score indicated that on images obtained by an intra oral camera, Model Z can correctly detect cavity presence on every image, regardless of the level of decay present. For no cavity cases (ICDAS class 0), Model Z achieved an accuracy of 45% (5/11 images), which is equivalent to a specificity score of .45. Finally, Model Z earned a precision score of .80, which indicated that for every 100 images/cases Model Z diagnoses as having a cavity, 80 would actually have cavity presence.
Explaining Model Output Using LIME Algorithm. The capabilities of LIME in providing explanations for a model’s cavity diagnosis were explored in PyTorch. LIME was applied on a pre-trained ResNet-18 model. The figures below represent outputs of the LIME algorithm, which are sample explanations that highlight the image regions that most influenced the ResNet-18 model’s diagnoses. Figure 16a. and Figure 16b. present two sample LIME explanations that were generated after the model correctly predicted cavity presence.

Discussion

Improvements on Previous Work (Berdouses et al., 2015). The table summarizes the differences between the AI cavity detection systems developed in this study and Berdouses et al, 2015.

<table>
<thead>
<tr>
<th>Study</th>
<th>Photographs Color Images</th>
<th>Type of Machine Learning</th>
<th>Top-1 Accuracy</th>
<th>Explainability</th>
<th>Occlusal Tooth Surface</th>
<th>Labial Tooth Surface</th>
<th>Lingual Tooth Surface</th>
<th>Buccal Tooth Surface</th>
<th>Teeth Present in Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Study</td>
<td>Yes</td>
<td>Deep Learning CNNs</td>
<td>82.8%</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Multiple</td>
</tr>
<tr>
<td>Berdouses et al., 2015</td>
<td>Yes</td>
<td>Random Forest</td>
<td>80.0%</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>One</td>
</tr>
</tbody>
</table>

LIME’s Potential. The ability to explain a diagnosis to a patient in an understandable manner is a crucial skill employed by dentists. By using LIME, the study’s cavity detection system has also acquired this communication skill. To our knowledge, there are no published studies that explain their classifier’s cavity diagnoses for photos. Using LIME, this study has closed the explainability gap previously present for AI cavity diagnoses. LIME treats the AI model as a black box, focusing on producing simple explanations that demonstrate how input data map into output classifications. While there are other complementary AI explainability algorithms available (ex. Grad-CAM), they’re more focused on providing insight into the inner workings of the AI model. The outputs of such methods tend to be consumed by AI researchers and data scientists [18]. In contrast, the LIME approach presented above may put us in a better position in providing user-centric explanations to patients and dentists.

Limitations of LIME. LIME was developed in 2016, and there are extensions of LIME that achieve similar results (i.e. SHapley Additive exPlanations (SHAP)). One limitation of these explainer algorithms is that they produce explanations at the pixel level. They don’t take into account semantic features (ex. texture). Another limitation is that these algorithms are applied on a trained model in a post-hoc manner. In the future, we aim to test a machine learning technique that considers explainability even during training (ex. Retain algorithm).

Regional Bias. Of the 192 images collected for the 2020 Field-Collected Dataset, only two stained tooth images were present, and both were in the validation dataset. We determined that the two stained images were part of the six no cavity images Model Z had incorrectly diagnosed. These stains were from tobacco and tea (Figure 17). As zero stained images were present in the no cavity training dataset, Model Z never learned how to differentiate between cavities and tea/tobacco staining. To improve Model Z’s specificity score of .45, additional images exhibiting tea/tobacco staining have to be collected. These stains have also prompted a conversation on regional bias, which is a prevalent topic in AI research. Oral hygiene habits can be divided based on geographical regions.
US, 88.2 million adults use tobacco (25.2%) [19]. However, in India, where all 2020 Field-Collected Dataset images were taken in, over 274 million adults (34.6%) use tobacco [20]. Furthermore, tea consumption per capita in India is also 1.4 times higher than the US [21]. Thus, it’s not surprising that we encountered both tobacco/tea staining in the 2020 dataset images (taken in India), but not in the 2019 dataset images (majority taken in the US). To reduce false positives cases, future training datasets must be more accommodating towards tobacco users, alcoholics, and heavy tea and coffee drinkers.

**Semi-Supervised Learning Techniques.** To enhance Model Z’s performance in the future, additional data needs to be obtained. However, an increase in image quality will also make the data annotation process for dentists more time consuming and tedious. As a result, dentists may become hesitant in providing the labels. In the future, it’s aimed that semi-supervised learning techniques can be used to decrease the reliance on dentists for labels. As Model Z is already trained on labelled data, future images obtained won’t need to be labelled. Instead, they simply will receive a pseudo label based on Model Z’s predictions. Once all unlabeled data receives a pseudo label, the labelled data (2019/2020 dataset images) and pseudo-labeled data can be combined into one training dataset. Using this technique, Model Z will have the opportunity to train on a vastly larger dataset.

**Application** The current results on a ResNet-27 with an accuracy of 82.8% indicate that an AI diagnostic system with ANN architecture used in practical setting can be created. As suggested by a practicing dentist, images of patients’ teeth can be acquired using an intra oral camera ($30). Classification using DL is achievable using a computing device like a $35 Raspberry Pi. This affordable setup could potentially be used as first level screening for dental triage in developing nations where there is a limited dentist availability. In developed countries, the setup could potentially detect cavities at home for people who are dentaphobic or can’t afford dental checkups.

**Conclusion**
Dental cavities, one of the most prevalent chronic disorders, impacts over 3.9 billion people worldwide. Typically, cavity detection requires the services of a trained dentist. However, barriers such as dentophobia, limited dentist availability, and lack of dental insurance prevent millions from receiving dental care. To address these issues, we created an AI diagnostic system that detects cavity presence and visually explains the rationale behind each diagnosis. The accessible system detects cavities of all levels of decay using an artificial neural network (ANN). With a single photographic color image, the system can provide a cavity diagnosis. Previously, there was a lack of studies focusing on whether AI can be used to detect early to advanced tooth decay on different surfaces (occlusal, lingual, buccal, and labial tooth surfaces) on photographic images. This study aimed to address this research gap. First, we aimed to select an ANN architecture suitable for the system. To facilitate this, we collected over 500 de-identified photos from online sources and consenting human participants using an intra oral camera. We then experimented with several neural network architectures and training techniques. Using transfer learning from an ImageNet1k dataset, the ResNet-27 architecture proved to be most optimal for cavity detection after earning an accuracy of 77.8% and sensitivity score of .69. ResNet-27’s accuracy and sensitivity score were then improved to 82.8% and 1.0 respectively using two-stage curriculum learning. Visual explanations for the system’s cavity diagnoses were generated using LIME. After applying LIME, the system now has the ability to detect cavity presence and explain its cavity diagnosis to the end-user in an understandable manner. This explainability feature was not present in previous work. By gaining two crucial skills typically employed by dentists, this study’s AI diagnostic system can now provide reliable cavity diagnoses to demographics that have constantly been unaccounted for in the past.

**References**

“Please Advise”: Understanding the Needs of Informal Caregivers of People with Alzheimer’s Disease and Related Dementia from Online Communities Through a Structured Topic Modeling Approach

Kerou Cheng¹, Zhijun Yin, PhD¹,²

¹Department of Computer Science, Vanderbilt University, Nashville, TN USA; ²Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN USA

Abstract

The informal or family caregivers of the Alzheimer’s disease or related dementia (ADRD) patients, also known as the “invisible second patients”, are often reported experiencing emotional and behavioral hardships. In recent years, the rapid development of online communities provides these caregivers a new opportunity for seeking information and emotional support. Comparing with offline social support services which have been constrained during the COVID-19 pandemic, online support allows caregivers to reach many peers in a convenient manner. This research aimed to examine the issues faced by ADRD caregivers through performing a structural topic modeling on posts from two online communities. Results revealed that the top concerns of the caregivers include getting along with Alzheimer’s patients, family issues, patients’ internal medical issues, stages of the disease, care facilities, etc. The results may have a further implication to the future implementation of psychological and social intervention of ADRD family care.

Introduction

Alzheimer’s Disease is a major cause of progressive dementia that involves memory loss and other cognitive disorders. At a late stage, a patient with Alzheimer’s Disease and related dementia (ADRD) may lose the ability to respond to their environment physically. ADRD is very common and according to Alzheimer’s Association, in 2021, more than 6 million Americans of all ages have ADRD, among which approximately 6.2 million are aged 65 and older⁴.

The family caregivers of ADRD patients are essential in providing them with assistance, including bathing, moving, dressing, etc. However, they are often faced with emotional and behavioral challenges. Many of them reported that they had to deal with issues like sleep deprivation, agitation, and depression⁵. The emotional challenge is strongly associated with caregivers’ anger toward the patients and the caregivers’ concerns about the limitation of social life as well as a restriction in personal time⁶. Although many in-person social support, such as psychological interventions and professional training, have been proved effective to improve caregivers’ emotional well-being and caring competence, these resources are not easily accessible to caregivers with geographic or time constraints, like caregivers living in rural areas⁷. Moreover, in-person courses bring about privacy concerns for some family caregivers and become especially less practical during the COVID-19 pandemic⁸. In recent years, the rise of online communities provides ADRD caregivers with an opportunity to seek informational and emotional support among tens of thousands of peers, without the restriction of geographical distance.

Comparing to offline support services or groups, online communities greatly reduce the cost to access resources, increase the convenience for communications between a large number of caregivers, and reduce the isolation and social stigma that many caregivers experience when caring for an ADRD patient⁹. The fact that caregivers are able to communicate with others with similar situations strengthens their sense of social inclusion and belonging to a group¹⁰. As a result, they are more encouraged to share their experience confidently¹¹. Examining a large number of posts composed by the caregivers in the online communities can potentially provide insights into their difficulties, which may provide directions for the development of support programs, or for the community moderators to help caregivers find the information they need in the future.

In the field of biomedical informatics, researchers are actively employing statistical and machine learning techniques to mine the online data for valuable insights into many health-related topics¹². For example, Chancellor et al. investigated the differences in social norms between two online weight loss communities by building a word embedding model on comment words¹³. Similarly, Lerrigo and Sinha used topic modeling and sentimental analysis techniques to explore the unmet need in patients with Inflammatory Bowel Disease on Crohn's and Colitis Foundation community forum¹⁴. However, there are not yet studies to examine the online communities of ADRD caregivers with topic modeling.
This research aimed to identify and examine the common topics in posts published by ADRD caregivers on two popular online communities, alzconnected.org (ALZConnected) and Reddit, and further compare the topic prevalence on these communities. Founded by Alzheimer’s Association, ALZConnected is the first online community that is dedicated to people affected by ADRD, including patients and their caregivers. By contrast, Reddit is a community-specific news aggregation and discussion platform that is widely used by users all over the world. We discussed the implication of our findings in supporting ADRD caregivers using online data in future research.

Method

Data

Online communities are often organized into different discussion threads. Each thread starts with an initial (or topic) post which is followed by many comments. This analysis focused on the initial posts on Reddit and ALZConnected. We used the Pushshift API to obtain the initial posts from two subreddits: r/dementia and r/Alzheimers, which are two main subreddits for Alzheimer's caregivers. We wrote a web crawler using Python Requests (version 2.25.1) and BeautifulSoup (version 4.9.3) libraries to obtain the initial posts from the caregiver forum on ALZConnected. This research was exempt from human subjects research by Vanderbilt University Institutional Review Board.

Objective

The goal of the research project is to apply structural topic modeling to identify the common issues posted by ADRD caregivers on two popular online communities, ALZConnected and Reddit, compare the topic prevalence across communities, and discuss the implication of the findings to the future intervention design.

Approach

Structural Topic Model In natural language processing, topic modeling is a probabilistic model to discover topics in a set of unstructured documents. Instead of the traditional topic models, we chose structural topic modeling (STM) for the analysis because it enables us to not only find groups of tightly co-occurring words that represent topics but also estimate how a covariate of interest is associated with the topic prevalence. In this project, the community of the posts (ALZConnected or Reddit) is the binary covariate of interest. With STM, it is possible for us to find common issues faced by ADRD caregivers as topics and discover the effect of the platform of the posts on the topical prevalence.

Model Selection We performed a model search across different numbers of topics \( K \) with the search\( K \) function from the stm package (v1.3.6). This function calculates metrics of each model for further evaluation. Specifically, we trained models with \( K \) ranging from 5 to 30 and plotted out the evaluation metrics including the held-out likelihood, residual dispersion, semantic coherence, EM iterations, exclusivity, and lower bound. The held-out likelihood of a model estimates the probability to generate the held-out documents. The residual dispersion provides an indication in favor of a larger \( K \). The exclusivity measures the extent to which the top words of a topic do not appear as top words in other topics, while the semantic coherence measures whether probable words for a topic tend to co-occur in documents (in either training documents or external documents). To obtain interpretable topics, we first used the elbow rule to narrow down the range of the possible candidate topic numbers for all the metrics (where the changing rate of metric value starts to decrease). Then, for each candidate topic number within this range, we conducted topic modeling and manually selected the model with the best interpretability.

Results

Data Statistics

Table 1. Platform, number of initial posts, number of users who published ≥1 initial post(s), start year, and latest year of data.

<table>
<thead>
<tr>
<th>Platform</th>
<th># Initial Posts</th>
<th># Users</th>
<th>Start Year</th>
<th>Latest Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALZConnected</td>
<td>33,132</td>
<td>10,175</td>
<td>2010</td>
<td>2021</td>
</tr>
<tr>
<td>Reddit</td>
<td>9,096</td>
<td>5,766</td>
<td>2010</td>
<td>2021</td>
</tr>
</tbody>
</table>

Table 1 shows the statistics of the data on each platform. Figure 1 shows the distribution of the number of initial posts published by each user on each community. While the number of initial posts published by each user on both communities follow a long-tailed distribution, the distributions are significantly different with a Kolmogorov–Smirnov test (\( p < 0.001 \)). As shown in Figure 2, while more initial posts were published on ALZConnected before
2019, the number of initial posts published on Reddit has increased over years since 2010 and started to surpass the number of initial posts on ALZConnected in 2019.

**Figure 1.** Number of initial posts published per user on each community.

**Figure 2.** Number of initial posts published on each community by year

**Model Selection**

Figure 3 illustrates how the six proposed metrics change over different $K$ values. From the figure, it can be seen that the changing rate of all the metrics started to decrease when $K > 10$. The “elbow” of each curve shows up until $K < 17$. Given this range, we run topic modeling for each topic number in \{11, 12, 13, 14, 15, 16\}. After examining the topics and the representative documents of the topic models, we found that the topic model with $K = 16$ produced the most exclusive but meaningful topics. Therefore, we chose the STM with $K = 16$, denoted by STM-16, as our final model for further analysis.

**Figure 3.** Model evaluation metrics with different $K$ values.

**STM-16 Topics**

**Table 2.** List of topics produced by STM-16 model.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Metric</th>
<th>Top Representative Words</th>
<th>Interpretation</th>
<th>Expected %</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>Kappa</td>
<td>Getting Along with Patients</td>
<td>PROB</td>
<td>Family Issues</td>
</tr>
<tr>
<td>----</td>
<td>-----------------------</td>
<td>-----------------------------</td>
<td>---------------</td>
<td>--------------</td>
</tr>
<tr>
<td>3</td>
<td>kindergarten, fluster, wanna, how, snide, fianc, nana</td>
<td>15.287</td>
<td>just, know, dont, like, get, say, want</td>
<td>10.179</td>
</tr>
<tr>
<td>12</td>
<td>Kappa</td>
<td></td>
<td>help, care, live, dad, mother, get, take</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>kappa</td>
<td>Participant Recruitment</td>
<td>post, read, help, thank, dementia, can, caregiv</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Kappa</td>
<td>Internal Medical Issues</td>
<td>said, back, day, got, told, went, get</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Kappa</td>
<td>Stage of Disease</td>
<td>choir, plateau, stage, experienc, year, seem, ago</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>kappa</td>
<td>Care Facilities</td>
<td>facili, resenti, facil, nurs, ltc, staff, reloc</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Kappa</td>
<td>Research Materials</td>
<td>oxid, amyloid, nitrat, neuron, neurotransmitt, httpwwwncbinlmnihgovpubm, mutat</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Kappa</td>
<td>Grief and Loss</td>
<td>fundrais, twitter, cure, bride, alzheimer’, brilliant, beth</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Kappa</td>
<td>Tips for Home Care</td>
<td>piano, gps, domino, game, toy, indoor, solitair</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Kappa</td>
<td>Religious Support</td>
<td>brave, uncondit, &quot;go, hero, writer, taught, soul</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Kappa</td>
<td>Financial and Legal Issues</td>
<td>landlord, debt, ex-husband, guardianship, conservatorship, harass, rent</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Kappa</td>
<td>Public Resources</td>
<td>premium, deduct, govern, recruit, beneficiari, asset, waiver</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Kappa</td>
<td>Aggressive Behavior Issues</td>
<td>antipsychot, hallucin, prescrib, seizur, aggress, auditori, paranoia</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Kappa</td>
<td>Diet Issues</td>
<td>fridg, butter, pasta, peanut, pantri, veggi, gag</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Kappa</td>
<td>Urinary Incontinence</td>
<td>bleach, flush, pee, diaper, teeth, robe, pant</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Kappa</td>
<td>Phone Scams</td>
<td>landlin, password, scam, voicemail, text, scammer, fraud</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>phone, card, text, mail, mil, call, email</td>
<td></td>
</tr>
</tbody>
</table>
The topics produced by STM-16 are listed in Table 2 and ordered by their expected proportion. The top representative words of each topic are listed with three different measurement metrics for better interpretation of the topic meanings. Specifically, the \textit{topic Kappa (Kappa)} returns words that are most associated with each topic\textsuperscript{12}. The \textit{PROB} metric measures the marginal probability of a topic word\textsuperscript{11}, while the \textit{Frequency-Exclusivity (FREX)} score is a univariate metric of topic words that averages their performance in both frequency and exclusivity\textsuperscript{15}. To better interpret the topics along with the representative documents, we built an interactive dashboard\textsuperscript{*} to conduct a qualitative analysis. For each topic, this dashboard displays a word cloud of its top representative words, and a list of representative posts that have the largest distribution in this topic. Each topic label in Interpretation column of Table 2 was generated by all the authors with the use of the dashboard. We summarized each topic as follows:

**Topic #3 Getting Along with Patients** The most frequent topic mentioned in all posts is Topic #3, which is related to issues with getting along with the Alzheimer’s patients. One of the scenarios that is often shared in this topic is the fluctuation in the patients’ emotions. For example, one user wrote “I am not sure how to handle my mom’s sudden change in the emotions area [...] She goes from crying and walking around saying she just wants to go home to being upset and out of [nowhere] telling me she didn’t say or do something [...] Should I just ignore it and let her go through this?” Another user mentioned about their father’s drastic personality change, “my mother and I started really worrying about my dad because the past couple of months he has had a major personality change like he had always been very shy and quiet, never really talking to strangers much and then he started talking to everyone and he would cry...” Users felt confused, helpless, and emotional when seeing their diagnosed family members getting depressed, aggressive, or forgetful. They used the forums as a vent, as a user wrote, “my husband doesn’t understand why I feel this way. I just need someone to talk to who understands.”

**Topic #12 Family Issues** Topic #12 is the second frequently mentioned topic. It is related to family issues with the caregivers. A number of users are in the “sandwich generation” who have both diagnosed parents and young children to take care of. They asked for advice on how to manage time and seek support on the forums. For example, one user wrote “we have been caring for my mother who is 84 with dementia in our home for over a year now, still raising 4 minor children. My husband [and] I are becoming overwhelmed and need some [advice] ...” Another user who took care of their father for two years decided to reach out to some support group, revealing that “I am a single mom of a 7yr old and take care of my dad full time and I work a full-time job as well.” Users also posted about feeling guilty for wanting to “creating a life for themselves as a young adult” while taking care of their parents. Other family problems mentioned in this topic include having self-centered siblings who were not willing to take care of the diagnosed parents and having parents who did not get along well while one of them was diagnosed with Alzheimer’s.

**Topic #14 Participant Recruitment** Topic #14 mainly includes posts from researchers who were looking for participants for their research projects. Most of them were looking for roles like caregivers, healthcare workers, and support workers to complete anonymous surveys online.

**Topic #2 Internal Medical Issues** Topic #2 is related to internal medical issues like chest pain, pneumonia, kidney infection, and so on. While these issues were not necessarily related to Alzheimer’s Disease, but they also drew the attention of the caregivers for their loved ones.

**Topic #16 Stage of Disease** Users discussed the progression of the disease. They posted about the stage of Alzheimer’s that the patient was in and asked about the duration of the stage. For example, one user wrote “my mother is in stage 6,” and asked “how long does this stage last? What stage is everyone in right now and for how long?” Although many users mentioned that they understood that “the progression of Alzheimer’s is impossible to predict and everyone is different”, they often described the patient’s condition and hoped someone with a similar experience could communicate with them.

**Topic #13 Care Facilities** In this topic, users posted about moving their diagnosed parents to memory care units, nursing homes, or assisted living facilities. Users often asked for suggestions on what facility to choose. For example, a user asked “I believe there are more activities in the nursing homes. Why do so many people pick ‘memory care’ for placement?” Another user listed the available choices of skilled care facility and memory care facility for their parents and asked, “what’s the best choice for both my parents?” Users also asked questions related to the functioning parent. For example, a user with a functioning father who was about to move their mother to the memory care unit wrote “For

\footnote{Interactive dashboard (posts presented were rephrased for privacy consideration): https://carol-cheng-98.shinyapps.io/alzheimers-topic-modeling/}
a lot of reasons I’m thinking it would be better if he stayed home day of the move. I would appreciate hearing about other’s experience with what to do with the functioning parent when moving to into care facility?”

**Topic #7 Research Materials** In this topic, users shared some knowledge about ADRD. Users usually shared links to articles they thought are helpful to understand the cause or potential treatments of the disease with a few sentences to summarize the content of the articles. For example, one user, who claimed themselves as a “non-scientist who has reviewed research into non-communicable diseases, including Alzheimer’s, for 30 years,” linked an article as “one of the best explanations for Alzheimer’s disease that is out there,” and said that “almost all are being caused by chronic, aberrant activation of the transcription factor, nuclear factor-kappa beta.” Another user who claimed to have been “studying Alzheimer’s disease for nearly fourteen years” shared an article that paralleled their own understanding of the disease and provided a summary of the findings in the article.

**Topic #11 Grief and Loss** In this topic, users expressed the grief they were feeling for their passing family with Alzheimer’s. Many users felt bad for the suffering that the patient had to go through. For example, one user whose mother passed away wrote that “I have been grieving horribly. Not for her passing. I am so grateful and thankful that she is finally free from this disease. But I am more than sad for the 12 years she suffered. For the 12 years I lost her piece by piece...” Another user expressed similar emotions, saying that “now that mom’s gone, I feel very sad and also relieved.”

**Topic #5 Tips for Home Care** In Topic #5, users shared tips for taking care of patients at home. For example, some users recommended installing security cameras at home to keep an eye on the activities of the patients. Some users who were taking care of patients with a fidgeting problem recommended fidget blankets for picking and anxiety or attaching ribbons and elastic to unused stuffed animals.

**Topic #10 Religious Support** In this topic, users shared some Bible verses in support of the fellow users.

**Topic #8 Financial and Legal Issues** Posts in this topic are related to the legal and financial issues of ADRD patients. One of the issues that are brought up frequently is the driving problem of Alzheimer’s patients. In a user’s post, their diagnosed father got into an accident with his car and was asking for suggestions in how to keep their father from driving. Another user wrote about their mom convinced DMV into issuing her a conditional driver’s license while they believe that their mom should not drive. Users also asked for advice on other financial and legal problems like patients being the power of attorney (PoA) of property, such as a house, revocable trust, and so on.

**Topic #6 Public Resources** Users shared resources that were related to Alzheimer’s disease in this topic, especially about federal programs like Medicare and Medicaid. For example, one user shared about the different coverage of Medicare Part A, Part B, and Part C in terms of adult daycare. Users also shared resources of some caregiver action networks, webinars on family caregiving. There are also posts about specific laws that were related to caregiving. For instance, a user wrote, “if you were not aware, this is the law in CA – California law prohibits untrained personnel from feeding residents...”.

**Topic #15 Aggressive Behavior Issues** In this topic, users posted about the aggressive behaviors of ADRD patients. The users wrote about problems in patients like attacking people, delusions, paranoia, and agitation. They also mentioned sleep problems in the patients along with these aggressive behavior issues. For example, one user wrote “my mom suffering from vascular dementia and she is as stubborn as a mule cause she paces back and forth day and night going on with only an hour or less of sleep, and then when she crashes the most she sleeps is 3-4 hours...She has tried different meds but nothing has worked.” Another user wrote “my FIL has become increasingly agitated, aggressive and combative with my MIL who is his primary caregiver since having a seizure about 7 weeks ago. We have been unable to determine a trigger for his aggression towards her for his attitude changes extremely quick like flipping a switch. You never know how long these episodes will last...” Although most of the users already turned to doctors for help and had some prescriptions, they posted to look for people facing similar issues.

**Topic #9 Diet Issues** In this topic, users posted about the diet issues of their patients with ADRD. These issues include overeating due to the dementia, inability to eat for patients in the late stage, and refusal to eat regular food because of losing taste. For example, one user mentioned that their mother who loved drinking coffee over-drunk coffee because she could not remember how many c1ps she drank. A user whose mother said that they had to hand feed their mother and be careful when preparing food so that their mother did not choke. Another user was looking for suggestions about their mother who added too much sugar in her meal because she could not taste the regular food.

**Topic #1 Urinary Incontinence** Users posted about the patient’s urinary incontinence issues in this topic. Many of them were worried because their parents refused to wear adult diapers or even hid their soaked pants. One user wrote
“mom can still go to the bathroom by herself, and most times put on her clothes not always right but she still can do this. Problem is she will either forget or choose not to put on the pull up depends and will of course wet herself…”

**Topic #4 Phone Scams** In this topic, users frequently asked for advice on how to deal with phone scams on the patients. The users complained about their diagnosed parents getting calls from some “charity” and giving their credit card information to the caller. Many users were helpless with the phone scams. A user wrote, “we have caller ID, we have unidentified calls blocked, but still, Mom gets calls daily from people trying to take advantage of her in one way or another.” Another user could not find an effective way to block the phone scams, saying that “do not call list is a joke.”

**Topic Correlation**

We used the `topicCorr` function in the `stm` package (v1.3.6) to estimate the correlation between each pair of topics. Figure 4 displays a network based on the topic correlation graph calculated by the function. In the figure, each vertex represents a topic, and each edge indicates a positive correlation between topics. The thickness of the edge indicates the strength of the correlation between topics.

![Figure 4. Topic correlation network of STM-16 model. A thicker edge suggests a higher correlation between two connected topics.](image)

The correlation network roughly divided the topics into several groups based on the connections between them. The first closely related group of topics include Topic #8 (Legal and Financial Issues) and its connected topics like Topic #12 (Family Issues), Topic #4 (Phone Scams), Topic #13 (Care Facilities). This topic group is related to the common issues that the caregivers are likely to run into. These problems may not be directly related to the symptoms of the disease, but they are brought up frequently by caregivers as problems that bothered them as well. Another group of topics, including Topic #2 (Internal Medical Issues) and its connected topics like Topic #1 (Urinary Incontinence), Topic #5 (Tips for Home Care), Topic #9 (Diet Issues), Topic #16 (Stage of Disease) and Topic #15 (Aggressive Behavior Issues), are directly related to the symptoms or health conditions of the ADRD patients. Although Topic #7 (Research Materials) is not directly related to the symptoms of the disease, it is connected to Topic #15 and Topic #16, suggesting the shared research articles covered the prescriptions and causes that are related to the symptoms. Additionally, Topic #14 (Participant Recruitment) and Topic #6 (Public Resources) are highly correlated because there are many research projects regarding the ADRD caregivers that are interested in examining the accessibility and
the impact of public resources (e.g., Medicare). Finally, it should be noted that despite a weak correlation, Topic #10 (Religious Support) only correlated with Topic #11 (Grief and Loss), which makes sense considering their natural connection in real life.

Effect of Community on Topics

The effect of post community on topic prevalence is illustrated in Figure 5. Among the most frequent topics by caregivers, the concerns that appear more on Reddit include Grief and Loss, Family Issues, Public Resources, Aggressive Behavior Issues, Care Facilities, Participant Recruitment, and Internal Medical Issues. The topics that appear more on ALZConnected are Religious Support, Getting Along with Patients, Diet Issues, Research Materials, Phone Scams, and Financial and Legal Issues. Topics that have no significant difference in these two communities are Urinary Incontinence, Tips for Home Care, and Disease Stage.

Figure 5. Effect of post community on topics. A negative (positive) coefficient suggests that the topic is more likely to be mentioned in Reddit (ALZConnected).

Discussion

There is a difference between the two online forums for ADRD caregivers in terms of the distributions of the number of initial posts by each user. While both distributions follow the pattern of a long-tailed distribution where most users created less than 3 initial posts, a larger percentage of users on ALZConnected created more than 3 initial posts compared to that on Reddit. This may imply that ALZConnected users are more engaged in taking the initiative to share information or seek support and that a number of users on ALZConnected are devoted to contributing to the online community of ADRD caregivers. Despite more frequent personal activities and more initial posts that were published on ALZConnected before 2019, the number of initial posts published on Reddit has increased over years since 2010 and started to surpass the number of initial posts on ALZConnected in 2019. This trend may indicate that the online community on Reddit is growing over years.

With the structural topic modeling, we found that the topics that are most frequently discussed in the online forums of ADRD caregivers include getting along with the patients, family issues, internal medical issues, stages of disease, etc. Our exploration on the topic correlations further indicated that these topics can be divided into the following categories: symptoms related to the disease, other potential issues for caregivers, as well as knowledge and resource sharing. The topic contents reveal some possible sources of concern for the ADRD caregivers. For example, the symptoms of disease certainly bring about concerns among the caregivers. The caregivers looked for advice on how to deal with symptoms related to the patients’ diet issues, internal medical issues, aggressive behavior issues, or urinary
incontinence. This finding is in accordance with the claim of Glueckauf et al. that the caregiving requirements of the ADRD patients involve managing potentially injurious behavior to self or others, issuing frequent reminders, and monitoring hygiene and self-care activities, which are performed at the cost of the caregivers’ physical, financial, and mental wellbeing. In addition to the physical issues of the ADRD patients, other issues related to finance, law, family relation, and finding proper facilities also cause problems to the caregivers. These findings suggested that online communities can serve as convenient platforms for caregivers to search for similar problems or challenges (and, likely, related solutions or information) that they face during caring for a person with ADRD. These communities may leverage the collective knowledge/experience from tens of thousands of caregivers to provide peer support. However, as our analysis of community effect in topics showed, online communities have a different context, and some topics may be more likely to be mentioned in one community, while other topics may be more likely to be mentioned in another community. Future intervention based on peer support in online communities may focus on designing efficient strategies to help new caregivers quickly locate a proper community to gain more deep, effective discussion.

It is worth noting that, although the discussion of the caregivers thoroughly covered most caring services provided for the patients, like medical diagnostic services, home care assistance, day care, and hospice, there is not a separate topic that distinguishes the caregiver’s own emotional needs. Assistance and programs for the caregivers like psychological consulting, stress management, and lifestyle enhancement were relatively less mentioned in the posts. However, it can be inferred, by examining their posts, that many of them were generally overwhelmed, helpless, or sad about their situations. This may imply that, although, as suggested by Brodaty, ADRD caregivers are experiencing high rates of burden and psychological morbidity that threaten the quality of life, they are yet to take the professional services into consideration to alleviate their psychological burdens. The inadequate knowledge about the resources for caregivers may be the result of the fact that these resources have been hard to access and even nonexistent. This may have some implications in the future development of the support programs for the caregivers. For example, more attention should be paid to raise the awareness of the mood issues among the ADRD caregivers given the gap between the emotional needs and the frequency that the emotion management services were asked about. More importantly, actions should be taken to improve the accessibility and affordability of these services such that they can be accessed by a large number of ADRD caregivers.

The topic contents also reveal some potential needs of certain groups of caregivers, which indicates the directions for the implementation of the intervention programs for ADRD home care. For example, based on the representative posts of Topic #12 (Family Issues), the caregivers in the “sandwich generation” may be especially in need of help with balancing their personal life, taking care of children, and taking care of the diagnosed parents. It would be helpful if there are resources like guidance from experts, support groups, or psychological consulting that are designed or customized for supporting their needs.

Despite the notable findings, we acknowledge that there are several limitations in the present study that can be addressed by future research. First, the study only investigated the topic contents and topic prevalence of the posts on each online community without evaluating the sentiments of the posts, which may help us better identify the emotions of those caregivers. For future studies, a sentiment analysis could be performed on each topic to allow a further comparison between the communities in terms of the post sentiments, as well as an evaluation of the psychological burdens that the caregivers are experiencing. Second, we directly applied STM to all the posts, and apparently some posts were not published by ADRD caregivers (e.g., Topic #14). While it would be great to exclude them when investigating caregivers’ needs, our results suggested that the impact of the existence of such posts on our analysis was not significant. Third, we only examined and compared the communities that target ADRD caregivers. ADRD caregivers can publish caring-related posts in any community that they think proper. It will be interesting to build a universal classification model to identify those posts and then compare their topic prevalence with the topic prevalence in ADRD-focused communities. Moreover, as we only examined the needs of caregivers through mining the initial posts, future research should consider examining the responses to these initial posts to understand the mechanism of peer support among ADRD caregivers in online environments. Finally, in future work, it would be helpful for sociologists, psychologists, and social workers focusing on the ADRD community to involve in the interpretation of the posts and turning the insights into actionable points.

**Conclusion**

With the technique of structural topic modeling, the present study examined the common topics posted by Alzheimer’s caregivers on two popular online forums. The differences between ALZConnected and Reddit can be found in terms of both posting frequency and topic prevalence. The STM-16 model identified 16 topics in the posts, which are further divided into several categories based on the correlation between them. These categories include symptoms related to...
the disease, other potential issues for caregivers, and knowledge and resource sharing. Based on the implications of the research findings, psychologists and support program developers could develop meaningful caregiver intervention programs by raising the awareness of emotion problems among the caregivers as well as improving the accessibility of resources for caregivers in online communities. The topic model also provides potential directions for the implementation of intervention programs by identifying the need of certain groups of caregivers.

References

Exploring the Readability of Ingredients Lists of Food Labels with Existing Metrics

Kathryn Cooper, PhD1*, William Gasper, BS1, Ricky Flores, BS1, Martina Clarke, MPH, PhD1, Erin Bass, MBA, PhD, Leslie Evans, MS, RD, LMNT2, Jana Ponce, PhD, RD, LMNT3

1University of Nebraska at Omaha, Omaha, NE, USA; 2Nebraska Medicine, Omaha, NE, USA, 3University of Nebraska Medical Center, Omaha, NE, USA

Abstract

Healthy diet and dietary behaviors are key components in prevention of chronic disease and management of chronic illness. Nutritional literacy has been associated with dietary behaviors and consumer choice of healthy foods. Nutritional literacy can be measured, for example, by examining consumer food label use, but current research focuses largely on the Nutrition Facts panel of a food product. Ingredients lists are critical for communicating food composition but are relatively unstudied in existing literature. The goal of this work is to measure the readability of ingredients lists on branded food products in the United States using existing metrics. We examined ingredients lists for all 495,646 products listed in the USDA Food Data Central database using four existing readability measures for text written in natural language. Each of these indices approximates the grade level that would be expected to comprehend a text; comparatively, patient consent forms are considered acceptable at an 8th grade reading level or lower. We report a broad variability for in readability using different metrics: ingredients lists recorded at a 9th grade reading level or higher to comprehend are found at rates of 16.5% (Automated Reading Index) to 74.9% (Gunning-Fog Index). Ingredients lists recorded at a 10th grade reading level or higher to comprehend are found at rates of 84.2% (using FRE Index). These results demonstrate the need to further explore how ingredients lists can be measured for readability, both for the purposes of consumer understanding as well as for supporting future nutrition research involving text mining.

Introduction

Nutrition is a known factor in the prevention and management of chronic diseases and impacts both short- and long-term health outcomes. Food label use is an important facet for many consumers when making nutritional choices for themselves and their families. In the United States, a food product label or food label is required to have 5 components, including: (1) product name or identity, (2) the package contents, (3) contact information for the food manufacturer, (4) the Nutrition Facts panel (NFP), and (5) a list of ingredients sorted in order of decreasing composite weight of the product.

Over 75% of consumers report that they read the ingredients list on the product label “sometimes” or “often” when they purchase a product for the first time per the 2019 Food Safety and Nutrition Survey Report (FSNSR). One major reason consumers read food labels is to identify unfavorable and potentially harmful ingredients. For example, the 2019 FSNSR indicates that consumers largely want to detect the presence of artificial ingredients in a food product: 47% of consumers surveyed were either very or extremely concerned about the use of artificial ingredients. Other reasons for focus on the ingredients list is to avoid or reduce intake of certain nutrients or ingredients, such as those with food allergies, individuals with cardiac disease managing their sodium, or a parent nursing an infant with food intolerances.

Poor nutrition is one of the key underlying causes of heart disease and improving nutritional literacy has been suggested as a way for individuals with heart disease improve their condition [1]. A study using a Healthy Eating Index (HEI) on US community-dwelling adults aged 70-79, found that 79.7% had a diet ranked as “poor” or “needs improvement” [2]. Approximately 25% of participants in this study later developed malnutrition. A more recent 2019 study found that 19% of US adults reported a food allergy, and around 11% of US adults are reported have a physician-diagnosed food allergy. 51.1% of physician-diagnosed food allergies were classified as severe, with 24% of individuals with physician-diagnosed food allergies having epinephrine pen prescriptions [3]. It is critically important for food labels to be understandable to the consumer at the point of first purchase.
Nutrition literacy can be generally described as the capacity an individual has for gathering and synthesizing nutrition information to make healthy decisions in their daily lives [4]. Higher nutrition literacy is associated with positive health behaviors, including making healthier dietary choices (i.e. eating more fruits and vegetables), and increasing one’s daily activity. Nutritional literacy falls under the larger umbrella term of health literacy, or one’s ability to make decisions impacting their overall health (Figure 1). Lower health literacy has been demonstrated to be associated with increases in emergency room visits and hospitalization, as well as higher mortality rates, and also with lower rates of preventative screening, medication adherence, label interpretation, and health messaging comprehension [5].

**Figure 1.** An overview of the relationship between health literacy, nutrition label literacy, and how they are measured against the 5 required components of a food product label in the US.

One important measure of nutritional literacy is usage of the food product label, or the frequency with which one uses a food label. A 2015 review of the effects of nutritional literacy on food label usage found that consumers who are familiar with how to read a food label are more likely to use its information to make healthy decisions [6]. Food label usage predicted dietary quality in 18- to 29-year-old persons (n=103), with increased usage expected to improve health outcomes such as dietary quality [7]. Women are known to be more frequent users of food labels [8], and there is some evidence to suggest that consumers identifying as Hispanic are frequent users as well [9]. Food label use is especially important for those trying to avoid or manage the incorporation of a nutrient or ingredient in their diet: Older adults are also more likely to avoid foods associated with known personal health issues than younger adults [9]. Low food label usage was found to be associated with difficulty following gluten-free diets for individuals with diagnosed Celiac Disease or Gluten Sensitivity [10]. The 2019 FSNSR notes that consumers read the food label upon first purchase, indicating that label readability is an important facet of consumer behavior [11]. Eighty-three percent of consumers checked the ingredient list upon first purchase of a product [11]. Therefore, it is critically important for food labels to be readable at the point of purchase so that consumers will be encouraged to use them.

In the United States, a food product label or food label is required to have 5 components, including: (1) product name or identity, (2) the package contents, (3) contact information for the food manufacturer, (4) the Nutrition Facts panel (NFP), and (5) a list of ingredients sorted in order of decreasing composite weight of the product. Consumer understanding of the composite food product label has been found to be correlated with income and education, but even highly educated subjects display difficulty understanding a food label (n=100) [12]. The NFP is definitively the most studied component of a food label in current literature, and was recently updated in 2016 to help consumers better understand the dietary implications of a product. Recent studies on the NFP have found evidence that supports a relationship between NFP use and positive dietary behaviors in young adults [8], in prediabetic adults [13], and in Latinx adults diagnosed with Type II diabetes [14]. However, readability of ingredient lists on a food label has not been extensively studied. The main purpose of this work is to explore the way ingredients are currently presented to consumers on packaged food products in the United States using existing measures of readability.
Methods

A flow diagram of the overall approach used in this method is shown in Error! Reference source not found.. The October 2020 release of the Branded Foods dataset from the USDA Food Data Central database [9] was downloaded via https://fdc.nal.usda.gov/download-datasets.html as a .CSV file. Information for a total of 498,182 products is contained in the Branded Foods dataset. Ingredients lists that were empty \( (n = 2,536, 0.51\%) \) were removed, leaving a total of 495,646 food products to be analyzed. Ingredients lists for these products were pulled from the CSV file, tokenized, and analyzed using the readability [1] and Natural Language Toolkit [2] libraries with Python version 3.6.10. The full process and Python scripts used to perform these tasks, along with data availability, can be found at https://github.com/kmcooper/il_readability_existing_measures. Ingredients lists were unaltered other than tokenization for the purpose of readability analysis. The goal was to analyze the readability of an ingredients list as it would be observed by the consumer without modification.

Readability Measures

Readability measures are used to determine the difficulty with which one can expect when reading a selection of text written in natural language. A number of readability metrics exist already; most of these incorporate factors such as word length, word complexity, number of syllables, text length, and others to determine readability, and are applied to documents such as manuals, textbooks, and patient consent forms, where an individual would be reading the document for comprehension. For context, in healthcare it is generally accepted that documents should be written at an 8th grade reading level or lower to be considered acceptable for documents needing to be written in plain language, such as consent forms [15].

In this analysis we examined ingredients lists of 496,646 distinct food products listed in the Branded Foods dataset from the USDA FoodData central database using four existing readability measures: the Flesh-Kincaid Reading Ease, Gunning-Fog, LIX, and the Automated Readability Index (ARI). Each of these indices approximates the grade level that would be expected to comprehend a text. After readability measures were calculated for each ingredients list, metrics were aggregated into grade levels according to the scale given by each measure; for example, the ARI measure ranks from Kindergarten to College. Each of those ranks are described below.

The Automated Readability Index (ARI): The Automated Readability Index or ARI was developed in 1967 by Smith and Senter [16] and has been used to measure readability of online websites for consumers on topics related to epilepsy [17], otolaryngology [18], breast lesions [19], hip surgery [20], and privacy policies [21]. The ARI measure is based on the number of words per sentence in text written in natural language, as well as characters per word to approximate word complexity [16]. The ARI index reports measures that align with readability from Kindergarten through College Students and Professorial Levels [16].

The Flesch-Kincaid Reading Ease Index (FRE): The Flesch-Kincaid Reading Ease Index, or FRE, was developed in 1975 to help the US Navy author technical documents and ranks texts from 5th grade up through College levels (College, College Graduate, and above) [22]. It is a popular index for measuring the readability of consent forms, with notable use in validating consent form reading levels for vulnerable individuals [23], for those participating in DNA sequencing analyses [24], and HIPAA compliant consenting materials [15], [25]-[27].
The Lasbarhetsindex Swedish Readability Formula (LIX): The Lasbarhetsindex Swedish Readability formula, or LIX, was developed in 1983 for measuring readability of newspapers and is applied to rank texts from 1st grade up through 12th grade, in addition to a final College reading level [28]. The LIX score is based on the number of words, and periods in a text, as well as word length (i.e. a “long word” is sometimes defined as one having 6 or more characters). It is noted for its applicability in readability analyses that are not necessarily rooted in syllable count [29] and has been used to measure readability in consent forms and online health information [30], [31], similar to other measures presented here.

The Gunning-Fog Readability Index: The Gunning-Fog Readability formula was developed in 1952 for measuring readability of newspapers upon first exposure to the text and is applied to ranks texts from 6th grade up through College Graduate level [32]. The Gunning-Fog index is based on word count, sentence count, and complex word count in a text, where complex words contain $\geq 3$ syllables. Similar to the other measures reported here, it is a popular measure of readability for health information presented to consumers and patients alike [33]-[37]. Mean and median scores as well as standard deviation for all readability metrics are automatically reported by the readability Python library. Conversion of readability scores to grade level is metric-specific and was performed using Python. Code to perform conversions for the measures reported here is also available on our Github repository for this project at [https://github.com/kmcooper/il_readability_existing_measures](https://github.com/kmcooper/il_readability_existing_measures).

Results

The mean and median scores for the four readability metrics used in our analyses are reported in Table 1. Mean, median scores for the 4 readability metrics, including mean reading level according to each metric, and the percent total of 495,696 ingredients lists analyzed falling at or below an 8th or 9th grade reading level. Table 1. Broadly, the median score reported by the four indices ranges between the 6th grade reading level and a college reading level, with little consensus between metrics. The Automated Reading Index (ARI) is by far the most magnanimous measure, reporting that ingredients lists for 83.5% products fall at or below an 8th grade reading level. By contrast, the Flesch-Kincaid Reading Ease metric is by far the most conservative measure, reporting that ingredients lists for 84.23% of products fall at or below a 10th grade reading level (FRE groups 8th and 9th grade together). At large, three of the four measures reported describe a mean reading level of 8th grade or higher for ingredients lists on branded foods.

Table 1. Mean, median scores for the 4 readability metrics, including mean reading level according to each metric, and the percent total of 495,696 ingredients lists analyzed falling at or below an 8th or 9th grade reading level.

<table>
<thead>
<tr>
<th></th>
<th>Mean Score</th>
<th>Mean Reading Level</th>
<th>Median Score</th>
<th>Std. Dev</th>
<th>% of IL at/below 8th Grade Level</th>
<th>% of IL above 8th Grade Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gunning-Fog Index</td>
<td>11.983</td>
<td>11-12th Grade</td>
<td>11.916</td>
<td>5.611</td>
<td>25.14%</td>
<td>74.86%</td>
</tr>
<tr>
<td>ARI</td>
<td>6.539</td>
<td>6th-7th Grade</td>
<td>6.474</td>
<td>3.357</td>
<td>83.50%</td>
<td>16.50%</td>
</tr>
<tr>
<td>LIX</td>
<td>37.295</td>
<td>8th Grade</td>
<td>36.694</td>
<td>15.558</td>
<td>61.57%</td>
<td>38.43%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean Score</th>
<th>Mean Reading Level</th>
<th>Median Score</th>
<th>Std. Dev</th>
<th>% of IL at/below 9th Grade Level</th>
<th>% of IL above 10th Grade Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flesch- Kincaid Reading Ease</td>
<td>37.931</td>
<td>College Level</td>
<td>37.026</td>
<td>26.889</td>
<td>15.77%</td>
<td>84.23%</td>
</tr>
</tbody>
</table>

The reading level distributions for each of the reported metrics are shown below in Figures 2-6. These distributions are presented as a total count of ingredients lists reported at each reading level using the given metric. It is not effective to compare metrics by their distributions due to their different reading level categorizations, however, visualizing their distributions in this way highlights the disagreement between metrics. For example, both the LIX and the Gunning-Fog indices capture ingredients lists at higher-than-normal-levels at the extreme ends of their distributions, and the Flesch-Kincaid Reading Ease is skewed toward the college ranked reading levels.
Figure 3. Ingredient list reading level distribution using the ARI measure. All ILs were analyzed using the ARI measure and categorized into grade levels between Kindergarten and Professor using the ARI. According to this measure, 83.5% of ingredients lists are presented at an 8th grade or lower reading level.

Figure 4. Ingredient list reading level distribution assessed using the FRE measure. All ILs were analyzed using the FRE measure and categorized into grade levels between 5th Grade and College Graduate. According to this measure, 15.8% of ingredients lists are presented at an 10th grade or lower reading level.

Figure 5. Ingredient list reading level distribution assessed using the LIX measure. All ILs were analyzed using the LIX measure and categorized into grade levels between 1st Grade and College using the LIX. Using LIX, 61.6% of ILs are presented at an 8th grade or lower reading level.
Discussion

Specifically, this work examines the readability of ingredients lists using existing readability indices, where readability can be generally defined to measure the level of education needed for an individual to understand a list of ingredients. There are limited studies on the readability of food products labels and few studies examining the readability of other commercial products, such as cosmetics [38] and dietary supplements [39]. This study highlights the documented challenges of analyzing the readability of ingredients lists using existing metrics; generally, text mining challenges such as these are already somewhat recognized in food composition research and information systems [40]-[43]. The disagreement between these measures used in this research suggests that existing readability metrics may not be sufficient to infer readability of ingredients lists, which are not traditionally written as natural language. Better metrics for measuring readability of ingredients lists can be developed to more accurately reflect the text structure of ingredients lists versus text written in natural language.

Another challenge directly impacting the consumer is that food production companies differ, sometimes widely, in the terms they use to present ingredients on a food product label, the preparation of those ingredients, and the purpose for which an ingredient is used. For example, soy lecithin is sometimes used as an emulsifier, in other foods it is used as a flavor protection agent. Additionally, a label might say “chocolate”, “milk chocolate”, or “chopped chocolate”, which could all be the same product, but prepared and or presented in different ways. A 2017 study on nutrition modeling recommended that food labels should denote (1) the ingredients in the food product itself and (2) how the ingredient was prepared to be used in the product, such as “chopped”, “raw”, or “pureed” [44]. There is building evidence that food preparation affects the gut microbiome[45], [46], which has implications for health outcomes. As the body of research in consumer access and use of nutritional information continues to grow, it is expected that there will be consumer-demand for information on ingredient preparation and provenance in food labeling policy. A 2014 review on challenges facing food science acknowledges the need for multidisciplinary teams to address these and other challenges, incorporating the fields of computer science, text mining, and informatics [47]. Resources and interdisciplinary teams are necessary to create consumer-centered information systems.

Conclusions

More research is needed to fully understand the consumer experience with ingredients lists on the food products label. With the digitization and aggregation of information on food products made and distributed in the United States, it is possible to apply informatics approaches that will support consumers in the pursuit of healthy dietary behaviors, like natural language processing. The study of the readability and consumer experience of ingredients lists can reveal insights into how consumers use and experience food products in their daily lives, but remains an under-utilized tool for improving health.
Limitations

This work uses existing metrics for measuring readability but it is not readily clear when the application of readability metrics becomes inappropriate, such as in the application of the metrics to a list of terms. There are a number of readability metrics that were not used in this work because they count sentences and punctuation which are typically used in narrative text [48, 49] that would not be appropriate in application to an ingredients list. The four metrics used here were chosen as they focus largely on word and character counts which are applicable in this context, however, future work could focus on the development of a readability metric specific to ingredients lists.

Acknowledgements

This work was supported with a Graduate Research and Creative Activity (GRACA) grant from the University of Nebraska at Omaha as well as a pilot award from the Business, Informatics, and Gerontology Big Ideas Award.

References


Using Natural Language Processing to Classify Serious Illness Communication with Oncology Patients

Anahita Davoudi, PhD, MS, MS, Hegler Tissot, PhD, Abigail Doucette MPH, Peter E. Gabriel MD, Ravi Parikh MD, MPP, Danielle L. Mowery, PhD, MS, MS, Stephen P. Miranda, MD

1 Biostatistics, Epidemiology, & Informatics, 2 Abramson Cancer Center, 3 Radiation Oncology, 4 Medical Ethics and Health Policy and Medicine, 5 Informatics for Biomedical Informatics, 6 Neurosurgery, 7 Computing & Information Science; University of Pennsylvania; 8 Information Science, Drexel University

Abstract

One core measure of healthcare quality set forth by the Institute of Medicine is whether care decisions match patient goals. High-quality “serious illness communication” about patient goals and prognosis is required to support patient-centered decision-making, however current methods are not sensitive enough to measure the quality of this communication or determine whether care delivered matches patient priorities. Natural language processing (NLP) offers an efficient method for identification and evaluation of documented serious illness communication, which could serve as the basis for future quality metrics in oncology and other forms of serious illness. In this study, we trained NLP algorithms to identify and characterize serious illness communication with oncology patients.

1 Introduction

For patients with cancer to receive care that aligns with their values, their clinicians must effectively explore their care preferences. Documentation of patient-specific goals and prognostic information earlier in the illness trajectory is critical for assessment of shared decision-making, goal-concordance, and healthcare utilization in oncology. High-quality serious illness communication (SIC) can enhance quality of life and goal-concordant care, while inadequate SIC is associated with greater psychosocial distress and aggressive end-of-life care that may be incongruent with patient preferences.

There is consensus that SIC documentation itself is a core quality measure that supports goal-concordance and therefore must be evaluated. However, it is well-documented that traditional forms of SIC documentation, including advance directives, are under-utilized and inconsistently applied, making it difficult to track SIC across inpatient and outpatient settings. High-quality SIC in oncology includes discussion of patient goals, prognosis, code status, and advance care planning. Routine assessment of documentation on these four topics is difficult because this information often exists as free-text in the electronic health record (EHR), which requires time-intensive, manual chart review to identify and abstract.

1.1 Natural Language Processing

Natural language processing (NLP) can offer an efficient, accurate alternative for identification of SIC in the EHR, and has been used to identify care-planning discussions and palliative care delivery. Despite early progress, more sophisticated approaches are needed to classify and evaluate SIC documentation. At this time, NLP approaches for identification of SIC predominantly rely on keywords derived from chart review. Such lexical approaches lend themselves well to identification of specific care-planning metrics, such as documentation of code status (e.g. “full code”, “do not resuscitate”) and discussions about hospice (e.g. “comfort measures only”). However, these algorithms are limited in their ability to capture nuanced documentation about patient priorities and prognostic communication, which does not always rely on representative keywords, is less prevalent in the EHR, and is highly variable from clinician to clinician, limiting identification of this documentation at scale.

Machine learning approaches that expand beyond keywords may support more accurate and automatic identification of these two critical SIC domains. In this study, we sought to leverage weakly-labeled EHR data from oncology...
patients to develop and validate an NLP algorithm that automatically identifies and classifies SIC documentation about prognosis and goals.

2 Methods

This study was approved by the University of Pennsylvania Institutional Review Board, protocol #842930. We first collected a weakly annotated dataset of free-text entries containing SIC documentation, and then trained several machine learning algorithms to automatically classify SIC documentation by domain and subdomain. Finally, we characterized the features associated with each SIC subdomain.

2.1 Dataset and Schema

In 2018, the University of Pennsylvania Abramson Cancer Center implemented the Serious Illness Care Program (SICP) developed by Ariadne Labs, a multi-component, systems-based intervention designed to enhance timing, frequency, and quality of SIC in oncology.19,20 Oncology clinicians are encouraged to document SIC using an EHR module, which generates a semi-structured “Serious Illness Conversation” note with subheadings by SIC domain. Prior to this implementation, all clinicians at Abramson Cancer Center were instructed to use an “Advanced Care

Table 1: Serious Illness Communication Subdomains for Prognosis and Goals.

<table>
<thead>
<tr>
<th>Subdomain</th>
<th>Prompt</th>
<th>Responses</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prognosis Domains</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prognostic Understanding (PU)</td>
<td>What is your understanding now of where you are with your illness?</td>
<td>Overestimates prognosis; Accurate understanding of prognosis; Underestimates prognosis; No understanding of prognosis;</td>
<td>“He knows he only has weeks to live.”</td>
</tr>
<tr>
<td>Information Preferences (IP)</td>
<td>How much information about what is likely to be ahead with your illness would you like from me?</td>
<td>Patient wants to be fully informed; Patient wants to be informed of big picture, but not details; Patient wants some information, but no “bad news”; Patient prefers information to be shared with ***</td>
<td>“She prefers weekly prognosis updates.”</td>
</tr>
<tr>
<td>Prognostic Communication (PC)</td>
<td>Information shared with patient about prognosis</td>
<td>Uncertain prognosis; Possibility of getting sick quickly; Limited time, may be as short as ***; May never get stronger or regain function</td>
<td>“He had questions about prognosis.”</td>
</tr>
<tr>
<td><strong>Goal Domains</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main Goals (MG)</td>
<td>If your health situation worsens, what are your most important goals?</td>
<td>Live as long as possible; Pursue every available treatment; Avoid hospitalizations/maximize time at home; Not be a burden/maintain independence; Be physically comfortable; Be mentally aware; Spent time with family</td>
<td>“The patient wants to live to see his daughter’s wedding.”</td>
</tr>
<tr>
<td>Fears/Worries (FW)</td>
<td>What are your biggest fears and worries about the future with your health?</td>
<td>Pain or other symptoms; Loss of control or dignity; Burdening others; Family concerns; Financial concerns</td>
<td>“He worries about becoming dependent.”</td>
</tr>
<tr>
<td>Strengths (ST)</td>
<td>What gives you strength as you think about the future with your illness?</td>
<td>Friends/family; Faith/spirituality; Prior experience with adversity</td>
<td>“Support of family and friends.”</td>
</tr>
<tr>
<td>Critical Abilities (CA)</td>
<td>What abilities are so critical to your life that you cannot imagine living without them?</td>
<td>Living independently; Being mentally aware; Interacting with others; Dressing, bathing, toileting; Eating and drinking</td>
<td>“Maintaining ability to interact with others is important.”</td>
</tr>
<tr>
<td>Tradeoffs (TO)</td>
<td>If you become sicker, how much are you willing to go through for the possibility of gaining more time?</td>
<td>Anything to prolong life incl. life support &amp; ICU care; Limited hospitalizations, some testing and treatments; No further life-prolonging care</td>
<td>“She doesn’t want to experience any major side effects unless there is a high likelihood of therapeutic benefit.”</td>
</tr>
<tr>
<td>Family/Friends (FF)</td>
<td>How much does your family know about your priorities and wishes?</td>
<td>Extensive discussion with family about goals and wishes; Some discussion, but incomplete; No discussion, but plans to address these issues; No discussion, wants help talking to family; Does not want family informed</td>
<td>“We talked about how he and his wife might begin to have conversations with their daughters.”</td>
</tr>
</tbody>
</table>
Planning” note template for free-text documentation of SIC. In the new “Serious Illness Conversation” note template, there are nine SIC domains, each with a menu of preset responses to choose from, based on the information acquired from the patient, as well as an optional, free-text comment box to insert free-text that provides more detail. The “Serious Illness Conversation” template outlines nine SIC subdomains, three regarding prognosis and six regarding goals. The SIC subdomains including prompts, the structured responses and fictitious, but exemplar free-text statements within the “comments” are listed in Table 1.

For this study, we queried the Penn Medicine cancer registry for all patients with stage III or IV cancer who were treated across all Penn-affiliated locations and whose records contained “Serious Illness Conversation” notes within our EPIC Clarity electronic data warehouse. Our cohort consisted of 3563 total patients from which 5,145 notes were identified, containing a total of 8,695 distinct “responses” and “comments”. The dataset was randomly split into 6,964 entries (80%) for training and 1,731 entries (20%) for testing.

2.2 Serious Illness Communication Classifier Development and Evaluation

Each entry from our dataset was preprocessed using the spaCy library: removing punctuation, eliminating stopwords, reducing case, and encoding n-grams (n=1-3 words).† We also encoded lexical categories using Empath.‡ Empath is an unsupervised tool trained using connotations between words leveraging a neural embedding derived from over 1.8 billion words of modern fiction. Empath can be utilized to generate lexical categories and contains over 200 built-in, topical and emotional categories generated from common dependency relationships in ConceptNet and Parrot. Topical categories include money, home, work, religion, health, death, etc. Emotional categories include sadness, anger, positive emotion, negative emotion, etc. Terms within both categories were verified using Amazon Mechanical Turk reviewers.

Using the comments from our training dataset, we trained four machine learning algorithms: Logistic Regression, XGBoost, BERT, and Bio+Clinical BERT.

- **Logistic Regression** learns a logit regression model that explains the relationship between the features and the class. Our model uses exhaustive grid search and L1-regularization to optimize performance while reducing the likelihood of over-fitting due to few training examples, many irrelevant features, and a large number of parameters.

- **XGBoost (extreme gradient boosted trees)** is a gradient descent algorithm that learns to predict the residual errors of prior models while minimizing the loss of adding new models before unifying models to make a final class prediction. These boosting models optimize speed and accuracy while reducing the likelihood of overfitting by penalizing trees and applying proportional shrinking of leaf nodes. The booster parameter was set to gblinear.

- **BERT (bidirectional encoder representations from transformers)** are pretrained deep bidirectional representations from unlabeled text fine-tuned using a “masked language model” that combines both left and right contexts. We leveraged the pre-trained BERT model to provide the vector representations of the embedding sets which were passed to a drop out layer (drop rate of 0.5); the default parameters were used.

- **Bio+Clinical BERT** is a BERT model that leverages pre-trained language representations initialized from BioBERT, a BERT model generated from PubMed article abstracts and PubMed Central article full texts and then fine-tuned using a clinical corpus of notes (e.g., discharge summaries, physician notes, nursing notes, radiology reports, etc.) from the Medical Information Mart for Intensive Care (MIMIC version III) dataset. The default parameters were used.

Using a data-driven approach, we trained each of the four algorithms as a SIC classifier to classify each comment according to SIC domains of goals or prognosis. As a proof-of-concept, we also trained only the logistic regression algorithm to classify 1 out of 9 possible SIC subdomains.

---

† [https://spacy.io/universe](https://spacy.io/universe)
‡ [https://github.com/Ejhfast/empath-client](https://github.com/Ejhfast/empath-client)
§ [https://github.com/googleresearch/bert](https://github.com/googleresearch/bert)
2.3 Serious Illness Communication Subdomain Characterization

For each SIC subdomain (e.g., the Goals domain has a subdomain of Strengths), we applied chi-square feature selection and selected the most significantly associated features (n-grams and Empath categories with \( p < 0.05 \)) associated to each class and applied a log-10 transform to each feature’s p-value. We visualized the associated features by transformed p-value using WordCloud. We also report and compare the distribution of Empath categories across subdomains.

3 Results

In this study, we leveraged weakly-labeled EHR data from oncology patients to develop and validate an NLP algorithm that automatically identifies and classifies SIC documentation about prognosis and goals.

In Figure 1, we report the percent distribution of comments by subdomain across the full corpus. Among all free-text comments, 61.4% belonged to the domain goals and 38.6% belonged to prognosis. For subdomains within goals, we observed proportions ranging from 6.3% Strengths to 13.2% Tradeoffs. For subdomains within prognosis, we observed proportions ranging from 7.4% Information Preferences to 17.1% Prognostic Communication.

3.1 Serious Illness Communication Classifier Development and Evaluation

In Table 2, we report the predictive performance of each machine learning algorithm on the test set. The highest F1-score was achieved by XGBoost for both prognosis (0.86) and goals (0.91). XGBoost achieved the highest precision for prognosis (0.86) and highest recall for goals (0.92). Conversely, Bio+Clinical BERT achieved the highest recall for prognosis (0.86) and highest precision for goals (0.92). In terms of deep learning algorithms, for both prognosis and goals, we observed higher recall (+6 points, +8 points) and precision (+16 points, +2 points) using Bio+Clinical BERT over BERT, respectively.

<table>
<thead>
<tr>
<th>Prognosis</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression (baseline)</td>
<td>0.81</td>
<td>0.85</td>
<td>0.83</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.85</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>BERT</td>
<td>0.80</td>
<td>0.64</td>
<td>0.71</td>
</tr>
<tr>
<td>Bio+Clinical BERT</td>
<td>0.86</td>
<td>0.80</td>
<td>0.83</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Goals</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression (baseline)</td>
<td>0.91</td>
<td>0.89</td>
<td>0.90</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.92</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>BERT</td>
<td>0.80</td>
<td>0.90</td>
<td>0.84</td>
</tr>
<tr>
<td>Bio+Clinical BERT</td>
<td>0.88</td>
<td>0.92</td>
<td>0.90</td>
</tr>
</tbody>
</table>
In Table 3, we report the predictive performance of the logistic regression algorithm on the test set for each SIC domain. Among prognosis, the highest F1-score was achieved for Prognostic Understanding (0.61) followed by Prognostic Communication (0.60). Among goals, the highest F1-score was achieved for Critical Abilities (0.71) followed by Strengths (0.65) and Tradeoffs (0.63).

Table 3: Logistic Regression SIC classifier performance by SIC subdomain on the test set.

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prognosis</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prognostic Understanding</td>
<td>0.58</td>
<td>0.64</td>
<td>0.61</td>
</tr>
<tr>
<td>Information Preferences</td>
<td>0.44</td>
<td>0.42</td>
<td>0.43</td>
</tr>
<tr>
<td>Prognostic Communication</td>
<td>0.57</td>
<td>0.63</td>
<td>0.60</td>
</tr>
<tr>
<td><strong>Goals</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main Goals</td>
<td>0.52</td>
<td>0.68</td>
<td>0.59</td>
</tr>
<tr>
<td>Fears/Worries</td>
<td>0.62</td>
<td>0.40</td>
<td>0.49</td>
</tr>
<tr>
<td>Strengths</td>
<td>0.75</td>
<td>0.58</td>
<td>0.65</td>
</tr>
<tr>
<td>Critical Abilities</td>
<td>0.70</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>Tradeoffs</td>
<td>0.60</td>
<td>0.65</td>
<td>0.63</td>
</tr>
<tr>
<td>Friends/Family</td>
<td>0.47</td>
<td>0.27</td>
<td>0.35</td>
</tr>
</tbody>
</table>

3.2 Serious Illness Communication Subdomain Characterization

In Figure 2, we present the most informative n-grams and Empath categories associated with each SIC subdomain. Features with high associations (low p-values) to a subdomain are larger in the WordCloud. Notable features by prognosis subdomain include: prognostic understanding (prognosis, understanding, curable, helpful, understands disease, know), information preferences (big_picture, detail, fully_informed), and prognostic communication (limited_time, short_months). Notable features by goal subdomain include: goals (spend_time, quality, home, live_long_possible, comfortable), fears/worries (fear, concern, loss, dying, suffering), strength (strength, friends, catholic, spirituality), critical abilities (walking, taking_care, reading, independently, self), tradeoffs (intubation, dnr, code, life_support, would_want, measures, considering), and friends/family (family, extensive, discussion, wife, daughter, conversation).

Figure 2: Wordclouds for Prognosis (1st row) and Goal (2nd/3rd row) Subdomains.
In Figure 3, we present the frequency distribution of observed Empath categories according to each SIC subdomain; any Empath category with less than 200 total counts is not shown. We observed 194 of the more than 200 built-in Empath categories in our full dataset. The most common Empath categories observed across the corpus include: *health*, *medical_emergency*, *positive_emotion*, *family*, *children*, *death*, *negative_emotion*, and *communication*.

![Figure 3: Frequency distributions of Empath category by subdomain; any Empath category with less than 200 total counts is not shown.](image_url)
4 Discussion

Accurate, reliable, and scalable identification of serious illness communication in the EHR is critical for measuring and improving the quality of oncology care.

4.1 Serious Illness Communication Classifier Development and Evaluation

We successfully utilized semi-structured EHR data to develop an NLP algorithm capable of classifying documented entries by SIC domain with high fidelity, identifying text about prognosis (0.86) and goals (0.91). Overall, performance of the classifier across all subdomains ranged from reasonable (0.71) to high (0.91). This study demonstrates promise for identifying SIC—and extracting more complex semantic constructs out of the EHR—without relying on keyword-based approaches. Automated methods for characterizing SIC documentation at scale are limited because clinical notes are variable and often unique to specific clinical situations, which narrow, lexical approaches might fail to anticipate. Here, we leveraged semi-structured data as “weakly labeled” text for classifier training, not only eliminating the need for annotation, but also enhancing the predictive power of the classifier by generating n-grams reflective of diverse lexical categories for training.

The SIC classifier was less effective at discerning individual subdomains within goals and prognosis likely because each subdomain represents overlapping constructs with shared terminology. The “Serious Illness Conversation” template was designed as a communication aid for clinicians to elicit patient values and support prognostic communication, so it is likely that individual subdomains are interrelated for the same patient. While distinguishing between subdomains may be less critical for clinicians using the template at the point of care, enhancing discrimination within each domain would improve classifier performance in free-text clinical notes going forward. It is possible that clinicians inadvertently documented information under the wrong subdomains, which would confound the classifier’s ability to distinguish between them. Notably, classifier performance identified goals better than prognosis, despite a broader range of subdomains, although this may be because the majority of documentation (61.4%) is about goals (Figure 1).

The next phase of this research will involve testing and validation of the algorithm’s ability to identify and classify SIC among undifferentiated clinical notes containing unstructured free-text. During this process, further work will be needed to explore why more supervised ML methods (e.g. logistic regression, XGBoost) outperformed deep learning algorithms in this study. Many of the comments and responses used for training and testing consisted of telegraphic phrases, so it may be that deep learning approaches will be more successful in further testing on longer free-text entries, where more contextual features are present. In fact, for both the prognosis and goals domains, we observed higher recall and precision using Bio+Clinical BERT over BERT, respectively, supporting the hypothesis superior performance can be achieved in part through the use of pre-trained models based on clinical documentation.

4.2 Serious Illness Communication Subdomain Characterization

Analysis of the most predictive features for each subdomain demonstrates that these features conceptually map very closely to the theme of each subdomain (Figure 2) while reflecting a broad range of etymologic categories (Figure 3), illustrating the utility of incorporating lexical terms and semantic grouping into the classifier training process. For instance, features associated with documentation about prognosis captured non-specific (terminal, curable, incurable) and time-based prognostication (limited time, short months, short weeks); the degree of prognostic understanding (overestimates, accurate, know, good understanding, understands cancer); how this information was communicated (office, internet/email) and to what extent (detailed, big picture, fully informed).

Similarly, subdomains within goals, features describe specific wishes or priorities (wedding, quality time) and even place of final rest (home, die house). Both negative and positive sentiments were reflected. For example, fears/worries contain features of negative emotion (worried, afraid, suffering, weakness, fearful, nervousness, concern, sadness); strengths contain features of positive emotion (comfortable, support, strong). Sources of strength include one’s faith (catholic, spirituality, divine) and support system (children, friends family). Critical abilities highlight activities of leisure (sports, walking, play, art, driving, reading, working) and daily living (living, breathing, eating) as well as terms related to autonomy (self, independent, dependence). To achieve these goals and maintain critical abilities, preferences for life-sustaining treatments were also captured, including code status (intubation, cpr, dnr, life support, 174
resuscitation, ventilation, full code). Both prognosis and goals were often shared with individuals representing family (wife, husband, son, daughter, sister, son) and those in decision-making roles (poa, power of attorney).

4.3 Clinical Applications

If further validated, the clinical implications of this SIC classifier are compelling. While documentation about goals of care and prognostic communication are known process measures of high-quality palliative care delivery, SIC is poorly captured by administrative claims data, and manual review of individual patient records is laborious and impractical at the population level—yet quality measurement in palliative care is still highly dependent on these two methodologies. A validated SIC classifier would offer a powerful tool for more useful quality metrics in oncology, either by evaluating communication quality or developing personalized measures of goal-concordance. Reliably tracking patient goals would provide useful context for assessing appropriateness of healthcare utilization, and characterizing narrative arcs in the disease trajectory could help frame quality improvement initiatives and psychosocial interventions during serious illness. In healthcare operations, explainable AI for logistic regression or XGBoost could even be used to inform clinician-facing EHR tools at the point of care, perhaps by visualizing positive coefficients or SHAP values across terms and Empath categories.

Although these results are preliminary, the methodology employed here allows for greater real-world applicability than other reports of NLP approaches to SIC identification thus far, which have all been keyword-based. Recent applications of these methods have seen success in patient groups drawn from pragmatic trials in oncology, but due to their lexical basis these efforts have required manual annotation of hundreds of clinical notes, and may be weighted towards inpatient admissions or medical crises requiring treatment decisions. Our method may lay the foundation for more nuanced identification of patient-specific priorities and prognostic communication more upstream in the disease trajectory, which would have significant utility across a wide array of clinical contexts.

4.4 Limitations and Future Work

This study has notable limitations. The SIC classifier was trained using semi-structured Epic EHR modules, which limits the replicability of this work in other settings where source text enriched with SIC may be lacking. Moreover, most SIC documentation in oncology exists within free-text clinical notes, requiring discrimination between relevant and irrelevant text. Performance may suffer in population-level datasets where SIC represents a minority of clinical documentation. In the next phase of this research, classifier training must be enhanced for application to free-text clinical notes. As a first step, we are actively applying the XGBoost classifier for goals and prognosis to sentences from free-text, de-identified clinical neuro-oncology notes that were manually annotated as part of ongoing research and quality improvement efforts at our institution. Preliminary results are promising (goals – F1: 0.72; prognosis – F1: 0.70). We anticipated a drop in performance because the schema used for annotation of these notes introduced additional subdomains under goals and prognosis for greater precision. Additional training and tuning will be needed to optimize the classifier for free-text notes and additional subdomains, which we plan to complete in the near future by leveraging free-text “Advance Care Planning” (ACP) notes obtained from our EHR.

This classifier is based on documentation from a limited number of oncology clinicians at one institution requiring further study in larger, more diverse populations to assess generalizability. In the future, we aim to better understand how patient preferences evolve over time, as well as any similarities or differences in SIC across gender, race, ethnicity, and culture.

Conclusion

Here we describe a novel application of NLP for classifying SIC documentation in oncology. If further validated, such an algorithm can retrieve and evaluate SIC documentation in routine clinical practice as a quality metric to assess key clinical and systems priorities in oncology.
Acknowledgements

We extend our gratitude to the University of Pennsylvania for partially supporting this important research through Dr. Mowery’s start-up funding.

References


[38] Reed-Guy L, Alexander TD, Biggiani G, Miranda SP, O’Connor N. Serious illness communication practices in glioblastoma care at an academic medical center. American Society of Clinical Oncology; 2019.


Towards Real-time Bone Drilling Simulation for Anchor Placement in VR Based Arthroscopic Rotator Cuff Surgery Simulation

Furkan Dinc 1, Khalil Oumimoun 1, William Kwabla 1, Sinan Kockara, Ph.D. 1, Tansel Halic, Ph.D. 1, Sreekanth Arikatla, Ph.D. 2, Shahryar Ahmadi, M.D.3
1University of Central Arkansas, Conway, AR, USA; 2Kitware Inc., Carrboro, NC, USA; 3UAMS, Little Rock, AR, USA

Abstract
Arthroscopic Rotator Cuff (ARC) is a minimally invasive surgery of the shoulder. ARC training for surgeons is challenging due to confined space, anatomical complexity, requirement of complex hands-eye coordination skills, subjectivity, and low fidelity in existing training mediums. We therefore offer a virtual reality based photorealistic medical simulation, Virtual Rotator Cuff Arthroscopic Skill Trainer (ViRCAST) for objective training. In this study, as a part of ViRCAST, we introduce a virtual reality-based bone drilling simulation. Bone drilling task is one of the most important tasks that surgeons need to perform before anchor placement in ARC. Realistic simulation of bone drilling with force feedback is complex due to real-time mesh modification and simulation constraints. We introduce a GPU based realtime bone drilling simulation for ViRCAST using an adaptive mesh refinement technique. Our GPU based solution improves the drilling simulation realism by enhancing mesh resolution without sacrificing the simulation performance.

Introduction
Arthroscopy is a minimally invasive surgical procedure that is performed via small incisions in the patient’s skin to examine, diagnose, and repair the injuries inside a joint [1]. Arthroscopic Rotator Cuff (ARC) is a surgical treatment for the group of muscles and tendons that connect the upper arm to the shoulder blade. Surgeons insert pencil-sized instruments with small lens and lighting into the joint and see the anatomy on a 2D monitor screen streamed from arthroscope, which is a small rigid fiber optic camera with light source. The goal of the anchor placement is to increase strength for tensile stress. The surgeon uses multiple anchors that are evenly distributed over the humeral head area to divide the load equally. One of the fundamental skills that a surgeon needs to master for the arthroscopic rotator cuff treatment is bone drilling for anchor hole [2]. Drilling the humeral head prior to placing suture anchors is a critical part of ARC. During the bone drilling step of the surgery, a drill toll is inserted through the cannula at a specific angle using the portals that the surgeon has created. The bone is then drilled enough to place the metallic or resorbable anchor. The screw-in anchors are placed in the drilled part and the torn tendon is sutured to the anchor (Figure 1).

Figure 1. ARC surgery, bone drilling, anchor placement, tendon suturing, and knot tying procedures illustrated. Circular views are arthroscopic views

Conventional ARC surgery education models
Due to non-natural hand-eye coordination, narrow field-of-view, and limited instrument control, ARC training is challenging and difficult to master. Conventional surgery education models except apprenticeship model neither possess the realism to teach the surgeons the joint anatomy nor aid surgical decision-making [3]. The most commonly used learning model is apprenticeship model; however, it incurs high risks on the patient’s health and has no tolerance for mistakes. The challenges and limitations of traditional learning modalities make virtual reality (VR) based solutions a good alternative. To this end, ViRCAST [4] offers a VR based alternative training solution for ARC surgery with realtime photorealistic rendering and haptic based tactile force feedback interactions. ViRCAST provides a realistic immersive VR platform for surgeon trainees to interact with the VR environment and volumetric models through actual surgical instruments to hone their skills.
Motivation
Bone drilling operations in ViRCAST need to be simulated photorealistically with accurate tactile sensation. In order to simulate the bone drilling task in a VR environment, 3D surface alteration algorithms need to be used for proper visualization and haptic feedback. These algorithms also need to perform in realtime (>60Hz for graphics, >700Hz for haptics) for high fidelity [5]. The most common approach for these algorithms is to represent the 3D shape with voxels. A voxel is a 3D representation of a pixel in 3D space. Voxels are preferable for confining volumetric spaces as opposed to the surface models that is only surface representation but void inside. The data stored in each voxel typically are the voxel’s center and its size. The data to be stored can be extended to specify visual or simulation properties; for example, the bone density in surgery simulations. These voxels can be removed at any time with interactions of some surgical instruments such as bone shaver in a 3D scene. The voxel removal step is followed by some surface reconstruction algorithms such as the marching cubes [6]. The next section explores some limitations and known issues with the iso-surface extraction technique.

Isosurface extraction
Isosurface extractors are used for visual access to experimental and theoretical volumetric data, such as medical images and physical simulations. Among those isosurface extraction techniques, the marching cubes algorithm [6] is prominent. It is a sequential-traversal method that has become the reference method for surface reconstruction and 3D contour creation of mathematical scalar fields. Using the marching cubes technique allowed the creation of very high-quality images by generating a set of triangles that approximate a mesh surface. Marching cubes algorithm is de-facto standard isosurface extraction algorithm in scientific visualization. The original marching cubes proposed by Lorensen and Cline [6] has many limitations. The algorithm suffers from what is known as “the hole/crack problem” illustrated in Figure 2. This is because marching cubes approximates isosurfaces with line segments which leads to cracks especially for the curved surfaces.

Figure 2. Hole problem in Marching Cube. There are visible cracks between the polygons.

Another major problem with the original marching cubes algorithm is speed. For a finer 3D volume representation very high resolution voxel grid is required. This drastically reduces its performance due to heavy computational loads. Wilhelms and Van Gelder [7] presented a technique that uses an octree to reduce the number of cubes traversed and hence improve the speed of the algorithm. Although this technique is effective, the results are still not adequate for large number of meshes with high resolution data sets especially if some sort of manipulation of the 3D surfaces is anticipated such as in our case; the bone shaving and drilling tasks.

Currently available iso-surface generation methods suffer from many limitations when it comes to real time large number of high-definition mesh deformations. The application presented in this study involves medical and surgery simulations that require many iso-surface regenerations and mesh deformations. Bone drilling task in VR is a complex application due to its requirements for needing very high speed, 3D calculation of the drilling area, realtime manipulations of vertices and triangles, and calculations for realtime collision detections. To overcome abovementioned issues and satisfy the requirements for VR based real time bone drilling task for ARC surgery, we adapt adaptive mesh refinement. This work aims on improving 3D rendering and generating continuous haptic forces by eliminating cracks on isosurfaces and resolving the performance issues of interactive voxel manipulations by incorporating parallel GPU computing and additional data structures.

Adaptive Mesh Refinement
For our purpose, we use Adaptive Mesh Refinement (AMR) technique introduced by Berger and Colella [8] [11]. AMR has since become one of the most widely used method for scientific simulations. AMR techniques are one of the key technologies for large-scale and high accuracy simulations, that require significant computational power and memory. They are highly effective for simulations that require adaptively changing the shape or resolution of a mesh. The main idea behind AMR techniques is to allow more intensive and refined computations on a specific area of interest (region of interest) on 3D or volumetric domain.

With AMR, it is possible to subdivide an initially coarse domain into adaptively refined (finer) subdomains where the simulation needs the most accurate and finer discretization. This approach significantly reduces the number of discretization elements (e.g. varied sized voxels) and thus, memory storage required to reach a required accuracy. In AMR data, different-resolution grids are layered on top of each other with certain rules in that refinement happens in power-of-two factors. AMR data values are defined only for the center of each cell, and the data also satisfy the rule regarding how refined cells have to align to coarser ones. Even with AMR, when extracting...
an isosurface using the marching cubes method, T-junctions will still lead to visible cracks in an isosurface, even if dangling nodes
(explained in Figure 3) have values that are consistent with the coarse level representation. T-junctions are where two finer surface
pieces meet one coarser piece.

![Figure 3](image)

**Figure 3.** (a) An original face at the boundary, (b) coarse level with red circles representing shared vertices between grids, (c) fine level
with additional gray vertices (squares) representing dangling nodes occur along with the vertices shared between grids (red circles). As
seen in (b) and (c) marching cubes (marching squares in 2D) approximates the isosurfaces (actual contour in (a); green curve) with line
segments which lead to cracks on the surface (yellow regions in (b) and (c)).

As seen from Figure 3, even though cracks get smaller in finer grid representation of an isosurface in AMR, they still persist. In addition,
during the bone drilling process, this scheme reduces rendering quality and visual fidelity and causes force discontinuity and instability,
which further hinders the quality of force feedback (even if better than using a standard voxel size with marching cubes).

Figure 4 a) shows a simple 15 bits AMR representation of humeral head, 5 bits for each dimension, which consists of 6 levels. Each level
in Figure 4 a) is different color coded. The root level is a single cube that encapsulates entire object. The more detailed grids at lower
levels are always enclosed by the less detailed grids at upper levels. Blue grid in Figure 4 a) is the next level after the root level.

![Figure 4](image)

**Figure 4.** a) A simple 3D AMR hierarchy (left) with 6 levels, b) AMR representation of the humeral head (right).

**Crack free isosurface generation with AMR**

Visualizing drilled bone via isosurface extraction is particularly difficult since AMR uses cell centered data while marching cubes,
which is a de-facto visualization method for isosurfaces, uses values at the vertices of the grid (dual representation). We use Wald’s dual
mesh method [9] stemmed from [7] that extracts crack free isosurfaces from cell-centered AMR. This approach interprets cell (a.k.a.
grids) centers of each patch of the AMR hierarchy as the vertices of a new patch, so that the dual grid to the original patch is generated.
Within these dual grids, isosurfaces are extracted utilizing the standard marching cubes method. The use of dual grids leads to gaps
between different levels of the AMR hierarchy. We use a novel procedural scheme called snapping (see Figure 5) from [9] to fill these
gaps with “stitch” cells (tetrahedra, pyramids, triangle prisms, and deformed cubes) ensuring that this step produces no T-junctions. The
snapping method given in [9] is stemmed from the stitching method given in Weber et al. [10]. Figure 5 shows snapping operation which
results in crack-free isosurface generation. Reader is referred to Wald et al. [9] for detailed explanation of the snapping operation.

![Figure 5](image)

**Figure 5.** A 2D example of snapping various finest-level logical dual cells to actual cells [9]

**Morton Code**

Morton code is a method usually used for block structured AMR (a.k.a octree AMR) representations. Morton code representation provides
an ordering along a space-filling while at the same time preserving the data locality. In our case, we also use block structured AMR
representation with Morton code. This technique enhances data-locality by mapping multi-dimensional data to one dimension. Morton
encoding is used with hash tables in building and generating trees because of its ease of computer implementation to map every voxel’s center. Its ease of implementation is due to its elegant binary nature and use of binary operations. Ensuring that the voxels with closer Morton codes are stored close to each other in the memory which allows a very efficient search. An important parameter of Morton encoding is the number of bits used for the Morton code. This number is computed by determining the quantized coordinates. Nodes are indexed using the x, y, and z coordinates. Three integers representing the coordinates are sent to a function that splits up the bits of each coordinate value by inserting two zero bits between each data bit. Figure 6 presents Morton code and Z-ordering, in that it shows the mapping of the n-dimensional points to one-dimensional scalars.

Figure 6. Visualization of the Morton code-based space filling curve in 2D. Notice that the order pattern represented with blue is called Z-order (blue line) given on the left, which intrinsically represents depth-first traversal (red dashed line) in the tree given on the right.

Efficient AMR based drilling

We have a 3D AMR tree implementation on the application side starting from a single root node. Root node has the coarsest resolution as seen from Figure 7 (left). Drilling process is a process of removing the finest AMR cell from the tree. Contact information is trivial due to the nature of AMR tree representation that can be performed in a logarithmic time. The cell to shave should have the finest AMR level. If the cell that is in contact with the colliding object such as bone shaver does not have the finest resolution, then that cell is divided into 8 more finer cells. If collision continues to happen even at the finer resolution, then the previous step is repeated and the cell is further divided into 8 more finer cells. This process is repeatedly conducted up until it reached to the finest resolution. The finest resolution is determined by the bit resolution per dimension (see Figure 7 for an exemplary representation). For instance, block structured AMR hierarchy is represented in 18 bits corresponds to 6-bit representation for each dimension in 3D. Thus, that will correspond to 3 levels in AMR hierarchy, $\lceil \log_2(6) \rceil$. At the finest level, shaving process proceeds and the contacting cell is removed from the tree. That is illustrated in Figure 7. In this example, there are 5 levels, and the top corner cell is removed from the hierarchy after a contact is detected.

Figure 7. An example AMR tree representation on a simple cube object. (left) Root cell, (right) single cell drilling operation output.

In order to use AMR for our bone shaving simulation in ViRCAST, direct volume visualization with AMR is needed that requires generation of isosurfaces during the mesh deformation processes. This is a time-consuming process. However, ViRCAST requires running realtime dynamic mesh updates. The serial processing nature of the CPU is not well suited to generate isosurfaces during the mesh deformation process. This is a highly parallel task, therefore, to generate isosurfaces with high levels of complexity in realtime, we look into the GPU as a more efficient solution. Speed and efficiency become a very important requirement when dealing with the high resolution models that are composed of many voxels and that undergo multiple mesh manipulations.

Array-based AMR hierarchy

We are using a scalable array to store the AMR cells. An AMR cell is a structure to hold a cell id, a boolean shaving variable, and a shaving scalar with the index of integers for the children cells of the current cell. There are 8 children in a cell in 3D, each cell is located
in the array with its corresponding id as the index. The cell count is exponential by the resolution in use. For a 15-bit resolution there will be up to $2^{15}$ possible cells held in memory. Modern memory allocation methods try to optimize the allocation of each single object in the memory but due to the dynamicity involved in the bone shaving task modern memory access patterns fail to maintain the unmanaged memory allocations. Thus, we present a scalable array-based tree data structure to dynamically manage the cells in the tree as illustrated in Figure 8. We assume that the cell count will be high enough for a possible rearrangement of a bulk of cells that will cause a frame drop. In our case, if a cell has an index, it will stay in that index unless it is deleted.

![Figure 8. An example array-based tree data structure.](image)

The first component of the array is used cells (See Figure 8). Instead of keeping a list of indices for the cells in use, each cell will know locations of its children. To start traversing the tree, there is a root node with the index number 0. If a child has an id other than -1, the traversal algorithm continues on that child recursively until a node that doesn’t have any children reached. The second component of the array only contains two index integers that indicate a bulk of free nodes starting from the first integer to the last. Let’s say the tree has 8 nodes and the first index for free nodes is 9, then there is another cell we want to add to the tree. The integer is incremented and the cell on the 9th index is given to the tree. If the first index is equal to the second integer, an array with the double size is created and the array is copied to a bigger one.

The last component of the array is the free queue. The need for this component is because we can delete some cells to simplify the tree after we shave (delete) some cells. Deleting means adding the integer of the cell into a queue that will be later allocated if needed. The cell allocation method first looks for a possible free cell in this queue, if the queue is empty the second component will be used.

![Figure 9. Neighbor finding algorithm for northern neighbor of node 17 is illustrated.](image)
We use a fast recursive cell neighbor finding algorithm that is used in the AMR refinement process. For simplicity of illustration, in Figure 9, a 2D version of AMR with 4 children at lower level is used instead of the 3D version with 8 children. Our algorithm is a modified version of finding next or previous node in the binary tree search algorithm. In the binary search algorithm, the goal is to find the parent node that could share the starting node and the next node. If the parent node is found, direction of searching area changes from left to right. In our case, we have two different axes instead of one axis as in the binary search. Siblings have their axis aligned neighbors as seen in Figure 9. In this example, finding the northern neighbor of the node 17 is demonstrated (see Figure 9). It starts from the node 17 and pushes its child number into a stack. In this case, 17 does not have any child node (no lower refined level); thus, its stack is empty. As it can be seen from the figure, the process visits parent nodes in order from bottom to top until the root node is found. For instance, parent node pushed \textit{a} into the stack since the previously visited node was 17. Then \textit{b} and \textit{d} are pushed into the stack in order when nodes 4 and 0 are visited respectively. After the root is found, searching process continues in reverse order, namely top to bottom. Apparently, after that point forward the inverse operation of push, pop is applied. For each \textit{a} path is visited, \textit{d} will be popped from the stack; \textit{c} path is visited \textit{b} is popped from the stack, and \textit{d} path is visited \textit{a} is popped from the stack. That way, we keep track of the \textit{Z}-ordering. The popped child number is reversed accordingly (\textit{a}<>\textit{d}, \textit{c}<>\textit{b}) than a recursive call is started from the corresponding node. The algorithm goes to the bottom (the most refined level) until either stack is empty or a node has no child. This neighbor cell search algorithm can be customized for each direction: east, north, west, and south in 2D. Also, for 3D -as the data structure we use-, two directions must be added per each dimension.

### GPU Implementation

The use of GPU with CUDA API allows to rapidly extract isosurfaces and render the results in realtime. The GPU based rendering algorithm takes AMR cell data array as an input and produces a single 3D surface model. Because our AMR implementation is a tree data structure, we first convert AMR tree to an AMR cell array. Each AMR cell stores a Vector3 coordinate for cell’s lower-left corner and an integer for the corresponding level. Each level has double size compared to its one finer level, assuming level 0 is the finest. The GPU based algorithm assumes that the input is AMR data in that each node has at most one level degree difference between its neighbors. Thus, there are predefined limited number of triangulation possibilities due to the limited cases with the neighboring cells as given in the predefined triangle orientations in marching cubes. For speed, these are held in the constant memory of the GPU since constant memory is cached. We adapt the CUDA implementation described in [9] to work in Unity3D\(^1\). A Dynamic-link library is used to send the 3D tree data and the method to get the output vertices and triangles. We sent 3D AMR tree as an array (see Figure 10 (a) for our array representation of AMR hierarchy) to CUDA implementation. Later, we produce vertices and triangles of the 3D object as an array which contributed to the mesh representation (Figure 10 (b)).

![Figure 10. a) Shaved AMR Tree representation of a cube, b) Marching cubes rendered version of the shaved cube.](image)

We start with a 3D tree class as a data structure to store and modify our data. Before creating an instance of it, we give a parameter to define the resolution. We have a shaver object to drill the parts of the tree in 3D space. Because our tree code is axis aligned, finding the node that should be drilled only takes a constant time, in the worst case, the number of iterations is equivalent to the height of the tree. For instance, for a 64-bit resolution in 3D, the height of the tree will be 5. After finding the node that contacts with the shaver, we delete that node, then refine the whole tree based on new modifications. To use CUDA marching cubes, we had to produce an input for the DLL-side of the implementation. It requires a cell array and a scalar array. We traverse the tree and place an instance to the array as these cell and scalar arrays only contain center point of the node and level in the hierarchy. Because our level hierarchy works in reversed order with the DLL side, we need to reverse every level by subtracting maximum level from the current level visited with max\_level-

\(^1\) https://www.unity.com/
level. Then, we send the array to the GPU with our modified interface. This interface will only return vertices and triangles’ array size and will store the data arrays inside the memory.

To get the data from DLL to Unity, we created an array in Unity and allowed the DLL to fill the array with its own data arrays. After this step, we can delete the array at the DLL side without any memory leaks. Unity-side arrays do not have any memory leaks since they are already managed by the Unity engine. With the result array, we separate vertices into multiple arrays because Unity version we use only supports certain number of meshes, up to 65,535 vertices. If the number of vertices of our 3D objects are greater than 65,535, we partition them and create a Unity object for each partition. In case there are some triangles which share two vertices in two different arrays, we populate those triangles in a spare array. When the mesh size goes high, the possibility of a triangle to share two vertices in two different arrays is low, so we don’t need to create another spare array.

![Figure 11](image)

**Figure 11.** Drilling demonstration on a humerus head representation. (a) 3D Humerus head generated by a predefined AMR tree instance, (b) a close-up screenshot before the drilling process, (c) slightly drilling area with bone shaver, (d) drilling continues, (e) a screenshot after the drilling process finishes.

**Results**

We have constructed a scene for a realistic shaving environment with a humerus head model to test our algorithms. Figure 11 demonstrates the step-by-step drilling process in realtime. The resolution we used for the demonstration is 18 bits. The increasing number of bits have an exponential computational power usage. The algorithms are tested on a machine with i7-9750H, GTX-1650, and 16 GB RAM specifications.

Table 1 provides the execution times for the algorithms that we have implemented. As it can be seen in Table 1, the fastest process is the combination of drilling and updating the tree. Tree extraction and mesh generation processes take more time to execute than the drilling phase; thus, those phases dominate the processor’s power consumption. For further research, any optimization in these processes will have a huge impact to the overall performance as presented in Table 2.

**Table 1.** Algorithm performance results (in milli seconds)

<table>
<thead>
<tr>
<th>AMR Tree Resolution</th>
<th>Drilling + Tree Update</th>
<th>Tree Extraction</th>
<th>CUDA Process + Mesh Update</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 bits</td>
<td>0.8</td>
<td>1.36</td>
<td>4</td>
</tr>
<tr>
<td>15 bits</td>
<td>1</td>
<td>6.98</td>
<td>6.39</td>
</tr>
<tr>
<td>18 bits</td>
<td>1.7</td>
<td>29.79</td>
<td>16.3</td>
</tr>
</tbody>
</table>

**Table 2.** Overall frame fates

<table>
<thead>
<tr>
<th>AMR Tree Resolution</th>
<th>Frame Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 bits</td>
<td>162 fps*</td>
</tr>
<tr>
<td>15 bits</td>
<td>70 fps</td>
</tr>
<tr>
<td>18 bits</td>
<td>20 fps</td>
</tr>
</tbody>
</table>

* Frames per second.
Even though developed bone shaving and drilling algorithms are designed specifically for rotator cuff surgery in ViRCAST, these methods can easily be adapted to other virtual surgical procedures. Some of these surgical procedures are common otologic surgical procedures such as mastoidectomy and acoustic neuroma resection [12], temporal bone dissection [13][14][15], petrous bone surgery [16], or dental drilling [17][18] to name a few.

Conclusion
We discussed the widely used isosurface extraction method called marching cubes and presented its limitations in terms of computational speed and memory usage especially in the domain of surgical simulations. Our proposed methods tackle these issues and significantly improve the performance of surgical simulation tasks involving mesh deformation such as bone drilling. The proposed method introduces a GPU-based technique combining Adaptive Mesh Refinement (AMR) and the marching cubes algorithm to create a real-time bone shaving/drilling application as part of our Virtual Rotator Cuff Arthroscopic Skill Trainer (ViRCAST). We have implemented an array-based AMR tree data structure, GPU-based crack-free isosurface generator, and neighbor finding algorithm to obtain a real-time bone drilling application. Resulting performance progress of our method provides a realtime mesh deformation at 18 bits of resolution and a smooth bone drilling task as having at least 70 frames per second at 15 bits or lower resolutions.

Acknowledgement
This project was made possible by NIH/NIAMS R44AR075481-01 and the Arkansas INBRE program, supported by the National Institute of General Medical Sciences (NIGMS), P20 GM103429 from the National Institutes of Health (NIH), and partially supported by NIH/NIAMS R44AR075481 – 01 and the Arkansas INBRE program, supported by the National Institute of General Medical Sciences (NIGMS), P20 GM103429 from the National Institutes of Health (NIH), and partially supported by NIH/NHLBI NIH/NIBIB 1R01EB025241, R56EB026490.

References
EHR Data Quality Assessment Tools and Issue Reporting Workflows for the ‘All of Us’ Research Program Clinical Data Research Network

Noah Engel, BA1, Hongjue Wang, MA2, Xinzhuo Jiang, MS2, Chun Yee Lau, BA2, Jason Patterson, BS2, Nripendra Acharya, BA2, Maura Beaton, MS2, Lina Sulieman, PhD3, Nishanth Pavinkurve, MS2, Karthik Natarajan, PhD2

1The Johns Hopkins University School of Medicine, Baltimore, MD, USA; 2Department of Biomedical Informatics at Columbia University Medical Center, New York, NY, USA; 3Vanderbilt University Department of Biomedical Informatics, Nashville, TN, USA

Abstract

The All of Us (AoU) Research Program aggregates electronic health records (EHR) data from 300,00+ participants spanning 50+ distinct data sites. The diversity and size of AoU’s data network result in multifaceted obstacles to data integration that may undermine the usability of patient EHR. Consequently, the AoU team implemented data quality tools to regularly evaluate and communicate EHR data quality issues at scale. The use of systematic feedback and educational tools ultimately increased site engagement and led to quantitative improvements in EHR quality as measured by program- and externally-defined metrics. These improvements enabled the AoU team to save time on troubleshooting EHR and focus on the development of alternate mechanisms to improve the quality of future EHR submissions. While this framework has proven effective, further efforts to automate and centralize communication channels are needed to deepen the program’s efforts while retaining its scalability.

Introduction

Overview of the All of Us Research Program

Clinical data research networks (CDRNs) enable researchers to access electronic health records (EHR) from thousands of patients and vast territories in an effort to answer questions in translational research. CDRNs have existed for many years, with each containing sets of patients from distinct populations, ages, or locations1.

Datasets used to inform clinical research, however, often lack the racial, ethnic, geographic, and gender diversity needed to advance personalized medical treatment for generalized patient populations2,3. Consequently, the National Institutes of Health launched the All of Us Research Program, which aims to collect health questionnaires, genetic data, physical measurements, digital health information, and EHR from at least 1 million persons in the United States4. The All of Us Research Program is nearly unparalleled in its diversity; as of June 2020, the program had over 224,000 participants in its workbench, with seventy-seven percent of its cohort identified as underrepresented in biomedical research5. With All of Us’s broadly-accessible dataset4, researchers from across the world will have access to robust information that can power the future of precision medicine.

Overview of the Data Network and Transfer

The All of Us Research Program’s data network is expansive; there are over 50 health provider organizations (HPOs) that coordinate the enrollment and EHR transfer of the program’s several hundred thousand participants. Firstly, HPOs must extract information from the local EHR and transform said data into the OMOP Common Data Model (CDM). This OMOP CDM is used to ensure that EHR, regardless of their original format, are put into a standardized set of data structures to enable end-users to index the data in a systematic and scalable fashion6. Finally, HPOs must load the data into a central repository of Google Cloud Buckets. In order to receive funding, HPOs are required to transfer EHR at least once a quarter after they are officially onboarded by the Data Research Center (DRC). The DRC is the receiving end of the central repository and develops a variety of tools, such as data cleaning rules, to ensure EHR are sufficiently de-identified and processed before they are provided to researchers.
There are several protocols to ensure EHR data from HPOs meets basic conformance standards for the OMOP CDM. Outside of this short list of checks, however, sites originally received little to no feedback upon provision of EHR to the repository.

**Data Quality Issues**

The decentralized approach of *All of Us*’s collection process comes with large-scale data quality issues. Similar to other CDRNs, the aggregation of healthcare information from sites with different data storage techniques, distinct EHR systems, and varying resources available to dedicate to healthcare informatics inevitably leads to differences in data quality across sites. Upon investigation, this phenomenon – combined with the aforementioned lack of systematic reviews of HPO submissions – resulted in a number of problems in submissions to the *All of Us* Research Program. These issues included, but not limited to, infrequent data uploads from HPOs, the omission of required fields in patient records, and the failure of sites to translate from native EHR codes to ‘standard’ OMOP codes. These issues are not uncommon to decentralized data environments and have led many within the scientific community to question if EHR data is suitable for research.

While existing CDRNs implemented a number of mechanisms to assess the quality of their EHR data, few have created systems to track the aggregate quality of data across all of their partner sites in a longitudinal fashion or implemented a systematic approach to provide feedback to their contributors. The scale of the *All of Us* program - with its 50 and counting HPO data partners and over 300,000 enrolled participants - provides a unique opportunity to implement a robust approach to decentralized EHR data quality reporting and communication.

In this study, we will detail the means by which the *All of Us* Research Program developed tools to run data quality checks on EHR submissions from enrollment sites. This paper will also detail *All of Us*’s communication practices to illustrate how findings were disseminated in a scalable manner and ultimately increased data partner engagement. We will then show that these protocols enabled the program to improve EHR data quality from HPOs as measured by metrics developed by the DRC and existing third-party informatics tools. Finally, we will detail means by which *All of Us* can incorporate additional tools and practices to further strengthen its data quality assessments.

**Methods**

**Creating Data Quality Metrics**

In order to assess data quality issues, the *All of Us* Research Program investigated the three-category framework described by Data Quality Harmonization Framework (DQHF) in 2016. DQHF’s first category is conformance, which refers to adherence to technical and syntactic constraints. Completeness is the second category, which validates that expected data values are present as one may expect within an observational context. The final tenet is plausibility, which ensures that data values provided are consistent with real-world applications of clinical medicine. DQHF was chosen as a means to establish data quality metrics and communicate said issues with sites given the expertise of its developers and its application in a variety of other informatics contexts.

With this framework, the DRC created a variety of metrics and goals to quantify the conformance, completeness, and plausibility of the EHR at both the site- and program levels. The inspiration for these assessments came from a variety of sources, all of which were guided by real-world EHR research applications. Firstly, *All of Us* program leadership explicitly specified minimum requirements regarding EHR data transfer and quality for HPOs to maintain their grants. Leadership chose these metrics as experts deemed them as foundational pillars to ensure that the EHR dataset would be suitable for medical research. Consequently, all of the program-defined metrics were included in the analyses run by the EHR Operations team. The second source of data quality metrics came from data interaction. The analyst responsible for developing the metrics previously took part in several projects in the *All of Us* research program and knew of the obstacles faced by the data’s end-users. The analyst also continued data exploration on both intra- and inter-site levels to identify issues that may arise when researchers engage with EHR. These experiences enabled the creation of metrics that could guide sites on how to address issues that obstruct the research usability of the EHR. Finally, investigations of the cleaning rules implemented by the DRC allowed the EHR Operations team to identify points of extensive data elimination as EHR moved through the pipeline. These evaluations allowed us to identify what types of data errors could be resolved prior to submission and thereby
improve the ultimate completeness and usability of the *All of Us* dataset. All of the data quality metrics were approved by program personnel before implementation and dissemination to sites. If needed, the EHR Operations team consulted physicians to determine the relevance and expected frequency of specific clinically-based metrics.

In total, these SQL queries were intended to evaluate 19 distinct metrics relating to the conformance, completeness, and plausibility of the EHR submissions provided by sites. Most of the metrics were applied across multiple tables - with some metrics containing an additional weighted average across six of the OMOP tables (condition_occurrence, observation, procedure_occurrence, measurement, drug_exposure, and visit_occurrence) - raising the total number of metrics to 94. All of the metrics discussed can be found in the figure below (Figure 1).

<table>
<thead>
<tr>
<th>Metric Name</th>
<th>Description</th>
<th>Data Quality Dimension</th>
<th>Applicable OMOP Tables</th>
<th>Number of Resultant Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visit Occurrence ID Failure Rate</td>
<td>Percentage of visit_occurrence_id fields in the selected table that do not exist in the visit_occurrence table</td>
<td>Conformance</td>
<td>1, 2</td>
<td>2</td>
</tr>
<tr>
<td>Route Concept ID Failure Rate</td>
<td>Percentage of route_concept_id fields that is not populated with a standard concept with domain route</td>
<td>Conformance</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Unit Concept ID Failure Rate</td>
<td>Percentage of unit_concept_id fields that is not populated with a standard concept with domain unit</td>
<td>Conformance</td>
<td>3, 4</td>
<td>2</td>
</tr>
<tr>
<td>Duplicate Records</td>
<td>Number of rows where all fields are identical to another row in the same table (with the exception of the primary key, which is forced to be distinct for each row)</td>
<td>Plausibility</td>
<td>1, 2, 3, 4, 5, 6, 7</td>
<td>7</td>
</tr>
<tr>
<td>Physical Measurements</td>
<td>Shows the percentage of participants in the visit_occurrence table with the following in the measurement table: 1. Body Weight 2. Body Weight 3. BMI 4. Heart Rate</td>
<td>Completeness</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>COVID Measurements</td>
<td>Shows the number of records where the 'concept ancestor' indicates that the record should contain information relating to COVID-19: arise 1. Total COVID measurements 2. The number of ancestor_concept_ids that were not properly mapped to a standard COVID-19 concept 3. Proportion of ancestor_concept_ids that were not properly mapped to a standard COVID-19 concept (by) with respect to the total number of COVID measurements</td>
<td>Conformance</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>EHR Transfer Rate A</td>
<td>Percentage of patients who are eligible for EHR transfer who have relevant EHR</td>
<td>Completeness</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>EHR Transfer Rate B</td>
<td>Percentage of patients who have had an in-patient visit (as indicated by a physical measurement or biotimespecimen) who have relevant EHR</td>
<td>Completeness</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>Concept ID Success Rate</td>
<td>Percentage of rows where the concept_id is a standard concept and matches the domain of the table</td>
<td>Conformance</td>
<td>1, 2, 3, 4, 5, 6, 7</td>
<td>7</td>
</tr>
<tr>
<td>Non-zero Concept ID</td>
<td>Percentage of rows where the concept_id does not equal zero</td>
<td>Conformance</td>
<td>1, 2, 3, 4, 5, 6, 7</td>
<td>7</td>
</tr>
<tr>
<td>Non-zero Person ID</td>
<td>Percentage of rows where the person_id is neither null nor zero</td>
<td>Conformance</td>
<td>1, 2, 3, 4, 5, 6, 7</td>
<td>7</td>
</tr>
<tr>
<td>Non-zero Type Concept</td>
<td>Percentage of rows where the type_concept does not equal zero</td>
<td>Conformance</td>
<td>1, 2, 3, 4, 5, 6, 7</td>
<td>7</td>
</tr>
<tr>
<td>Non-zero Source Concept ID</td>
<td>Percentage of rows where the source_concept_id does not equal zero</td>
<td>Conformance</td>
<td>1, 2, 3, 4, 5, 6, 7</td>
<td>7</td>
</tr>
<tr>
<td>Non-zero Status Concept ID</td>
<td>Percentage of rows where the status_concept_id is neither null nor zero</td>
<td>Conformance</td>
<td>1, 2, 3, 4, 5, 6, 7</td>
<td>7</td>
</tr>
<tr>
<td>Non-zero Concept ID</td>
<td>Percentage of rows where the concept_id is neither null nor zero</td>
<td>Conformance</td>
<td>1, 2, 3, 4, 5, 6, 7</td>
<td>7</td>
</tr>
<tr>
<td>End Date Preceding Start Date</td>
<td>Percentage of rows where the 'end date' precedes the 'start date'</td>
<td>Plausibility</td>
<td>1, 3, 5</td>
<td>3</td>
</tr>
<tr>
<td>Data after Death</td>
<td>Percentage of rows where the date in question follows the 'death date' found in the 'person table' (if the person is deceased)</td>
<td>Plausibility</td>
<td>1, 2, 3, 4, 5, 6, 7</td>
<td>7</td>
</tr>
<tr>
<td>Erroneous Dates</td>
<td>Percentage of rows where the date or datetime field is 1. Before 1900 (for all dates other than observation) 2. Before 1980 (for observation)</td>
<td>Plausibility</td>
<td>1, 2, 3, 4, 5, 6, 7</td>
<td>7</td>
</tr>
<tr>
<td>Date/Datetime Discrepancy</td>
<td>Percentage of rows where the 'date' and 'datetime' fields are more than one day apart</td>
<td>Conformance</td>
<td>1, 2, 3, 4, 5, 6, 7</td>
<td>7</td>
</tr>
</tbody>
</table>

Figure 1. Different data quality metrics with their descriptions, relevant data quality dimension, and applicable OMOP table(s). Table key: 1 = condition_occurrence, 2 = procedure_occurrence, 3 = drug_exposure, 4 = measurement, 5 = visit_occurrence, 6 = observation, 7 = tables 1-6 (aggregate metric).
Organization, Calculation, and Presentation of Metrics

After defining data quality metrics, we retrieved data quality attributes on the OMOP-transformed tables provided by sites by using queries constructed on Google BigQuery. This query method construction enabled fast and scalable searches while maintaining the general syntax of Structured Query Language (SQL). These metrics were calculated for each site and the results would be output to a Microsoft Excel file.

After creation of the Excel files, another Python script was developed to reorganize the data and create aggregate data quality metrics. This program provided multiple views of seeing the data quality over time. For each metric, the program created the following:

1. One Excel sheet for every OMOP table. In said sheet, each site occupied a row. The bottom (final) row was the ‘aggregate’ data quality metric, which was an average across all sites with each site being weighted by its number of rows. The columns of said sheet represented different dates when the data quality metric was calculated. This sheet was created for each of the OMOP tables for which the data quality metric was calculated.

2. One Excel sheet for every data partner. In said sheet, each OMOP table occupied a row. The bottom (final) row was the ‘aggregate’ data quality metric, which was an average across all OMOP tables with each table being weighted by its relative number of rows. The columns of said sheet represented different dates when the data quality metric was calculated. This sheet was created for each of the HPOs for which the data quality metric was calculated. One additional Excel sheet was created to create an ‘aggregate’ data quality metric across all of the data partners. The format of said additional sheet, including the meanings of the rows as OMOP tables and columns as dates, was otherwise identical.

After the creation of these sheets, the information generated was fed into a Jupyter notebook, in which the Python-based Seaborn package created heat maps for said data quality metrics. These heat maps enabled visualization for sites to understand the progression of their data quality over time and the precise areas where they could resolve errors in their submissions.

Finally, another script generated messages to send to data stewards via email. The emails sent to the sites communicated which data quality metrics were not passing the DRC’s standards and provided direct links to resources – such as video tutorials, sample queries, and text explanations - where sites could learn about the exact meaning of these data quality metrics. These emails would contain the aforementioned heatmaps so sites could precisely locate the source of their data quality issues and see how their data quality trended over time. These emails and the corresponding heatmaps would be sent to sites on the first Monday of every month to ensure that sites were routinely notified about their data quality and held accountable for their submissions.

All of the aforementioned scripts are publicly available online within the All of Us GitHub at the following link: https://github.com/noahgengel/curation/tree/develop/data_steward/analytics.

Automating Data Quality Reports

In an effort to improve the scalability and timeliness of data quality reports, the DRC implemented an interactive data visualization tool from Tableau Software®. Starting in June 2021, the Tableau software automatically ran the previously-defined queries on Google BigQuery every day and calculated the corresponding data quality metrics. Once these metrics were calculated, the program created a series of data visualizations – including bar graphs, stacked bar graphs, line graphs, and heatmaps – so data stewards and program leadership could log into their Tableau accounts and receive feedback within 24 hours of each submission. This framework also enabled the DRC to integrate new data quality metrics into the dashboard after testing a metric on several volunteer sites.

The visualization tools were implemented with versatility, readability, and context in mind. These visualizations were separated into five dashboard tabs and contained interactive filters by which sites could select the metrics they wanted to prioritize and the exact calculations of each metric. Similar to past iterations of the data quality reports, Tableau provided the option to trend a data partner’s metrics in a longitudinal fashion so data stewards could assess their sites’ progression. Finally, unlike past reports, the data quality dashboard displayed the arithmetic mean of each
regional consortium’s data partners for each data quality metric. This display (Figure 2) allows data stewards to compare their site’s progress to the status of other sites and see how each consortium compares to the DRC’s targets.

Figure 2. Example screenshot (sites de-identified) from the interactive data visualization tool. Screenshot also contains a pop-up that data stewards could bring up to better understand the data quality metric at hand.

Additionally, in an effort to validate EHR submissions against proven informatics tools, the DRC leveraged ACHILLES. ACHILLES is a free-to-use and open-source program that characterizes and visualizes data quality rules regarding patient demographics, clinical observations, conditions, procedures, and drugs on OMOP-conformant observational health databases. By leveraging ACHILLES, the DRC could analyze and tabulate a robust set of potential errors in each EHR submission for each site. ACHILLES served as an appropriate means to externally validate data quality as many of its checks were completely distinct from the DRC’s reported data quality checks. Therefore, reductions in the absolute number of ACHILLES errors over time could signify that improvements in EHR data quality were not confined to specific data quality errors targeted by the DRC.

After each submission, the All of Us pipeline would generate a results.html output file within the next 24 hours. The results.html file would appear in the bucket where sites submitted their data. This file contained information about the most recent upload, including a report on the errors, notifications, and warnings generated by ACHILLES. Sites could then view the output on these files and adjust their subsequent submissions to resolve or reduce the number of errors in the results.html reports.

Mass Communication and Resource Creation

Beyond reporting on data quality metrics, the DRC created means through which it could disseminate data quality information at scale. Firstly, the DRC hosted bi-weekly calls in which data stewards from all sites would join and learn about newly-implemented data quality checks, initiative-wide metrics, and issues that may pertain to their specific EMR systems. Next, in 2019 Q3, the DRC set up regular video conferences in which we would discuss data quality issues with individuals who oversaw EHR transfer from multiple sites within the same regional consortium. The DRC also created a series of data steward email threads based on commonalities between said data stewards—such as EMR systems—to ensure that we could efficiently send out information to pertinent parties. Finally, in 2021 Q2, the DRC leveraged ZenDesk – a customer service management system – to ensure data partners’ questions would be effectively stored, accessible, and organized when compared to traditional email.

The DRC also wanted sites to independently understand and potentially resolve many of their data quality issues. The DRC uploaded all of the code that calculated the data quality metrics—with the exception of any potentially-identifying information—to a public GitHub repository. Pending the technical abilities of each data steward, this transparency allowed partners to recapitulate the DRC’s metrics and identify example rows of EHR data that
negatively impacted their data quality. In an effort to improve user experience, said GitHub links – along with definitions of each data quality metric and a means by which sites could provide feedback to the DRC– were incorporated into the Tableau dashboard (Figure 2). To facilitate understanding of the metrics, the DRC created several iterations of knowledge repositories. These repositories allowed users to easily identify a particular metric and find a page to help them recapitulate the DRC’s results. Depending on the iteration of the repository, each page would contain information such as sample SQL queries, text descriptions of the metrics, EMR system-specific information, common points of failure, or an instructional video recorded by the DRC.

Results

Impact on Site Engagement

Prior to implementation of regular data quality communications, sites frequently failed to comply with quarterly data transfer requirements. For instance, in the six quarters prior to the systematic communications, there were two quarters where over 20% of the sites did not transfer any EHR to the DRC (Figure 3).

After implementation of regular communications regarding data quality in Q3 2019, sites were more likely to make at least one submission per quarter. In the first half of 2019, about 81% of sites were compliant with EHR transfer. The following half-years, however, saw average compliance rates of 78%, 90%, 96%, and 96%, respectively. Of note, once the data quality communications were largely established in Q2 2020, the compliance rate for a given quarter never fell below 93% - a rate higher than any of its preceding quarters (Figure 3).

During this same timeframe, the DRC expanded the number of participating HPOs. Between 2020 Q2 and 2021 Q2, the DRC started accepting submissions from 14 new sites, representing a 39% expansion of the program when compared to 2019 Q2 (Figure 3). The sites added during this time frame represented a diverse array of data partners – ranging from federally qualified health centers (FQHCs) to large academic medical centers.

![Figure 3. Number of sites (both total and those with a successful submission) for each quarter.](image)

Improvements with Respect to Internal Data Quality Standards

Given that conformance serves as the foundation for the DQHF, many of the DRC’s high-priority metrics target the adherence of submissions to the OMOP Common Data Model (CDM). One set of metrics evaluated the ‘concept_id’ field for each of six the widely-used OMOP tables previously mentioned to ensure the ‘domain_id’ of the concept_id matched the destination table and that the ‘concept_id’ was a ‘standard’ non-zero value. This metric serves as a proxy to determine whether data partners successfully and completely converted information stored in...
their native EHR systems to the OMOP CDM. This metric intentionally excludes the first quarter of submissions for each site in order to allow data partners to become familiar with OMOP specifications and data transfer protocols.

In 2019 Q1, the median concept success rate for each site across the aforementioned tables was below the DRC’s target concept success rate of 90%. The spread of the median concept success rates also proved large, with a standard deviation of 32.9% and a 25th-75th percentile range of 20.8%. Upon implementation of the data quality communications in 2019 Q3, the median concept success rate jumped up to 98.5%, with a standard deviation of 19.5% and a 25th-75th percentile range of 9.7%. This trend of increasing OMOP adherence and decreasing site variability in data quality persisted for all subsequent quarters, which all had median values over 99.2%, and high standard deviation and interquartile ranges of 27.4% and 4.07, respectively. The most recent calculations, which reflect 2021 Q3, show an initiative-wide median concept success rate for the six tables of 99.98% with a standard deviation of 13.8% and a 25th-75th percentile range of 1.27% (Figure 4). This trend of improvement in data quality emerged despite the fact that the number of rows nearly tripled from over 1.1 billion rows in 2019 Q1 to nearly 3.3 billion rows in 2021 Q3.

![Median Concept Success Rate Across Six OMOP Tables After First Quarterly Submission](image)

**Figure 4.** Median conformance rate for designated OMOP tables across all sites for each quarter. No data available for 2020 Q1 or 2020 Q2.

**Improvements with Respect to External Data Quality Standards**

The ACHILLES analyses similarly showed an improvement in data quality across a large number of sites. During the first implementation of the ACHILLES checks in 2019 Q1, of the 30 sites with at least one submission at that time, 67% of sites had over three ACHILLES errors in their final submission of the quarter. While this percentage dropped to 41% of sites in 2019 Q2, it continued to fall and eventually bottomed out to around its most recent value of 35% of sites with over three ACHILLES errors as of the first half of 2021. This relative stabilization in terms of the number of sites with few ACHILLES errors came despite the continual addition of new sites, indicating an increase in the total number of sites with fewer than three ACHILLES errors (Figure 5).

The only exception to the aforementioned flatline was in 2020 Q1, in which 61% of sites reported three or more ACHILLES errors (Figure 5).
Figure 5. Number of sites (both total and those with more than three ACHILLES errors) for each quarter.

Discussion

Ultimately, the implementation of data quality metrics and regular communications from the DRC to its data partners increased HPO engagement as measured by compliance with data transfer requirements. The implementation of regular communications from the DRC, regardless of each site’s submission status, reversed the trend of many sites not meeting their requirements for quarterly data transfer (Figure 3). Of note, this increase in site submissions arose despite the fact that the DRC took no efforts to further encourage submissions from HPOs. These findings indicate that merely implementing automated reports provided our data stewards with incentive to invest in dedicating their time and resources to the All of Us program. Consequently, the implementation of these protocols should allow the initiative to continue expanding its number of partner sites without experiencing an overwhelming number of non-compliant data partners who would require manual interventions from the DRC and leadership.

The results from this study also indicate that these practices hold potential to improve the quality of EHR data used for research purposes. As indicated by checks on ‘concept success rates’, consistent reporting on a particular issue resulted in a substantial decrease in the number of sites with suboptimal data quality and led to decreased variability in data quality across sites. While these results appear intuitive, they show that initiative-led communications can correct potential data errors without manual interventions on behalf of the DRC.

Equally promising are the results from the ACHILLES analyses. While some overlap existed between the metrics reported by the DRC and the messages displayed by ACHILLES, the ACHILLES tool contained far more checks. It should also be of note that the DRC’s communications with HPOs mostly focused on the initiative-defined metrics and that the ACHILLES reports were oftentimes considered of low-priority to sites, with many data stewards reporting that they seldom or never accessed the reports due to the long processing time of the results.html files. The results from our study, however, indicate that the implementation of data quality checks and communications reduced the relative frequency of sites with more than three ACHILLES errors and an absolute increase in the number of sites with three or fewer ACHILLES errors (Figure 5). The only exception to the trends previously discussed arose in 2020 Q1. This exception likely occurred, however, because data partners were instructed to prioritize increasing their frequency of EHR transfer over resolving potential data quality issues in March 2020 as program leadership wanted recent data in the midst of the COVID-19 pandemic. As a whole, these findings indicate that the process of engaging sites and creating data quality metrics to encourage sites to interrogate their EHR had off-target improvements in data quality as measured by a third-party informatics tool.
The outlined EHR data quality workflow has opportunities for expansion and improvement. As previously mentioned, the DRC specifically targeted issues relating to conformance as this dimension of data quality forms the foundation of the DQHF.\textsuperscript{11} Future iterations of the workflow, however, should expand the number of metrics calculated to increase the means by which we assess conformance and expand our emphases to the additional dimensions of completeness, plausibility, and recency of the data. Furthermore, the DRC could benefit by working with \textit{All of Us} researchers to understand the issues faced by users and how it could prioritize existing metrics for our data partners. Finally, the DRC’s checks only work on OMOP-conformant EHR data – limiting the ability of non-OMOP conformant sites to investigate their data quality or the ability of other CDRNs to replicate our protocols.

The DRC could also expand the robustness of its data quality metrics. For instance, future metrics could create notifications if a site’s values for a particular measurement or procedure – such as patient heights or frequency of the flu shot – fall well below the values found in other HPOs. These analyses are already conducted by the DRC but could benefit from automation in order to improve the communication and frequency of these findings. Analyses against external datasets from other institutions could similarly assess the readiness of EHR data for research purposes. The DRC could also leverage data quality tools other than ACHILLES, such as DQe-c\textsuperscript{8}, in order to assess the quality of site submissions against a battery of analyses known to work with OMOP-conformant clinical data. Finally, the DRC hopes to integrate other information about the processing of the data, such as the amount of data eliminated by its data cleaning rules, into its visualization tools. This information would ensure data stewards can appreciate the importance of data quality and allow leadership to assess if certain sites are disproportionately affected by cleaning.

The DRC could also borrow practices established by other CDRNs in order to improve its data quality workflow. Similar studies implemented a diversity of ways in which they analyzed and quantified data quality interactions – such as characterizing the cause of the issue, quantifying the number of interactions with different sites, and creating an overall quality ‘score’ for each submission to reflect the relative prioritization of different data quality errors\textsuperscript{15}. Past studies could also provide the DRC with insights as to potential future data quality metrics and the relative frequency with which particular issues may arise\textsuperscript{9}.

In total, the current workflow for \textit{All of Us} has and will continue to undergo many iterations. Future analyses of the effectiveness of the DRC’s workflows should expand upon and strengthen the findings of this study. These additional findings would have the potential to further elucidate the benefit of creating a regular, automated, and approachable means of reporting on data quality within large CDRNs.

\textbf{Conclusion}

Large decentralized data networks inevitably face challenges in their ability to reconcile data from disparate sources and ensure resultant EHR is of sufficient quality for research. \textit{All of Us} provides a strong case study as to how networks could implement a systematized approach to reporting data quality at a nearly unprecedented scale. Thus far, these protocols have proven effective at increasing the engagement of partner sites and decreasing data quality issues as measured by internal and external assessments. Going forward, the \textit{All of Us} program could expand on its current efforts in order to further quantify the effectiveness of its data reporting processes and improve the usability of its EHR for use in research.
References


Assessment of Data Quality Variability across Two EHR Systems through a Case Study of Post-Surgical Complications

Sunyang Fu, M.H.I\textsuperscript{1}, Andrew Wen, M.S.\textsuperscript{1}, Gavin M. Schaeferle B.S.\textsuperscript{3}, Patrick M. Wilson M.P.H\textsuperscript{3}, Gabriel Demuth, Ph.D.\textsuperscript{2}, Xiaoyang Ruan, Ph.D.\textsuperscript{1}, Sijia Liu, Ph.D.\textsuperscript{1}, Curtis Storlie, Ph.D.\textsuperscript{2,3}, Hongfang Liu, Ph.D.\textsuperscript{1}

\textsuperscript{1}Department of AI and Informatics Research, \textsuperscript{2}Department of Quantitative Health Sciences, \textsuperscript{3}Kern Center for the Science of Healthcare Delivery, Mayo Clinic, Rochester, MN, USA.

Abstract

Translation of predictive modeling algorithms into routine clinical care workflows faces challenges in the form of varying data quality-related issues caused by the heterogeneity of electronic health record (EHR) systems. To better understand these issues, we retrospectively assessed and compared the variability of data produced from two different EHR systems. We considered three dimensions of data quality in the context of EHR-based predictive modeling for three distinct translational stages: model development (data completeness), model deployment (data variability), and model implementation (data timeliness). The case study was conducted based on predicting post-surgical complications using both structured and unstructured data. Our study discovered a consistent level of data completeness, a high syntactic, and moderate-high semantic variability across two EHR systems, for which the quality of data is context-specific and closely related to the documentation workflow and the functionality of individual EHR systems.

Introduction

The rapid adoption of electronic health record (EHR) systems incentivized by the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 has enabled the digital transformation of clinical data and large-scale data-driven research\textsuperscript{1,2}. In particular, the longitudinal, voluminous, and dense data offered by the EHR fuels the development of modern machine learning (ML) techniques for predicting disease trajectories and health outcomes, offering unique opportunities for real-time clinical decision support, risk management, and personalized patient monitoring\textsuperscript{3-4}. In alignment with the vision of evidence-based care and precision medicine, medical decisions can be tailored to the individual patient leveraging predictive models trained from longitudinal EHR data\textsuperscript{3}. One famous example was the Dual Antiplatelet Therapy (DAPT) study that used multiple predictive models to estimate the risk of ischemic events and bleeding to identify unique clinical factors, providing data-driven knowledge insights for maximizing patient treatments\textsuperscript{5}. More recently, Dikilitas et al leveraged both EHR and eMERGE data to derive risk prediction models for coronary heart disease based on three major racial and ethnic cohorts\textsuperscript{7}. Leveraging the power of unstructured data, Oliwa et al developed a predictive model from EHR-based clinical notes to identify patients who are at risk for falling out of HIV care\textsuperscript{8}.

Despite the increasing volume of research related to clinical predictive modeling, the translation of prediction algorithms into routine clinical care remains challenging\textsuperscript{6,9-10}. A recent study done by Wong et al published in JAMA Internal Medicine evaluated the Epic sepsis model on a large-scale cohort of 27,697 patients. The observed model performance (AUC, 0.63) was substantially lower than the reported performance in the internal documentation (AUC, 0.76-0.83)\textsuperscript{11}. Further analysis revealed that primary issue for the Epic sepsis model was not the degradation of performance, but rather the direct deployment of the model without a proper definition of EHR data elements, implementation transparency, detailed instructions on how model can be used, customized, and interpreted, as well as best practices to deal with data quality issues in the context of the clinical problem that the model is designed for.

These findings support the intuitive idea that models cannot be trusted without a good understanding of the data being fed into them. Consequently, the validity and portability of predictive models are dependent on the data on which it is derived\textsuperscript{10}. EHR data is known to suffer from several data quality issues\textsuperscript{12,13}. Unlike data being prospectively collected in a controlled environment such as clinical trials, EHR systems are primarily designed for
patient care, and data documentation patterns can be easily affected by numerous contextual factors, including clinical setting (e.g., ICU vs. non-ICU), human factors (e.g., varying levels of medical expertise and training), patient characteristics, and practice guidelines (e.g., whether to document incidental findings)\textsuperscript{14-16}. Furthermore, the EHR system itself has a significant impact on the form and format of clinical data. Built-in documentation functionality such as templates, copy and paste, auto-documentation, and transcription can affect the EHR-specific syntactic and semantic definition for any data contained therein\textsuperscript{17, 18}. The issues with data quality can therefore be further exacerbated when working with different EHR systems. An evaluation study done by Madigan et al. discovered that 40% of results from ten different clinical databases vary significantly in terms of data heterogeneity, which measures the variability of information quality and semantic definition across heterogeneous data sources\textsuperscript{14}. In the context of predictive modeling, if models are trained on data that cannot be reproduced due to a high level of variability, models may suffer the issues of portability and generalizability.

To help further investigate and quantify the data quality issue caused by the heterogeneous EHR systems in the context of predictive modeling, we retrospectively assessed the variability of data generated from two different EHR systems. As EHR system functionality and information documentation patterns are deeply embedded within the clinical workflow and practice, the quality of data needs to be examined for the given context (e.g., clinical setting, data documentation environment, and use cases). We considered three dimensions of data quality in the context of EHR-based predictive modeling for three distinct translational phases: model development (data completeness), model deployment (data variability), and model implementation (data timeliness). The data quality-related measurements were implemented in a real-world study of predicting post-surgical complications (PSC) that comprised a wide range of clinical modalities collected from three stages of surgery (pre-operative, intra-operative, and post-operative). To the best of our knowledge, this is the first study that compares data heterogeneity of two EHR systems using the case matching design. We believe the pragmatic informatics methods presented by the study can be considered as potential data quality assessment methods for the implementation and translation of future predictive models.

**Methods and Materials**

**Study Setting** This study was approved by the Mayo Clinic Institutional Review Board. In May 2018, Mayo Clinic completed a large EHR migration and workflow standardization. The effort for Mayo Clinic Rochester campus included the conversion of the GE Centricity/LastWord EHR system (Centricity) to Epic EHR system (Epic). This migration offers an ideal scenario to study the difference between two EHR systems because confounding factors from inter-institutional variation can be mitigated due to the entire study being conducted within a single institution\textsuperscript{14}. In addition, we used the case matching design to account for potential confounders contributed by patient population variation\textsuperscript{19}. Bins were created for the age variable with a fixed range of 5 years. We performed exact matching for age, sex, and type of surgery (Table 1). Two study cohorts with a colorectal surgery as the primary procedure performed at Mayo Clinic Rochester were retrospectively constructed. Each cohort contains a total of 811 patients.

<table>
<thead>
<tr>
<th>Table 1. Matching Criteria of Two EHRs</th>
<th>GE Centricity (Pre-migration)</th>
<th>Epic (Post-migration)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total matched patients</td>
<td>811</td>
<td>811</td>
</tr>
<tr>
<td>Matching criteria</td>
<td>Age, sex, type of surgery (CPT: 44140, 44141, 44143-44147, 44150 - 44153, 44155 - 44158, 44160, 44204 – 44208, 44210 - 44212, 45110 - 45114, 45116, 45119, 45123,45395, 45397)\textsuperscript{20}</td>
<td></td>
</tr>
</tbody>
</table>

**Study Variables** All study anticipates are part of the ACS National Surgical Quality Improvement Program (ACS NSQIP\textsuperscript{5}) based on the Mayo Clinic Rochester campus. The program conducts a monthly evaluation of a sample (approximately 20%) of the colon and rectal surgery (CRS) practice based on standard procedure sampling methodology\textsuperscript{21}. The NSQIP variables were defined by the ACS NSQIP abstraction guidelines and can be summarized into three stages: pre-operative, intra-operative, and post-operative (including post-hospitalization) (Figure1). The key variables include patient demographic, comorbidities, preoperative labs (90 days before surgery), clinical, intraoperative elements, and postoperative occurrences/complications for 30 days after surgery.
McNemar’s test was performed to determine the statistically significant difference in the data completeness between Centricity and Epic.

**Data collection** Definitions of data collection and abstraction were standardized and aggregated with 18 other participating institutions across the nation\(^2\). The structured data consists of 102 unique variables falling under categories of demographic data, patient-provided information (PPI), symptoms, comorbidities, physiologic measurements, laboratory tests, observational assessments, and operative factors. The data was retrieved from the Mayo Unified Data Platform (UDP) using an R-based application programming interface (API). The UDP is an enterprise data warehouse that loads data directly from the Mayo Clinic EHR. Patient comorbidities were found from ICD-9 and 10 codes recorded within one year of surgery. PPI was measured and collected at the time of admission. Symptoms, physiologic values, laboratory tests, and observational factors were abstracted from two weeks before surgery until the start of surgery. In addition to the 102 variables abstracted from EHR, another 16 were generated using NLP as a service\(^2\), a Mayo Clinic internal natural language processing (NLP) platform for extracting medical information from unstructured text. This system was developed based on an open-source NLP framework MedTaggerIE\(^2\). In total, 118 variables were created for the final data set.

**Measurements of EHR Variability** To examine the potential variability of data quality caused by two EHR systems, we consider three dimensions of data quality: data completeness, data variability (syntactic and semantic), and data timeliness, as listed in Table 2.

**Table 2. Definitions of Data Quality Dimensions in the Context of EHRs**

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data completeness</td>
<td>A record contains all observations, all desired types of data, and a specified frequency of data over time(^2),(^2).</td>
</tr>
<tr>
<td>Data variability (syntactic)</td>
<td>The structure (or syntax) of data(^1),(^2),(^7).</td>
</tr>
<tr>
<td>Data variability (semantic)</td>
<td>The meaning (or semantics) of data(^1),(^2),(^7).</td>
</tr>
<tr>
<td>Data timeliness</td>
<td>The measurement of time expectation for accessibility and availability of data(^2).</td>
</tr>
</tbody>
</table>
**Data Variability** The health level seven (HL7) messages of unstructured clinical notes within one month before and after the surgery date were retrieved for the two matched cohorts. The HL7 Clinical Document Architecture (CDA) is a standard XML format for the syntactic representation of clinical documents based on the Reference Information Model (RIM)\(^5\). The general structure of a CDA document is comprised of 1) document header or metadata information such as document date, document creator, and service location, and 2) narrative text in the body of the document. Based on the definition proposed by Elkin et al and Sohn et al, syntactic variability was examined by comparing meta-structure, documentation sections of the HL7 messages, and calculating corpus statistics\(^15, 27\). The following metrics were considered: tokens/section, tokens/document, tokens/patient, sections/document, sections/patient, and documents/patient\(^5\). The statistically significant difference between the two sites was determined using Wilcoxon signed-rank test.

The semantics variability was examined by comparing the number of PSC concepts per patient across two EHR systems. The PSC concepts were extracted by an existing NLP algorithm\(^30\). Since the original algorithm was developed and evaluated using the Centricity data only, we conducted corpus annotation and NLP refinement on 100 patients with roughly 1200 Epic clinical notes (within one month before and after surgery index date). Corpus annotation is a process of marking interpretative linguistic and predefined clinical concepts. The annotation was conducted by the same annotator (DMI) who participated in the previous study and had gone through training and consensus development. The same annotation guideline, annotation software (MAE), and schema were applied. The 100 patients were randomly split into 50 training and 50 test sets. The out-of-box (i.e., directly applied with no refinement) precision, recall, and f1-score for NLP were 0.72, 0.84, and 0.79, respectively. After the refinement on the training data, the final performance on test data was 0.92 in f1-score. Two versions of the NLP algorithm were applied separately to two cohorts (Centricity and Epic).

Furthermore, we assessed the variability of semantic textual similarity (STS) of the positive mentions extracted from the NLP algorithm. We measured the sentence pair textual similarity using the averaged value of three surface lexical similarities, which include a string-matching algorithm proposed by Ratcliff and Obershexp, cosine similarity of two-word vector space, and Levenshtein distance\(^31-33\). The method was utilized and evaluated in the 2018 BioCreative/OHNLP clinical semantic textual similarity challenge\(^32\). A high similarity sentence pair is determined when the average score was greater or equal to 0.40. Based on the concept distribution, we examine the textual similarity of two frequent concepts - Anemia and Abcess and the two least frequent concepts – Purulent Drain and Wound Infection (with minimal 50 sentence pairs per section). We calculate both intra-EHR similarity (i.e., comparison within the same EHR system) and inter EHR similarity between Epic and Centricity. The distributions of the unique clinical expressions were visualized using histogram charts.

**Data Timeliness** In the era of achieving real-time clinical decision support and prospective risk detection, information timeliness becomes an important quality criterion since the timeliness of the model is dependent on data. Data timeliness was defined as the time expectation of whether information can be accessible given each patient encounter\(^28\). The analysis of data timeliness for structured data was focused on lab variables due to their high prevalence and importance to the prediction of PSC\(^24\). We retrieved the lab result record date (i.e., the date when a record loaded to the source system) and compared it with the patient encounter date. For unstructured data, we retrospectively collected and measured the time spent on the documentation of clinical notes for each patient visit (i.e., comparison of note date on source system and encounter date). To simplify the measure, we define timely information as the data that can be accessed within 24 hours of a CRS-related clinical encounter.

**Results**

**Data completeness** The overall comparison of RoM across the two EHRs was illustrated in Figure 2. Among a total of 118 variables studied, the median rate of missing for Centricity is 0.011 (1st IQR 0, 3rd 0.71), whereas it was 0.007 (1st IQR 0, 3rd IQR 0.665) for Epic. We observed a high RoM among the intraoperative variables (green dots) compared with the postoperative variables (red dots). There was no significant pattern discovered for the comparison of measurement and temporal variables. A zero to mild difference was discovered for both highly complete variables (RoM < 0.1) and highly incomplete variables (RoM > 0.85). On the other hand, there was a high variation among variables with RoM between 0.1 to 0.85 across two EHRs (Figure 2- Area of High Heterogeneity).
Between the Wilcoxon signed test, variables with a significant difference in the level of RoM were Serum albumin (p<0.001), BUN (p<0.001), Bilirubin (p<0.001), Alkaline phosphatase (p<0.001), C. Diff (p<0.001), Transfer status (p=0.002), International Normalized Ratio (INR) of PT values (p=0.012).

Table 3. Comparison of Information Completeness by Operation Stage and Data Source

<table>
<thead>
<tr>
<th>Stage</th>
<th>Original Data Source</th>
<th>Total No. of Variables</th>
<th>No. of Significant Variables (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preoperative</td>
<td>ADT</td>
<td>4</td>
<td>0 (0)</td>
</tr>
<tr>
<td></td>
<td>Billing (D)</td>
<td>16</td>
<td>0 (0)</td>
</tr>
<tr>
<td></td>
<td>CN</td>
<td>26</td>
<td>1 (4)</td>
</tr>
<tr>
<td></td>
<td>Demo</td>
<td>6</td>
<td>2 (33)</td>
</tr>
<tr>
<td></td>
<td>Lab</td>
<td>36</td>
<td>16 (44)</td>
</tr>
<tr>
<td></td>
<td>Vitals</td>
<td>4</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Intraoperative</td>
<td>Billing (P)</td>
<td>1</td>
<td>0 (0)</td>
</tr>
<tr>
<td></td>
<td>Surgery</td>
<td>11</td>
<td>1 (9)</td>
</tr>
<tr>
<td>Postoperative</td>
<td>ADT</td>
<td>10</td>
<td>0 (0)</td>
</tr>
<tr>
<td></td>
<td>Billing (P)</td>
<td>3</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

*Variables were organized by the stage of surgery and summarized by the original data source. Statistically significant difference was determined by paired McNemar’s test of the data completeness between Centricity and Epic, a significant variable was defined as p<0.05; Abbreviation: ADT: admit, discharge, and transfer status, CN: clinical notes, Billing (D): diagnosis code, Billing (P): procedure code.

Data Variability The comparison of the corpus statistics between Centricity and Epic was provided in Table 4. We observed a larger number of clinical documents, tokens, and sections (total) in Epic compared with Centricity. Because the total number of documents and sections for Epic has increased, the number of sections/patient and documents/patient were higher than Centricity. On the other hand, the median of the number of tokens/patient for Epic was lower than Centricity despite the fact that the total number of documents and tokens were higher. Based on the Wilcoxon signed-rank test, all five corpus statistics metrics were found to be significant for the comparison between two EHRs. Overall, it is evident that two systems have different ways of organizing clinical documents.
Table 4. Corpus statistics of GE Centricity and Epic

<table>
<thead>
<tr>
<th>Original data source</th>
<th>Centricity</th>
<th>Epic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of patients</td>
<td>811</td>
<td>811</td>
<td></td>
</tr>
<tr>
<td>No. of documents</td>
<td>18,648</td>
<td>30,476</td>
<td></td>
</tr>
<tr>
<td>No. of tokens (Total)</td>
<td>8,273,327</td>
<td>11,383,088</td>
<td></td>
</tr>
<tr>
<td>No. of sections (Total)</td>
<td>94,645</td>
<td>116,399</td>
<td></td>
</tr>
<tr>
<td>No. of sections (Unique)</td>
<td>64</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>No. of tokens/section, median (IQR)</td>
<td>29 (95)</td>
<td>64 (90)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>No. of tokens/document, median (IQR)</td>
<td>243 (440)</td>
<td>229 (328)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>No. of sections/patient, median (IQR)</td>
<td>35,927 (42,828)</td>
<td>26,744 (36,367)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>No. of sections/patient, median (IQR)</td>
<td>4 (5)</td>
<td>3 (5)</td>
<td>0.0012</td>
</tr>
<tr>
<td>No. of documents/patient, median (IQR)</td>
<td>81 (69)</td>
<td>94 (92)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>No. of documents/patient, median (IQR)</td>
<td>15 (14)</td>
<td>26 (26)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

*IQR: interquartile range. Statistically significant difference was determined by paired Wilcoxon signed-rank test.

Based on the semantic mapping and analysis of the document sections across two EHRs, there was a high similarity of the clinical document sections across the two systems. Among the total 94,645 sections in Centricity, the top three sections were “Impression/Report/Plan” (14,203), “Chief Complaint/Reason for Visit” (7,106), and “Physical Examination” (5,853). The three most prevalent sections for Epic were Impression/Report/Plan” (20,392), “Procedure Information” (8,267), and “Physical Examination” (6,628). Among the top 15 sections, 9 sections were matched, including ‘Impression/Report/Plan’, ‘Chief Complaint/Reason for Visit’, ‘Physical Examination’, ‘History of Present Illness’, ‘Vital Signs’, ‘Subjective’, ‘Diagnosis’, ‘Procedure Information’, and ‘Social History’. The overall summary statistics of the number of PSC-related clinical concepts extracted by NLP was provided in Table 5. Overall, the concept/document ratios and concept/patient ratios for Epic (blue and orange columns) were lower for all concept types and significantly lower for Anemia, Abscess, Cellulitis and Painful incision; this pattern indicating that Centricity has potentially higher semantic breadth in the context of PSC.

Table 5. Semantic Concept Distribution of Two EHRs

<table>
<thead>
<tr>
<th>PSC Concept</th>
<th>Concept/Document Ratio</th>
<th>Concept/Document Ratio</th>
<th>Concept/Patient Ratio</th>
<th>Concept/Patient Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Centricity</td>
<td>Epic</td>
<td>Centricity</td>
<td>Epic</td>
</tr>
<tr>
<td>Anemia</td>
<td>0.1058</td>
<td>0.0524</td>
<td>1.8947</td>
<td>1.1690</td>
</tr>
<tr>
<td>Abscess</td>
<td>0.0813</td>
<td>0.0218</td>
<td>1.4562</td>
<td>0.4856</td>
</tr>
<tr>
<td>Cellulitis</td>
<td>0.0575</td>
<td>0.0097</td>
<td>1.0292</td>
<td>0.2158</td>
</tr>
<tr>
<td>Painfulincision</td>
<td>0.0201</td>
<td>0.0017</td>
<td>0.3592</td>
<td>0.0371</td>
</tr>
<tr>
<td>Purulentdrain</td>
<td>0.0076</td>
<td>0.0013</td>
<td>0.1356</td>
<td>0.0293</td>
</tr>
<tr>
<td>Woundinfec</td>
<td>0.0063</td>
<td>0.0015</td>
<td>0.1126</td>
<td>0.0339</td>
</tr>
<tr>
<td>Wounddehiscence</td>
<td>0.0020</td>
<td>0.0006</td>
<td>0.0365</td>
<td>0.0136</td>
</tr>
<tr>
<td>Fascialdehiscence</td>
<td>0.0017</td>
<td>0.0008</td>
<td>0.0302</td>
<td>0.0170</td>
</tr>
<tr>
<td>Abdomtender</td>
<td>0.0010</td>
<td>0.0007</td>
<td>0.0188</td>
<td>0.0165</td>
</tr>
<tr>
<td>Infecabdom</td>
<td>0.0009</td>
<td>0.0004</td>
<td>0.0162</td>
<td>0.0097</td>
</tr>
<tr>
<td>Fever</td>
<td>0.0008</td>
<td>0.0001</td>
<td>0.0146</td>
<td>0.0017</td>
</tr>
<tr>
<td>Reopen</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0010</td>
<td>0.0015</td>
</tr>
</tbody>
</table>

Based on the concept distribution from Table 5, we further examined the textual similarity of two most frequent concepts, anemia and abscess, and two least frequent concepts, purulent drain and wound infection. Figure 3 presents the summarized textual similarity scores for intra and inter EHR comparison. Since the document section plays an important role in contextual information, this analysis was further stratified by document sections. Compared with Epic, the intra-EHR textual similarity of Centricity was higher for all disease categories and majority of the sections. On the other hand, Epic yielded a substantially higher similarity under the ‘Secondary Diagnoses’ section. Among the section dimension, the similarity difference of ‘Diagnosis’, ‘Past Medical/Surgical History’, ‘Secondary Diagnosis’, and ‘History of Present Illness’ was substantial. For most clinical concepts and document sections, there was no substantial drop of inter-EHR similarity.
**Figure 3.** Comparison of the Textual Similarity between Centricity and Epic

The distribution of the unique clinical expressions related to abscess and anemia (Figure 4) revealed two completely opposite patterns: Epic has more standardized language for abscess, whereas Centricity is lengthy and descriptive. On the other hand, the language representing anemia for Epic was more variant than Centricity. For example, the expression of “*Anemia Posthemorrhagic Acute (Blood Loss Anemia)*” was repetitively documented for more than 30% of the total sample size. The varying similarity patterns affirms that the characteristics and patterns of data is context dependent.

*Figure (left): distribution of abscess-related concepts under the section of Secondary Diagnosis, figure (right): distribution of anemia-related concepts under the section of past medical/surgical history between Epic (blue) and Centricity (orange). X-axis: unique clinical expressions related to abscess and anemia; Y-axis: frequency of expression. Bars skewed to the left: indication of high language repetition; bars scatter to the right: indication of low language repetition.

**Figure 4.** Comparison of the Concept Distribution between Epic and Centricity

**Data timeliness** There was no delay of information found in the structured lab variables across two EHR systems. Amongst the total 811 patients who had CRS under the centricity EHR, there were a total of 1855 visits and 1673 instances of on-time documentation, and 182 instances of delayed documentation. For the other 811 matched patients under Epic EHR, there were a total of 3260 encounters, of which 44 encounters had documentation delay. The delayed documentation rate for Centricity and Epic cohorts were 0.098 and 0.013, respectively.
Discussion

The translation of predictive modeling algorithms into routine clinical care faces challenges in the form of various data quality issues caused by the heterogeneity of EHR systems. To better understand this barrier, we retrospectively assessed the variability of data from two EHR systems in the context of PSC. We discovered a consistent level of data completeness across two EHR systems with an exception for lab data. To further understand Epic’s significant improvement of capturing lab data in the context of CRS, we investigated the workflow difference between two EHRs. We learned that after EHR migration, there was a process change with how laboratory tests can be ordered. The migration enables the primary providers to order the laboratory test directly through the Epic EHRs, which may explain the lower RoM score. Conversely, there was a high syntactic variation suggested by the corpus statistics. There was also a moderate high difference in the semantic type and frequency of document sections. Textual similarity revealed a consistent pattern for roughly half of the concept-section pairs. High language variation was found for the sections of Secondary Diagnosis (Abscess) and Past Medical/Surgical History (Anemia). The data timeliness of clinical notes documentation for Epic was improved when compared with Centricity. The improvement of information timeliness from the Epic system suggests a potential higher utilization of auto- or assisted documentation. However, confirmation of this finding requires additional on-site evaluation, which we have left to a future study.

The validity and reliability of clinical data are crucial for the development of robust, safe, and scalable predictive models. However, data is often being viewed as the least incentivized aspect by ML researchers\textsuperscript{35}. Dealing with data can indeed be challenging; for example, data curation and wrangling can be time-consuming and tedious, especially in the context of secondary use of EHRs, where researchers have limited control of how the data is documented and standardized. On the other hand, because these latent factors (e.g., variant patterns of documentation) may introduce systematic bias and measurement error, a solid understanding of how data is documented, defined, and collected is required prior to adoption of any predictive models relying upon it. Solid data understanding can promote a good data curation plan and solutions for mitigating potential biases or confounders prior to model development and re-deployment. The transparency of information documentation has a direct implication to the explainability, implementability, and ultimately the trust of the models derived from the data. Based on our case study investigation on CRS patients, the EHR system plays an important role in how data is documented, defined, and organized. In a situation when a model will be translated to care practice or deployed to a different environment, proper data quality assessment needs to be conducted including the comparison of data characteristics and variability between the destination environment and the development environment.

Our investigation confirmed that the quality of data needs to be viewed from the context of data being generated and documented. For example, the results from Figure 3 discovered high similarity patterns in two EHR systems under different contextual factors (disease-section combination). Chart review was conducted to confirm expressions with high textual similarity were associated with the use of documentation templates. Although the use of templates may enhance the documentation standardization, the clinician’s reasoning process may be eliminated. The direct implication to machine learning models may be a varying level of contextual knowledge loss\textsuperscript{36}. The varying results from the analysis also strongly indicated a proper model re-training, refinement, and re-evaluation are needed.

Our study has several limitations. Since the study was conducted on a single-case scenario, the generalizability of the findings is limited by the scope of the study. We aim to expand our investigation on multiple different institutions with diverse case scenarios as part of future work. Furthermore, we plan to leverage qualitative methods to study the workflow of data documentation and transformation across multiple EHR systems.

Conclusion

To better understand the potential data heterogeneity caused by different EHR systems, we proposed and applied a standardized set of informatics methods to retrospective assess the variability of data quality contributed by two EHR systems. We discovered a varying level of data quality across two EHR systems, for which the quality of data is context-specific and closely related to the documentation workflow and the functionality of individual EHR systems. We recommend that data understanding should be equally incentivized as model development.
Acknowledgment

We gratefully thank Donna M Ihrke for performance corpus annotation and Mayo Foundation for Medical Education, NIH (#R01EB019403, U01TR002062-04) for supporting this project.

References


gene2gauss: A multi-view gaussian gene embedding learner for analyzing transcriptomic networks

Sudhir Ghandikota\textsuperscript{1,2}, Anil G. Jegga, DVM, M.Res\textsuperscript{1,3}

\textsuperscript{1}Division of Biomedical Informatics, Cincinnati Children’s Hospital Medical Center, Cincinnati, Ohio, USA; \textsuperscript{2}Department of Electrical Engineering and Computer Science, University of Cincinnati College of Engineering, Cincinnati, Ohio, USA; \textsuperscript{3}Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA

Abstract

Analyzing gene co-expression networks can help in the discovery of biological processes and regulatory mechanisms underlying normal or perturbed states. Unlike standard differential analysis, network-based approaches consider the interactions between the genes involved leading to biologically relevant results. Applying such network-based methods to jointly analyze multiple transcriptomic networks representing independent disease cohorts or studies could lead to the identification of more robust gene modules or gene regulatory networks. We present gene2gauss, a novel feature learning framework that is capable of embedding genes as multivariate gaussian distributions by taking into account their long-range interaction neighborhoods across multiple transcriptomic studies. Using multiple gene co-expression networks from idiopathic pulmonary fibrosis, we demonstrate that these multi-dimensional gaussian features are suitable for identifying regulons of known transcription factors (TF). Using standard TF-target libraries, we demonstrate that the features from our method are highly relevant in comparison with other feature learning approaches on transcriptomic data.

Introduction

Network-based module detection methods are widely used in the analysis of large-scale gene expression datasets to identify groups of co-expressed gene clusters\textsuperscript{1-3}. These methods improve upon the classical clustering methodologies by modelling local co-expression effects among the genes in a given set of samples\textsuperscript{4}. Some of them also use the gene expression similarities between regulators and target genes to infer regulatory networks\textsuperscript{5-7}. The resultant gene modules and transcription factor (TF) regulons from these methods are biologically more relevant in comparison to classic methods\textsuperscript{8}. However, they are sensitive to the inherent noise and batch effects within the transcriptomic studies. Further, most of these methods typically analyze a single study and are not capable of integrating or aggregating multiple datasets. On the other hand, graph embedding methods learn distinct representations for each node in a graph\textsuperscript{8-12} by sampling multiple random walk paths centered around each node and using them in the learning step. These representations can then be used in a variety of supervised and unsupervised tasks. Although effective, they end up using only the local contexts in a path and do not utilize the overall global structure. Graph neural networks (GNNs) are robust to structural noise within the input networks\textsuperscript{13} and have the ability to combine graph topology with multi-dimensional feature vectors assigned to each network node. Convolutional GNN models\textsuperscript{14, 15} have been applied to expression networks in supervised applications such as classifying tumor samples and have been shown to achieve impressive classification metrics\textsuperscript{16-19}. They learn representations for each node by aggregating feature vectors of their immediate neighbors\textsuperscript{14, 15}. Additionally, attention-based GNN models have the added advantage of assigning weights to the neighboring nodes and using them in the feature aggregation step\textsuperscript{20}. Multiple such graph convolutional layers could be stacked together to learn high-level node representations by encoding both local and global structural information. While, this seems viable in theory, empirical studies have shown that features from deep implementations of graph convolutional layers tend to converge to a stationary point making them indistinguishable between the different nodes\textsuperscript{21, 22}. Additionally, extending such implementations in multi-view approaches enables us to learn from multiple transcriptomic networks (i.e., similarity views) simultaneously, while being robust enough to handle the inherent noise often associated with these studies\textsuperscript{23-25}.

GNNs have also been shown to be capable of modeling “uncertainty” about a node’s aggregated features by encoding each node as a gaussian cloud (distribution) instead of a single point\textsuperscript{26}. The proximity with the neighborhood feature vectors is used to encode the parameters associated with each gaussian node representation. In this study, we propose gene2gauss, an GNN-based feature learning framework that can encode genes in a gene network as full multivariate gaussian distributions with mean and variance vectors. Additionally, gene2gauss can combine both short (one-hop) and long-range (two or more hops) interaction neighborhoods within the same layer, making it ideal for shallow neural network implementations. Recently, the Set2Gaussian\textsuperscript{27} embedding approach was used to encode gene sets (pathways...
or protein complexes) as gaussian distributions where each gene was represented as a single point. These gaussian embeddings were shown to be effective in accurately stratifying the tumor samples. Our proposed framework also supports multi-view applications where two or more gene networks can be analyzed together to learn pooled gene features from multiple studies. Overall schematic of our proposed methodology is shown in Figure 1.

**Figure 1**: Schematic representation of the gene2gauss framework. Gaussian gene features are learnt from one or more gene co-expression networks \((G^1, G^2, ..., G^k)\) representing multiple transcriptomic studies. Individual convolution operations are first applied to aggregate gene features from distinct neighborhoods in each network. These features are then pooled together and encoded into gaussian distributions.

As a proof-of-concept, we analyzed different transcriptomic datasets associated with idiopathic pulmonary fibrosis (IPF) using our gene2gauss model. IPF is a rare and severe lung disease characterized by irreversible scarring of lung parenchyma, leading to a progressive decline in lung function, and respiratory failure. Using gene2gauss, we learned multivariate gaussian embeddings that encode gene co-expression neighborhoods associated with each gene. These features were further statistically validated using several different TF-target libraries (see Methods). Finally, we also compared our framework with other neural network-based gene representation learning methods.

**Methods**

Our network learning framework requires one or more adjacency matrices corresponding to the input networks where each network \(G^k = (V, E^k)\) is an undirected, weighted (or unweighted) graph. The vertex set \(V\) is shared among all the networks while the edge sets \(E^k\) differ, representing the different similarity views being analyzed. Additionally, our framework also uses a multi-dimensional attribute matrix \(X \in \mathbb{R}^{n \times d}\) associated with the network nodes as an optional input. If not provided, gene2gauss generates and uses a matrix of identity vectors as the initial set of node features during the learning process.

A critical step in our proposed framework is the convolution operation applied individually on each network using distinct filters or feature maps. We use a novel convolutional layer that can combine aggregated node features of both direct (one-hop) and long-distance (two or more hops) neighborhoods for a given central node (Figure 2). Given two
For any given node $v$, the multiGCN convolutional layer proposed in our framework uses the Graph Convolutional Network (GCN)\textsuperscript{15} propagation rules given below:

\[
\begin{align*}
    h_v^1 &= \text{ReLU}(\sum_{u \in \overline{N}_1(v)} \deg(u) \deg(v)[\deg(v)]^{-1/2} W^1 x_u) \quad (\text{eq. 1}) \\
    h_v^2 &= \text{ReLU}(\sum_{u \in \overline{N}_2(v)} \deg(u) \deg(v)[\deg(v)]^{-1/2} W^2 x_u) \quad (\text{eq. 2})
\end{align*}
\]

where $\deg(.)$ denotes the degree of a node, $\overline{N}_1(v)$ and $\overline{N}_2(v)$ are the one-hop and two-hop neighborhoods of node $v$ in the normalized adjacency matrices $\overline{A}_1^k$ and $\overline{A}_2^k$ respectively. Also, $h_v^1 \in \mathbb{R}^d$ and $h_v^2 \in \mathbb{R}^d$ represent the aggregated feature vectors for node $v$ from both sets of neighbors while $W^1$ and $W^2$ are the respective convolutional filters. As seen in equations above, the GCN propagation step assigns a pre-determined weight to quantify the importance of each neighboring node. Conversely, the Graph Attention Networks (GATs)\textsuperscript{20} adopt self-attention mechanisms to learn the relative weights between two connected nodes\textsuperscript{28}. Therefore, we also used an attention-based convolutional layer called multiGAT that can be used to aggregate neighborhood features using learnable weights as shown below:

\[
\begin{align*}
    h_v^1 &= \text{ELU}(\sum_{u \in \overline{N}_1(v) \cup v} \alpha_{vu}^1 W^1 x_u) \quad (\text{eq. 3}) \\
    h_v^2 &= \text{ELU}(\sum_{u \in \overline{N}_2(v) \cup v} \alpha_{vu}^2 W^2 x_u) \quad (\text{eq. 4})
\end{align*}
\]

where $\alpha_{vu}$ is the attention coefficient that quantifies the connective strength between node $v$ and its neighbor $u$. These attention weights are computed using a single-layer feedforward network as seen below:

\[
\begin{align*}
    \alpha_{vu}^1 &= \text{softmax}_u(\text{LeakyReLU}(a_1 [W^1 h_v^1 \| W^1 h_u^1])) \quad (\text{eq. 5}) \\
    \alpha_{vu}^2 &= \text{softmax}_u(\text{LeakyReLU}(a_2 [W^2 h_v^2 \| W^2 h_u^2])) \quad (\text{eq. 6})
\end{align*}
\]

where $a_1$ and $a_2$ are vectors of learnable parameters associated with self-attention mechanisms for the one-hop and two-hop neighborhoods respectively. The $\text{softmax}(\cdot)$ function ensures that the weights sum up to a value of one over all neighbors of a given node.

The aggregated features from both the one-hop and two-hop neighborhoods are directly combined using a gating mechanism\textsuperscript{29} (Figure 2). The combined feature vector for a node $v$ from a network $k$ is computed as below:

\[
\begin{align*}
    h_v^{(k)} &= \text{gate}(h_v^2), h_v + (1 - \text{gate}(h_v^2)), h_v^2 \quad (\text{eq. 7}) \\
    \text{gate}(h) &= \text{ReLU}(W^{\text{gate}} h + \text{bias}) \quad (\text{eq. 8})
\end{align*}
\]

where $\text{gate}(\cdot)$ is the gating function parameterized with a weight matrix $W^{\text{gate}}$ and ReLU non-linearity. It acts as a control mechanism for combining one-hop and two-hop neighborhoods of a given gene node. While we only use up to two-hop neighbors in this paper, our framework can be extended to longer neighborhoods. Additionally, we also experimented with the vanilla GCN\textsuperscript{15} and GAT\textsuperscript{20} convolution layers where only the direct neighbors will be used in the feature aggregation step.
Finally, the combined feature matrices \( \{Z^{(1)}, Z^{(2)},...,Z^{(k)} \} \) from each network view are pooled together and encoded into mean and covariance matrices corresponding to each input gene using feed-forward non-linear neural network layers (Figure 1). These transformations could be considered as parameterized mean and variance functions \( \mu_{\theta}(h_v) \) and \( \Sigma_{\theta}(h_v) \). An energy-based link prediction loss function given below is minimized during the training step.

\[
\mathcal{L} = \sum_{v,u \in \text{pos}} E_{vu}^2 + \exp^{-E_{vu}^\text{neg}} \quad (eq. 9)
\]

The objective in eq. 9 brings together the gaussian distributions of neighboring genes closer while distancing those of unconnected ones. Therefore, to train this objective, we use sets of positive (strongly connected) and negative (unconnected) edges. Then, for each gene node \( v \), we randomly sample one highly connected node \( u_{\text{pos}} \) and one negative unconnected node \( u_{\text{neg}} \) during each epoch and use them to train the loss function. These gene sets can be based on prior biological knowledge (supervised) or can be constructed from within the input networks (self-supervised). A symmetric version of the Kullback-Leibler (KL) divergence between the learned gaussian distributions of two nodes is the energy value between them and is calculated as shown below:

\[
E_{vu} = D_{KL}[\mathcal{N}(\mu_{\theta}(Z_v),\Sigma_{\theta}(Z_v)) \| \mathcal{N}(\mu_{\theta}(Z_u),\Sigma_{\theta}(Z_u)) ] + D_{KL}[\mathcal{N}(\mu_{\theta}(Z_u),\Sigma_{\theta}(Z_u)) \| \mathcal{N}(\mu_{\theta}(Z_v),\Sigma_{\theta}(Z_v))] \quad (eq. 10)
\]

where \( D_{KL}(\cdot) \) function computes the KL divergence between any two given multi-dimensional gaussian distributions. We used the same notation used in Bojchevski et al.25 as given below:

\[
D_{KL}\left(\mathcal{N}(\mu_{\theta}(Z_v),\Sigma_{\theta}(Z_v)) \| \mathcal{N}(\mu_{\theta}(Z_u),\Sigma_{\theta}(Z_u))\right) = \frac{1}{2} \left[ \text{tr}\left(\Sigma_{\theta}^{-1}(Z_v)\Sigma_{\theta}(Z_u)\right) + (\mu_{\theta}(Z_v) - \mu_{\theta}(Z_u))^T \Sigma_{\theta}^{-1}(Z_u)(\mu_{\theta}(Z_u) - \mu_{\theta}(Z_v)) \right] - \log \left( \frac{\det(\Sigma_{\theta}(Z_u))}{\det(\Sigma_{\theta}(Z_v))} \right) \quad (eq. 11)
\]

where \( \mu_{\theta}(Z_v) \in \mathbb{R}^L \) and \( \Sigma_{\theta}(Z_v) \in \mathbb{R}^L \) are the encoded mean and variance vectors of gene \( v \), det (\cdot) and tr (\cdot) are the determinant and trace operators on a given matrix. Eventually, each gene is considered as a multivariate gaussian cloud, parameterized by mean and variance vectors that are learnt based on proximity with convolved neighborhood node features. Further, KL divergences computed between the final gene embeddings can be used to construct gene-gene similarity matrices useful for identifying consensus modules and deciphering regulatory networks.

**Results**

We demonstrate the applicability of our \textit{gene2gauss} framework by analyzing four different whole lung transcriptomic datasets in IPF (Table 1). Processed gene expression data from each of the four studies were downloaded from the NCBI’s Gene Expression Omnibus (GEO)30. In case of studies with raw counts, we applied variance stabilizing transformations31 during the pre-processing phase. Prior to the learning step, we build a gene co-expression network from each study. Any of the standard correlation metrics can be used for this step. In this case study, we computed pairwise Pearson correlations between two gene expression profiles and used them to build a weighted gene network. For computational reasons, we filtered the final gene networks to retain only the edges passing a correlation threshold of \( \geq 0.5 \). In total, 8454 genes were used as network nodes and Table 1 below lists the number of edges from each input network used in our case study.

**Table 1.** List of whole lung transcriptomic studies used in our IPF case-study. The individual sample counts along with the edge coverage within the co-expression networks are also included.

<table>
<thead>
<tr>
<th>GEO ID</th>
<th>Sample Counts</th>
<th>Edge Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE47460</td>
<td>160 IPF; 108 Controls</td>
<td>2,092,012</td>
</tr>
<tr>
<td>GSE10667</td>
<td>31 IPF; 15 Controls</td>
<td>4,799,152</td>
</tr>
<tr>
<td>GSE53845</td>
<td>40 IPF; 8 Controls</td>
<td>1,832,248</td>
</tr>
<tr>
<td>GSE48149</td>
<td>13 IPF; 9 Controls</td>
<td>2,412,172</td>
</tr>
</tbody>
</table>

Figure 3 below shows the distributions of node degrees within each input network. Additionally, cell type associations from a recent single-cell lung RNA-sequencing (scRNA-seq) study35 were used as multi-dimensional node attribute vectors where the dimensionality is equal to 59 distinct cell types studied in Travaglini et al.35. A value of 1 indicates...
that the gene is found to be expressed (FDR p-value ≤ 0.05; log Fold Change ≥ 0.5) in a given cell type and a value of 0 indicates otherwise.

Figure 3: Distributions of node degrees among the input gene co-expression networks.

We implemented a self-supervised approach to train our models by extracting the edge sets for minimizing the link-prediction loss function (eq. 9) from within the training networks themselves. For the positive set, we collected all edges with strong co-expression correlation values (≥ 0.8) while for the negative set, we randomly sampled the same number of disconnected gene pairs. To choose the various model hyperparameters involved, we employed a grid search over four different choices of convolutional layers (GCN, GAT, multiGCN and multiGAT), four different pooling schemes (CONCAT, MEAN, MAX and SUM), and convolutional filter counts (\(D\)) (64, 128, 256, and 512) and finally three different gaussian embedding dimension (\(L\)) values (64, 128, and 256). Each model is trained for a maximum of 100 epochs (early stopping with a tolerance value of 5 epochs) using the Adam optimizer with a learning rate of 0.01. All the weight matrices are initialized using the method described in Glorot et al. We also applied L2 regularization for convolutional layer weights using the default regularization parameter of \(\lambda = 0.01\).

To compare the embeddings learned in different gene2gauss experiments, we have generated the neighborhoods of more than 700 TFs and statistically compared them to multiple putative TF-target libraries from ChEA3. For each experiment, we first computed a gene distance matrix by calculating pairwise KL divergences between the gaussian distributions. Then, for each TF, we retrieved 300 closest genes as its regulatory neighborhood to keep it consistent with the regulon size in ChEA3 library. Fisher’s exact tests were performed to test the enrichments of the generated neighborhoods against the six primary libraries (ARCHS4_Coexpression, ENCODE_ChIP-seq, Enrichr_Queries, GTEx_Coexpression, Literature_ChIP-seq, and ReMap_ChIP-seq) and a lung tissue specific TF-target library published in ChEA3. Finally, we compared the observed enrichment levels of all TFs by performing one-sided Mann-Whitney tests of mean enrichment ranks across all libraries. We followed the same approach in our comparisons with two other feature-learning baseline methods.

In our evaluations, surprisingly, we observed that the multiGAT convolutional layers significantly underperformed when compared to the other three (Figure 4), with multiGCN and GAT layers performing slightly better in our pairwise comparisons. Similarly, we did not find any significant differences between the different feature pooling mechanisms, although concatenating the aggregated features from each input was found to be slightly better (Figure 5a). We also did not observe any significant impact of convolution filter counts (\(D\)) on our final gene neighborhoods except in the case of 256-dimensional gaussian embeddings (Figure 5b).
Figure 4: Panel of heatmaps showing the negative log p-values from pairwise Mann-Whitney tests used for analyzing the impact of different pooling mechanisms on the proposed convolutional layers.

Figure 5: Heatmaps of pairwise Mann-Whitney test log p-values for (a) assessing the impact of pooling operations among the 128-dimensional gene gaussian embeddings generating using multiGCN convolutional layers, (b) identifying the impact of hidden layer (i.e., convolution layer) dimensions on the final multiGCN-based gaussian embeddings of different sizes (64, 128 and 256), and (c) for comparing features from gene2gauss with two other baselines and randomly generated gene embeddings.
Next, we compared our method with gene features from two other feature learning methodologies\textsuperscript{39, 40} applicable for gene transcriptomic studies. In case of gene2vec\textsuperscript{46}, we generated both 128-dimensional and 200-dimensional gene features (recommended by authors) using co-expressed gene pairs ($\geq 0.5$) from the 4 input studies. Similarly, we extracted 128 and 256 dimensional gene features using the multi-task learning method (labeled as gene2vec+PPI) proposed in one of our previous works\textsuperscript{39}, again using the same set of gene pairs. Further, we compared our features with randomly generated gaussian mean and variance vectors. We repeated the same procedure explained above to compare the different gene neighborhoods coming out of the baseline methods. From our comparisons, neighborhoods based on both multiGCN, and GAT convolutional filters significantly outperformed the rest of the baselines (Figure 5c). This potentially displays the advantage of utilizing the geometric structure through weighted, attributed co-expression networks as inputs to our gene2gauss framework. Interestingly, features from gene2vec+PPI method (128-dimensions) were found to be relatively comparable to our method, although still significantly less relevant. This is probably because it also uses protein-protein interactions (PPI) during training thereby including a certain level of high-level structure.

Finally, we analyzed the gene regulatory neighborhoods from the gene2gauss implementation by computing their enrichments against single-cell marker genes in IPF lung tissues\textsuperscript{41, 42}. To do this, we used around 400 cell type-selective TFs in human lung tissues\textsuperscript{35} and identified their IPF-specific neighborhoods based on the learned gaussian embeddings. Subsequently, we tested these neighborhoods for enrichments of single cell RNA-seq markers from two independent studies\textsuperscript{41, 42} in fibrotic lung tissue. We then ranked the cell types enriched in each TF neighborhood based on Fisher’s exact test p-values. In our results, we observed that the enriched cell types in the identified gene neighborhoods were in concordance with the cell type of the queried TF. For instance, FOXJ1 is a known canonical marker and a master regulator of the motile ciliogenic program that controls the production of motile cilia\textsuperscript{43}. The gene2gauss-derived gene neighborhood of FOXJ1 showed significant enrichment for cilia cell type markers (129/300 genes; p-value = 2.86E-100). Table 2 below lists other lung TFs whose gene2gauss-derived gene neighborhoods showed enrichment for the respective cell types from human lung. Additionally, we also generated disease-specific gene neighborhoods of known IPF genes compiled from GWAS Catalog\textsuperscript{44} and implemented the marker enrichment analysis. The list of IPF candidate genes along with the full enrichment results can be found in our GitHub repository (https://github.com/SudhirGhandikota/gene2gauss). These results illustrate that our network learning framework is able to bring together multivariate gaussian distributions of genes that are expressed in the same broad cell types within the lung tissue. Hence, it can be hypothesized that the final gene clusters constructed based on our features represent biological knowledge through distinct signaling pathways or biological processes involved in IPF.

Table 2: Cell marker enrichments among regulatory neighborhoods (300 genes) of 8 different lung-specific TFs that are also known canonical markers of different lung cell types. Only the top 3 enriched (FDR p-value $\leq 0.05$) cell types (from Adams et al.\textsuperscript{41}) for each TF are listed here. The actual cell types associated are highlighted in bold.

<table>
<thead>
<tr>
<th>TF (cell type)</th>
<th>Enriched cell types of the gene2gauss-derived gene neighborhood (300 genes) of TFs</th>
<th>Overlap Count (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSHZ2 (Myofibroblasts)</td>
<td>Fibroblasts</td>
<td>45 genes (4.94E-25)</td>
</tr>
<tr>
<td></td>
<td>Myofibroblasts</td>
<td>38 genes (7.03E-20)</td>
</tr>
<tr>
<td></td>
<td>Smooth muscle cells</td>
<td>27 genes (3.10E-12)</td>
</tr>
<tr>
<td>LEF1 (Pericytes)</td>
<td>Fibroblasts</td>
<td>52 genes (2.94E-32)</td>
</tr>
<tr>
<td></td>
<td>Myofibroblasts</td>
<td>47 genes (7.53E-29)</td>
</tr>
<tr>
<td></td>
<td>Pericytes</td>
<td>34 genes (2.75E-17)</td>
</tr>
<tr>
<td>MYC (Bronchial Vessel 1, Fibroblasts)</td>
<td>Mesothelial cells</td>
<td>36 genes (3.69E-17)</td>
</tr>
<tr>
<td></td>
<td>Peribronchial cells</td>
<td>34 genes (3.09E-14)</td>
</tr>
<tr>
<td></td>
<td>Fibroblasts</td>
<td>31 genes (9.64E-13)</td>
</tr>
<tr>
<td></td>
<td>Ciliated cells</td>
<td>129 genes (2.86E-100)</td>
</tr>
</tbody>
</table>
where the attention weights can correspond to the view (learnable) feature pooling approaches that can weigh used in neighborhoods for the standard convolutional layers indistinguishable are particularly not ideal for attention limiting the maximum number of k-hop neighbors that can be used for feature aggregation. Such dense neighborhoods are particularly not ideal for attention-based approaches where the attention weights assigned can become indistinguishable due to the large node counts. This was partially observed in our IPF case study where the multiGAT convolutional layers, which make use of dense two-hop neighbors, significantly underperformed in comparison with the standard GAT layers. We plan to explore GNN mechanisms that can handle and leverage such dense long-range neighborhoods for supervised or semi-supervised graph learning applications. Secondly, the feature pooling methods used in gene2gauss assign equal weights to each of the input networks. As future work, we plan to include weighted (learnable) feature pooling approaches that can weigh the feature vectors coming out of each individual similarity view, acting as an additional self-attention layer. Such approaches are especially useful for supervised applications where the attention weights can correspond to the contribution of a particular input network to the supervised task at

| FOXJ1 (Ciliated cells) | Goblet cells | 19 genes (6.00E-05) |
| SPDEF (Goblet cells) | Ciliated cells | 105 genes (2.84E-69) |
| | Goblet cells | 26 genes (4.62E-09) |
| | Ionocytes | 20 genes (2.25E-03) |
| ASCL1 (Neuroendocrine cells) | Ciliated cells | 40 genes (2.80E-10) |
| | Pulmonary neuroendocrine cells | 25 genes (1.01E-07) |
| | Ionocytes | 27 genes (2.02E-06) |
| TP73 (Ciliated cells) | Ciliated cells | 119 genes (8.27E-87) |
| | Goblet cells | 19 genes (6.00E-05) |
| | Ionocytes | 19 genes (5.03E-03) |
| PPARG (Macrophages) | Macrophages | 50 genes (1.52E-25) |
| | Alveolar Macrophages | 30 genes (6.61E-19) |
| | AT2 cells | 19 genes (2.90E-05) |

**Conclusion**

In this study, we propose gene2gauss, a novel gaussian embedding learning framework than can be used to analyze one or more gene networks together to encode individual gene nodes as multi-dimensional gaussian distributions. It utilizes graph convolutional layers that are capable of aggregating features from both direct (one-hop) and long-range (two or more hops) neighbors. Additionally, different gene functional annotations can be included in the analysis step as node attributes. Our framework can be applied in both supervised and unsupervised (self-supervised) settings. The learned gaussian gene features can be used in a variety of use cases including, identifying consensus gene clusters, and deciphering regulatory neighborhoods of known candidate genes (e.g., Genome-Wide association studies), TFs, or ligands. We showed the utility of our framework by analyzing four transcriptomic networks associated with IPF and identified IPF-specific gene neighborhoods of lung TFs and known IPF candidate genes. We also assessed the impact of different convolutional layers, filter counts, feature pooling mechanisms and output dimensionality. Additionally, we also tried to test the impact of network density by running our framework on full (correlation ≥ 0) gene co-expression networks and repeated the same set of validation experiments. In our evaluations, we did not observe any significant impact of increased neighborhood density on the final learned features. Finally, we compared the learned features in IPF with two other neural network-based representation learning frameworks. By analyzing these gene neighborhoods, we found several candidate TFs displaying a significant amount of the cell-specific regulon activity within IPF. In practice, we used TensorFlow (GPU-based version) for efficient implementation of our model which can be freely accessed at https://github.com/SudhirGhandikota/gene2gauss.

Our gene2gauss framework has some limitations. Firstly, gene co-expression networks are inherently dense, thus limiting the maximum number of k-hop neighbors that can be used for feature aggregation. Such dense neighborhoods are particularly not ideal for attention-based approaches where the attention weights assigned can become indistinguishable due to the large node counts. This was partially observed in our IPF case study where the multiGAT convolutional layers, which make use of dense two-hop neighbors, significantly underperformed in comparison with the standard GAT layers. We plan to explore GNN mechanisms that can handle and leverage such dense long-range neighborhoods for supervised or semi-supervised graph learning applications. Secondly, the feature pooling methods used in gene2gauss assign equal weights to each of the input networks. As future work, we plan to include weighted (learnable) feature pooling approaches that can weigh the feature vectors coming out of each individual similarity view, acting as an additional self-attention layer. Such approaches are especially useful for supervised applications where the attention weights can correspond to the contribution of a particular input network to the supervised task at 213
hand. Finally, since gene2gauss is a transductive framework and does not naturally generalize to unseen gene nodes, additional re-training (for the new nodes) may be required to overcome this limitation.

References


A Day-to-Day Approach for Automating the Hospital Course Section of the Discharge Summary

Vince Hartman, BA¹, Thomas R. Campion, Jr., PhD²

¹Cornell Tech, New York, New York; ²Weill Cornell Medicine, New York, New York

Abstract

Optimal solutions for abstractive summarization of electronic health record content have yet to be discovered. Although studies have applied state-of-the-art transformers in the clinical domain to radiology reports and information extraction, little is known of transformers’ performance with the hospital course section of the discharge summary. This paper compares two summarization approaches for automating the hospital course section within the discharge summary: (1) a truncation approach that uses all clinical notes and (2) a day-to-day approach that segments the notes per clinical day. We pair both approaches with different transformer encoder-decoder based-models - BART, BERT2GPT2, ClinicalBERT2GPT2, and ClinicalBERT2ClinicalBERT - and evaluate the transformers that work best for each approach using ROUGE metrics. The results demonstrate that the day-to-day approach can overcome the limitations of longform document summarization for the patient clinical record.

Introduction

Over the past 20 years, various studies have investigated automating patient summaries using electronic health record (EHR) content [1, 2, 3, 4, 5]. Automated clinical notes could help address the issue of physician burnout [6, 7] as EHRs have led to physicians spending more time in front of a computer than with a patient. One study found that physicians spent 44% of their time using the EHR during their workday, with much of that time spent on clinical documentation [8]. EHR templates have improved some efficiency for physicians’ note writing, but at the expense of lower quality notes compared to ones written through dictation. To overcome this challenge of onerous tasks in an EHR, Natural Language Processing (NLP) has been posited as a technology that will help alleviate the burdens of regulatory and clinical requirements.

Transformers are currently the state-of-the-art approach for NLP and generative text [9]. Given that encoder-decoder transformers have an input limit of either 512 or 1024 tokens because of modern-day memory constraints, it is infeasible to train a transformer using the entirety of the clinical chart to summarize a patient’s course of treatment. So, implementing a transformer with textual input that is longer than the transformer’s input limit has typically required one of three strategies: (1) truncation [10, 11], (2) extractive summarization of individual sections and then combining and performing a second abstractive summarization step [12, 5, 13], or (3) using a recurrent neural network (RNN) based model that allows for unlimited sequential data as either the input or output to the transformer [14, 15]. In some applications such as movie reviews [16], a truncation method of using the beginning and end of a long document works well for textual summarization. In a clinical context though, truncation has generally not been pursued since the middle sections of the patient record can have as much salience as the beginning and end. The extractive-abstractive pipeline has been the recommended approach for longform document summarization of the patient record [5, 17] as it mirrors the physician workflow where physicians first extract important concepts from the patient record followed by them generating a unique summary in their own words.

In this study, we compare two approaches of using the patient record to automate the hospital course section in the discharge summary. The discharge summary is the primary document for communicating clinical information to outpatient providers at the conclusion of an inpatient stay [18]. In addition to the hospital course section, the discharge summary contains other elements including but not limited to the principal diagnosis, the past medical and social history, allergies, medications, and laboratory data (See Figure 1). These other elements of the discharge summary have been mostly automated through data extraction using clinical note templates in the EHR[19]; the hospital course section has not [17]. Automating the hospital
course section would be useful as inpatient physicians would save time compared to the current approach of summarizing EHR content manually [8], discharge reports would be sent faster and more consistently to outpatient physicians [20, 21], and the hospital course section would be more concise and not be overburdened by “note bloat” [22]. While useful, automating the hospital course section is challenging as the narrative is composed from a large corpus of clinical data. For this reason, no adequate approaches currently exist for automating the hospital course section. We explore these limitations through two approaches: one that relies on document truncation and a second that segments the input data into three unique components. Both of these approaches use state-of-the-art transformers to create a textual abstractive summary.

![Figure 1.](image)

### Methods

**Dataset**

MIMIC-III [23] consists of 53,423 de-identified digital hospital admissions from 38,597 distinct patients that were seen in the critical care unit of Beth Israel Deaconess Medical Center in Boston, Massachusetts between 2001 and 2012. Within each patient record, clinicians captured different types of data during the course of a patient’s admission including but not limited to lab results, medications, progress notes, discharge summaries, and vital signs. We focused on the data listed in Table 1.

For the discharge summary notes in the MIMIC-III dataset, we extracted the hospital course section using a regular expression. We identified 45,167 hospital course sections from 52,000 discharge summary notes. We filtered on these 45,167 hospital course sections further based on the summarization approaches discussed in the Truncation and Day-to-day sections. For all of the datasets, we constructed a train, validation and test structure with a ratio of 80:10:10.
Table 1. Data types used from the MIMIC-III data set

<table>
<thead>
<tr>
<th>Types of Data</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive</td>
<td>Admission diagnosis, admission and discharge times, admission and discharge location, record of death, event times, insurance, religion, age, gender, marital status, and ethnicity</td>
</tr>
<tr>
<td>Interventions</td>
<td>Procedures performed such as dialysis and imaging studies</td>
</tr>
<tr>
<td>Notes</td>
<td>Echo, ECG, nursing, physician, rehab, case management, respiratory, general, nutrition, consult, social work, pharmacy, radiology, other, and discharge summary</td>
</tr>
<tr>
<td>Reports</td>
<td>Electrocardiogram and imaging studies</td>
</tr>
</tbody>
</table>

**Truncation Summarization Approach**

As a baseline, we constructed a dataset that used as input the complete clinical record of a patient and as output the full hospital course section.

In a previous unpublished study, we used all 45,167 hospital course sections and their corresponding notes as a baseline, but performance was poor due to limitations of the MIMIC-III dataset in that the patient clinical record is incomplete. The MIMIC-III dataset more or less only includes content during the patient’s stay in the ICU rather than data from the full hospital course of treatment.

In an effort to attain meaningful results, we limited the dataset to the hospital course sections that had a corresponding admission note, which is a required note a physician completes for each hospital admission [24]. This strategy produced 5,877 records. For these patient records, we combined the unstructured data from all other clinical notes in chronological order alongside the structured data from Table 1. We combined the unstructured and structured data as seen in Figure 2.

**Day-to-day Summarization Approach**

Given the challenges with longform document summarization with transformers, we implemented an approach that segments notes per day. We refer to this approach as day-to-day summarization for predicting the hospital course section of the discharge summary. Similar approaches have demonstrated improved performance with summarization and clinical extraction in previous research [25, 13].

We first segmented the 45,167 hospital course sections of the MIMIC-III dataset into three separate datasets (see Figure 1): (1) the history of present illness (HPI), (2) the daily narrative, and (3) the discharge plan. By segmenting the hospital course section into these three separate sections, we can use distinct data from the
clinical chart and prioritize it accordingly. Thereby, we reduce the amount of input text for the transformer model (see Figure 3). The three segmented datasets are constructed as follows:

1. HPI Segment: the first sentence of the hospital course section
2. Daily Narrative Segment: any sentence within the hospital course section that has a date listed in the format of month-day
3. Discharge Plan Segment: any sentence within the hospital course section that contains one or more of the following words: discharge, death, deceased, died, follow-up, or AMA

Figure 3. The day-to-day summarization data flow of the clinical data input through the transformers to the summary sentences. The sentences are combined in chronological order to form the hospital course section.

Segment 1: History of Present Illness (HPI)

Example 1: Ms. [**lastname**] is an 88 year old female with history of CAD, hypertension who presents with epigastric and chest pain with evidence of pneumonia on CXR.

Example 2: 47 yo M with history of seizure disorder and EtOH abuse brought to ED by EMS after a seizure at his shelter.

The hospital course section of the discharge summary begins with one to two introductory sentences about the patient and presents the patient’s clinical history and chief complaint (see examples). These two sentences of the hospital course section are the most formulaic in the summary and are derived from either the admission note or the history & physical (H&P) note [17]. For the MIMIC-III dataset, the structure is generally as follows: “Mr/s. [Patient Last Name] is an [Age] [Gender] with a history of [Medical Hx] who presented with [Admission Diagnosis] on [Admit Date] and was found to have [Detailed Reason for Admission].” Given how formulaic the first sentence of the hospital course section is, the NLP transformer models would be expected to perform well given the improved performance of transformers when supplied with structured data [13].

To construct the HPI summarization dataset, we limited the dataset to hospital admissions that have a corresponding admission note. The 45,167 hospital course sections in the MIMIC-III dataset had 5,877 associated admission notes (as was stated in the truncation approach). With these admission notes, we constructed a regex to extract the HPI. In addition to the HPI sentences from the admission note, the input
dataset included the admission type, admission location, marital status, ethnicity, admission diagnosis, patient gender, and patient age for each hospital admission (See Figure 3).

**Segment 2: Daily Narrative**

Example 1: On [**12-22**] coumadin was adjusted to previous dose of 1 mg daily.

Example 2: On the evening of [**11-19**], the patient was in Afib with RVR.

The middle portion of the hospital course section is an overall summary of what occurs each day while the patient is in the hospital. The day-to-day approach takes the clinical input for a specific day and produces one to two summary sentences for each of those days. These daily summaries are combined to create a complete narrative from admission to discharge.

The primary documents for each clinical day included in the day-to-day approach are the physician progress notes, consult notes, and procedure notes. These notes were chosen based on a study which has found that physicians spend the vast majority of their time reviewing the assessment, impression and plan sections of these notes compared to other clinical data such as medications, vital signs, and laboratory results [26].

For the daily narrative dataset, we extracted from each hospital course section any sentences that included a date during the hospital admission (see examples). For the MIMIC-III dataset, dates are listed in the format of [**DATE**]. We filtered on corresponding notes during the hospital admission that occurred on that date. We saved all of these notes with the note category, description, and provider title (MD, RN, LPN, etc.). If a sentence from the hospital course section for that date did not have any corresponding clinical notes, it was excluded from the dataset. The daily narrative dataset included 12,058 sentences from the hospital course sections with dates and corresponding clinical notes. In addition to the corresponding clinical notes occurring for that date, the input dataset included the admission type, admission location, marital status, ethnicity, admission diagnosis, patient gender, and patient age for each hospital admission (See Figure 3).

**Segment 3: Discharge Plan**

Example 1: She was discharged in stable condition to home with home oxygen as she had ambulatory desats to 87%.

Example 2: Due to possible ligamentous injury, he will wear a soft cervical collar for four weeks; he will follow-up in the trauma clinic if his pain persists beyond four weeks.

Example 3: Patient seemed comfortable and immediate cause of death was cardiac arrest secondary to sepsis.

The hospital course section concludes with one to two sentences describing either the discharge plans or the expiration of the patient (see example). Discharge plans primarily consist of stating the post-discharge instructions for the patient so downstream providers can follow-up with the patient to ensure compliance and monitor continual treatment [27]. Discharge planning is a clinical team approach and includes staff members such as physicians, nursing, case managers, and social work with all of them documenting these plans within their own respective daily notes [28]. Lastly, if the patient expires, providers have to document a brief statement and cause of death in the hospital chart; this information is restated at the end of the hospital course section for expired patients.

With the discharge plan summarization dataset, we filtered each hospital admission that contained one of the following notes within 2 days of discharge: case management, social work, physician, nursing, or a death note. We filtered these notes further to include at least one of the following discharge related words: discharge, transfer, death, deceased, died, follow-up, priest, or AMA. 5,077 of the hospital course sections had one of these associated notes. In addition to the discharge plan notes, the input dataset included the discharge location, and a hospital expiration indicator (See Figure 3).
Summarization Models

To understand and compare the two methods, we paired the approaches with five different summarization based-models: TextRank, BART, BERT2GPT2, ClinicalBERT2GPT2, and ClinicalBERT2ClinicalBERT. These different models were interchanged and tested in the transformer steps illustrated in Figures 2 and 3. Although the focus is on abstractive summarization, we used a rule-based extractive summary model, TextRank, as a benchmark for comparison with the other transformer models.

TextRank is an algorithm that uses the structure of sentences in a document to determine key-phrases that appear central to conceptual understanding for a summary [29]. TextRank does not rely on training data nor supervision, but selects sentences based on their relative importance. The algorithm can be extended quite easily to large documents but performs poorly when the output text does not contain many words from the original source documents.

BART was used as a model for abstractive summarization [30]. We used the pre-trained BART-large-CNN transformer model and BART-large tokenizer from Hugging Face, built with PyTorch, and trained on non-domain specific text. We performed transfer learning using our dataset for 3 epochs and 4 beams with a max input token length of 1,024 and a max output token length of 50.

BERT2GPT2 and ClinicalBERT2GPT2 were used as models from Hugging Face [31]. We constructed two encoder-decoder models using BERT-base cased and ClinicalBERT [32] as the encoders and GPT2 [33] as the decoder. We performed transfer learning for both models for 3 epochs with a max input token length of 512, max output token length of 142, 4 beams, and a length penalty of 2.

ClinicalBERT2ClinicalBERT was used as an encoder-decoder model from Hugging Face [31], with ClinicalBERT in place of the BERT-base cased checkpoint [32]. We performed transfer learning for 3 epochs with a max input token length of 512, max output token length of 142, 4 beams, and a length penalty of 2.

Evaluation

We evaluated the performance of the transformer models used in the truncation and day-to-day summarization approaches with ROUGE, or Recall-Oriented Understudy for Gisting Evaluation. ROUGE is a collection of metrics that measure the precision and recall of the computer-generated text against the original human text (the reference). ROUGE is a common metric for measuring the performance of abstraction using matching n-gram pairs and serves as a benchmark for evaluating NLP models across domains. We used three ROUGE metrics: ROUGE-1, ROUGE-2, and ROUGE-L. ROUGE-1 measures the overlap of unigrams between the computer-generated summary and the reference summary; ROUGE-2 is the overlap of bigrams; and ROUGE-L measures the longest matching sequence of words using the longest common subsequence.

ROUGE-2 is generally identified as the best ROUGE score for comparing two summarization systems [34]. In particular, Rouge-2 recall is cited most often by researchers within the domain of text summarization as it does not allow for artificially inflating ones results by producing an excessively long summary. So, we determined the best transformer model for each summarization approach as given by the model with the highest ROUGE-2 recall scores. As a reference for understating good ROUGE recall scores with longform document summarization, state-of-the-art performance ranges from 41-48 for ROUGE-1, 18-21 for ROUGE-2, and 36-42 for ROUGE-L [35, 36, 37].

Results

We fine-tuned the BART, ClinicalBERT2ClinicalBERT, BERT2GPT2, and ClinicalBERT2GPT2 models for the training and validation datasets for the truncation and day-to-day summarization approaches for 3 epochs. After fine tuning the models, we obtained results on the test datasets which we list in Table 2.

The truncation summarization approach had a best performance with the BART model with a ROUGE-2 score of 10.58. Overall, scores for the truncation approach models were much lower when compared to the state-of-the-art scores as discussed in the Evaluation section. The input data was quite large at 28,539 word tokens on average and 99.95% of the dataset having more than 1,024 tokens. Since BART has a word token maximum of 1,024 tokens and BERT at 512, this confirms that the dataset was truncated.
### Table 2. ROUGE recall scores \{R-1/R-2/R-L\} for the truncation and the day-to-day summarization models with each segmented section of the hospital course section (HPI, daily, and discharge plan).

<table>
<thead>
<tr>
<th></th>
<th>R-1</th>
<th>R-2</th>
<th>R-L</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Truncation approach:</strong> complete clinical record</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TextRank</td>
<td>21.13</td>
<td>1.97</td>
<td>10.24</td>
</tr>
<tr>
<td>BART</td>
<td><strong>39.57</strong></td>
<td><strong>10.58</strong></td>
<td><strong>18.41</strong></td>
</tr>
<tr>
<td>BERT2GPT2</td>
<td>11.48</td>
<td>2.59</td>
<td>6.83</td>
</tr>
<tr>
<td>ClinicalBERT2GPT2</td>
<td>12.89</td>
<td>3.45</td>
<td>7.69</td>
</tr>
<tr>
<td>ClinicalBERT2ClinicalBERT</td>
<td>10.61</td>
<td>3.03</td>
<td>6.74</td>
</tr>
<tr>
<td><strong>Day-to-day approach:</strong> HPI segment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TextRank</td>
<td>24.33</td>
<td>10.42</td>
<td>20.42</td>
</tr>
<tr>
<td>BART</td>
<td><strong>29.65</strong></td>
<td><strong>16.49</strong></td>
<td><strong>28.05</strong></td>
</tr>
<tr>
<td>BERT2GPT2</td>
<td>33.31</td>
<td>10.02</td>
<td>26.60</td>
</tr>
<tr>
<td>ClinicalBERT2GPT2</td>
<td>36.61</td>
<td>13.51</td>
<td>29.19</td>
</tr>
<tr>
<td>ClinicalBERT2ClinicalBERT</td>
<td>30.36</td>
<td>12.13</td>
<td>26.74</td>
</tr>
<tr>
<td><strong>Day-to-day approach:</strong> daily segment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TextRank</td>
<td>16.66</td>
<td>6.17</td>
<td>13.98</td>
</tr>
<tr>
<td>BART</td>
<td>30.31</td>
<td>13.68</td>
<td>23.74</td>
</tr>
<tr>
<td>BERT2GPT2</td>
<td>35.92</td>
<td>6.17</td>
<td>27.80</td>
</tr>
<tr>
<td>ClinicalBERT2GPT2</td>
<td>42.52</td>
<td>13.93</td>
<td>31.22</td>
</tr>
<tr>
<td>ClinicalBERT2ClinicalBERT</td>
<td><strong>39.11</strong></td>
<td><strong>16.87</strong></td>
<td><strong>30.24</strong></td>
</tr>
<tr>
<td><strong>Day-to-day approach:</strong> discharge segment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TextRank</td>
<td>8.78</td>
<td>0.50</td>
<td>6.54</td>
</tr>
<tr>
<td>BART</td>
<td>24.62</td>
<td>8.35</td>
<td>19.22</td>
</tr>
<tr>
<td>BERT2GPT2</td>
<td>31.65</td>
<td>7.57</td>
<td>21.94</td>
</tr>
<tr>
<td>ClinicalBERT2GPT2</td>
<td><strong>35.73</strong></td>
<td><strong>10.65</strong></td>
<td><strong>25.33</strong></td>
</tr>
<tr>
<td>ClinicalBERT2ClinicalBERT</td>
<td>30.84</td>
<td>9.84</td>
<td>23.09</td>
</tr>
</tbody>
</table>

The day-to-day summarization approach saw a combination of different transformers perform better depending on the segmentation task. The HPI segment performed best with BART, the daily segment best with ClinicalBERT2ClinicalBERT, and the discharge plan best with ClinicalBERT2GPT2. For the HPI and daily segmentation tasks, the ROUGE-2 scores were close to state-of-the-art results for longform document summarization. The results for the discharge plan segmentation task were overall much lower across all models which implies the dataset was not as well constructed as compared to the HPI and daily segments. As the input text was filtered and curated with the day-to-day approach and did not include the full clinical chart, the average word tokens was significantly lower for each segmentation task when compared to the truncation approach. The daily segmentation task had the highest average input with 1,307 word tokens and 80% of the dataset having less than 1,024 tokens. This implies that none of the three segmentation tasks with the day-to-day approach saw issues with truncation.

In both approaches, the BERT2GPT2 model performed relatively poorly compared to the other transformers. At times the BERT2GPT2 model performed no better than simplistic extraction as seen with it’s comparable scores with TextRank.

**Discussion**

Even if we selected BART as our transformer model for the truncation approach, the approach would not be sufficient for automatically summarizing the hospital course section in the discharge summary given how low our ROUGE scores are. Since we tested with a full collection of state-of-the-art transformers for the clinical domain and filtered the dataset for completeness, it seems unlikely that another transformer model would perform significantly better using the truncation approach. This implies that the poor results are related to truncation and a different strategy is indeed required. Given the promising results, we saw with
the day-to-day approach, it has the potential to overcome the issues with truncation given the improved performance of the ROUGE-2 scores for the HPI and daily segments.

The poor performance of the models for the discharge segment in the day-to-day approach appears to be caused from a poor construction of the input data or a limitation of the dataset. Since TextRank performed markedly badly with a score of 0.5, there seems to be little extractive similarity between the input data and the sentences from the hospital course section. In future work, we would recommend further filtering strategies for the discharge segment task or using a different dataset than MIMIC-III. A recent dataset published of clinical action items annotated from MIMIC-III discharge summaries may also prove useful [27].

For future studies, we would recommend two main improvements related to the limitations of this study: using a dataset with more robust and complete patient records than MIMIC-III, and recruiting clinicians to perform a convenience sample of the computer-generated hospital course sections to measure quality and faithfulness metrics. We plan to implement both of these improvements in a future study in partnership with Weill-Cornell under the direction of an Institutional Review Board (IRB). While a ROUGE score is a useful benchmark for gauging the performance of an abstractive summarization model, ROUGE is poorly constructed for communicating quality. Abstractive summaries use textual synthesis, so they may inherently have few matching n-grams between the human and computer summaries, but, in fact, the summaries relay the same factual information. Thus, measuring the quality of a model generally requires performing a convenience sample with domain experts comparing the computer-generated summaries against the human-generated ones [30, 38, 3].

Conclusion

Transformers have become the model of choice to solve NLP tasks such as text summarization. We present two approaches which incorporate transformers for summarizing the patient record. The results demonstrate that the approach of dividing the hospital course section into segmented smaller components and then reconstructing an ensemble chronological summary can overcome the issues of truncation that was encountered when feeding the full patient record into a transformer model. Through the experiments performed, we show the potential of this approach and how it could assist physicians with writing the hospital course section via automation.

Acknowledgements

This study received support from New York-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


Temporal Subtyping of Alzheimer’s Disease Using Medical Conditions Preceding Alzheimer’s Disease Onset in Electronic Health Records

Zhe He, PhD1,*, Shubo Tian, BS1,*, Arslan Erdengasileng, MS1, Neil Charness, PhD1, Jiang Bian, PhD2
1Florida State University, Tallahassee, Florida USA
2University of Florida, Gainesville, Florida USA

Abstract

Subtyping of Alzheimer’s disease (AD) can facilitate diagnosis, treatment, prognosis and disease management. It can also support the testing of new prevention and treatment strategies through clinical trials. In this study, we employed spectral clustering to cluster 29,922 AD patients in the OneFlorida Data Trust using their longitudinal EHR data of diagnosis and conditions into four subtypes. These subtypes exhibit different patterns of progression of other conditions prior to the first AD diagnosis. In addition, according to the results of various statistical tests, these subtypes are also significantly different with respect to demographics, mortality, and prescription medications after the AD diagnosis. This study could potentially facilitate early detection and personalized treatment of AD as well as data-driven generalizability assessment of clinical trials for AD.

Introduction

Alzheimer’s disease (AD) is a progressive neurodegenerative disorder that affects an estimated 6.2 million Americans age 65 and older in 2021. This number is likely to reach 13.8 million by 2060.¹ It is a highly heterogeneous disease that varies not only in symptoms and progression but also in the risk factors for different phenotypes. AD has been considered as associated with a wide range of risk factors including age, genetics, head injuries, vascular diseases, infections, environmental factors, and many other medical conditions such as cardiovascular disease, obesity, and diabetes.² Subtyping of AD considering temporal changes of patients’ medical conditions prior to clinical onset of AD can be useful to facilitate understanding the heterogeneity of AD and its development (e.g., what conditions or diagnoses often precede AD onset and how they are presented across what AD subpopulations) which in turn can help improve diagnosis, treatment, prognosis and disease management.

Today, the increasing availability of large-scale datasets and development of machine learning algorithms make it possible to explore AD heterogeneity in a data-driven manner. Subtyping AD patients into homogeneous groups may lead to more differentiated disease stratification that may facilitate personalized diagnosis and prognosis.³ Previously, most data-driven AD subtyping studies have focused on creating AD subtypes with neuroimages,⁴ neuropsychological data,⁵ and neuropathological data.⁶ Electronic health records (EHRs) capture a variety of important information collected in the encounters, including demographics, diagnoses, symptoms, prescriptions, lab tests, procedures, etc. The wide adoption of EHR systems enables EHR data to be a promising source for AD subtyping. However, very few studies have used EHRs with fine-grained encounter information to identify subtypes of AD.⁷ Among the handful of studies using EHR data for AD subtyping, Xu et al. identified probable AD patients in EHRs from Weill Cornell Medicine (WCM)/NewYork-Presbyterian Hospital (NYP) and then applied hierarchical clustering with variables that can predict different clinical outcomes of AD and found four subtypes of probable AD.⁷ Their subtypes were derived by the clustering algorithm using the features identified in a supervised case-control-based binary prediction model. Landi et al. first employed deep learning to derive representations for AD patients in the Mount Sinai Health System Data Warehouse, and then applied hierarchical clustering to split patients with AD into three subgroups.⁸

In this study, we used the EHRs obtained from the OneFlorida Data Trust to find subtypes of AD using 40 top medical conditions diagnosed by the physicians or self-reported by the patients before first AD diagnosis among 29,922 AD patients. Our hypothesis is that occurrences and progression of the health conditions prior to AD diagnosis are correlated with different phenotypes of AD.⁷ Spectral clustering algorithm was used to cluster the AD patients into subgroups based on the conditions before first AD diagnosis. We further extracted and analyzed the demographics, mortality outcome, and prescription information for the patient cohort and performed Chi-square tests and multinomial logistic regression to (1) evaluate the quality of the clustering results; and (2) identify clinically meaningful associations between the demographics, mortality, and the identified clusters. The contribution of this work is two-

* Equal-contribution first authors
folds: (1) we demonstrated the feasibility of using longitudinal condition information prior to first AD diagnosis to find clinically meaningful subtypes of AD; and (2) we identified patients with certain demographics that are more likely to be clustered into certain subtypes.

Materials and Methods

Data source and cohort selection

The data used in this study were acquired from the OneFlorida Data Trust, a centralized data repository for the OneFlorida Clinical Research Consortium. OneFlorida is one of the 9 clinical data research networks in the United States funded by the Patient-Centered Outcomes Research Institute that constitute the National Patient-Centered Clinical Research Network (PCORnet).10 The OneFlorida Data Trust, which follows the PCORnet Common Data Model, contains fine-grained and longitudinal encounter information along with diagnoses, medications, lab tests, procedures, etc. for over 15 million Floridians. We defined the patient cohort as those who were diagnosed with AD between January 2012 and January 2021 in the OneFlorida EHRs. A total of 122,669 AD patients were identified with International Classification of Disease Ninth Revision (ICD-9) code 331.0 and ICD-10 CM codes of G30, G300, G30.0, G301, G308, G30.8, G309, G30.9.

The objective of this study is to identify the subgroups of AD patients using conditions diagnosed or self-reported across 6 consecutive timeslots of 6 months each before the first diagnosis of AD. After AD patients were extracted from the OneFlorida database, a comprehensive process, as shown in Figure 1, was used to select the AD patient cohorts. We selected the patient cohort with the following selection criteria: (1) a patient should be at least 20 years old on the first AD diagnosis date, (2) has diagnosis or condition records pertaining to the top 40 prevalent conditions (using the phecode11) in the cohort in any of the 6 timeslots prior to the first AD diagnosis date.

Phenotype Identification

We used a product of the Phenome-wide association studies (PheWAS) called phecode11 to determine the other conditions that the AD patients often also had.12 Phecodes were introduced to present the most frequent health conditions in EHR data. Phecode groups relevant ICD-9 and ICD-10-CM codes into clinically meaningful phenotypes, thereby enabling us to leverage accumulated ICD-9 and ICD-10-CM data for PheWAS in the EHR.13 In the OneFlorida Data Trust, ICD-9 and ICD-10 CM codes were used for records of diagnoses and self-reported health conditions. We mapped the ICD-9 and ICD-10 CM codes to phecodes and identified the most common conditions. Top 60 phenotypes were reviewed manually and the top 40 were selected as the most interested phenotypes, which are listed in Table 1.

Table 1. Top 40 phecodes and phenotypes

<table>
<thead>
<tr>
<th>Phecode</th>
<th>Phenotype</th>
<th>Phecode</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>290.1</td>
<td>Dementias</td>
<td>480</td>
<td>Pneumonia</td>
</tr>
<tr>
<td>401.1</td>
<td>Essential hypertension</td>
<td>740.9</td>
<td>Osteoarthritis NOS</td>
</tr>
<tr>
<td>272.1</td>
<td>Hyperlipidemia</td>
<td>244.4</td>
<td>Hypothyroidism NOS</td>
</tr>
<tr>
<td>591</td>
<td>Urinary tract infection</td>
<td>496</td>
<td>Chronic airway obstruction</td>
</tr>
<tr>
<td>798</td>
<td>Malaise and fatigue</td>
<td>348.8</td>
<td>Encephalopathy, not elsewhere classified</td>
</tr>
<tr>
<td>285</td>
<td>Other anemias</td>
<td>428.1</td>
<td>Congestive heart failure (CHF) NOS</td>
</tr>
<tr>
<td>292.4</td>
<td>Altered mental status</td>
<td>563</td>
<td>Constipation</td>
</tr>
<tr>
<td>745</td>
<td>Pain in joint</td>
<td>427.21</td>
<td>Atrial fibrillation</td>
</tr>
<tr>
<td>250.2</td>
<td>Type 2 diabetes</td>
<td>760</td>
<td>Back pain</td>
</tr>
<tr>
<td>411.4</td>
<td>Coronary atherosclerosis</td>
<td>433.31</td>
<td>Transient cerebral ischemia</td>
</tr>
<tr>
<td>530.11</td>
<td>GERD</td>
<td>038</td>
<td>Septicemia</td>
</tr>
<tr>
<td>532</td>
<td>Dysphagia</td>
<td>741.3</td>
<td>Difficulty in walking</td>
</tr>
<tr>
<td>296.22</td>
<td>Major depressive disorder</td>
<td>276.14</td>
<td>Hypopotassemia</td>
</tr>
<tr>
<td>418</td>
<td>Nonspecific chest pain</td>
<td>509.1</td>
<td>Respiratory failure</td>
</tr>
<tr>
<td>785</td>
<td>Abdominal pain</td>
<td>585.3</td>
<td>Chronic renal failure [CKD]</td>
</tr>
</tbody>
</table>
To create the temporal condition information for patients, we investigated the top conditions in six consecutive 6-month timeslots in the 3 years before the first AD diagnosis for each patient, as illustrated in Figure 2. Here, Timeslot 1 is from six months before the first AD diagnosis to the first AD diagnosis; Timeslot 2 is from 12 months to seven months prior to the first AD diagnosis, so on so forth. As long as a patient was diagnosed with or self-reported a condition in a particular timeslot, we considered this patient has this condition in that timeslot.

Figure 2. Timeslots defined for creating the temporal condition information for patients

**Clustering analysis**

Clustering is a fundamental machine learning technique that uses clustering algorithms to exploit underlying structure in the data and group the data points with similar characteristics into clusters. There is a wide range of clustering algorithms. The most popular ones include K-Means, latent class analysis (LCA), and Spectral Clustering. K-Means is a widely used centroid-based clustering algorithm that groups data points into clusters by minimizing the squared distances of points to the centroid in each cluster. LCA is a popular model-based clustering approach that discovers the unobserved (or latent) classes based on the distributions of observed variables in the data with iterative, maximum likelihood method. Despite their popularity, K-Means can only separate clusters that are linearly separable in the original space, while LCA requires the assumption that the observed variables within the data are conditionally independent of each other. In recent years, spectral clustering has emerged as another popular clustering algorithm providing better performance than other algorithms in many cases, especially when the instances cannot be linearly separated in the original space. Therefore, we selected spectral clustering for this study.

Spectral Clustering is a technique originated from graph theory, but is often used on both graph data and other data. It uses the top eigenvectors of an affinity matrix derived from the data to partition the data points into clusters. In this study, we used the scikit-learn Python package for the implementation of spectral clustering. The affinity matrix was constructed from the Hamming distances between data points using a Laplacian kernel. K-Means was used as the strategy for assigning labels in the Laplacian embedding space.

To investigate the consistency of grouping the cohorts into different clusters, we organized the top other conditions in aggregate and temporal formats. In the aggregate format, each patient was represented by a vector of the top 40 conditions and temporal conditions, while LCA requires the assumption that the observed variables within the data are conditionally independent of each other. In recent years, spectral clustering has emerged as another popular clustering algorithm providing better performance than other algorithms in many cases, especially when the instances cannot be linearly separated in the original space. Therefore, we selected spectral clustering for this study.
demographics, (3) mortality outcome, and (4) prescriptions after the first AD diagnosis. We also used multinomial logistic regression to model the nominal clustering results using demographic features as independent variables.

Prescription after the first diagnosis of AD can provide useful information about the treatment patterns characterizing the subgroups of AD patients. We extracted the prescription medication information after the first AD diagnosis which was available for 2,012 patients in the cohort. Prescription medication information in the OneFlorida Data Trust is coded with RxCUIs in RxNORM. To focus on the drug types instead of drugs along with dosage and dose form information, we mapped RxCUIs to Level 3 concepts of the Anatomical Therapeutic Chemical (ATC) Classification System, a drug classification system that classifies the active ingredients of drugs in five different levels where Level 3 concepts are often used to identify chemical, pharmaceutical or therapeutic subgroup.19 We selected 13 most frequently prescribed drug types in terms of Level 3 ATC concepts that are relevant to AD patients and analyzed the drug types for each cluster.

Results

Characteristics of the Cohort

After the careful selection, a total of 29,922 AD patients were included in this study. The demographic information of the patient cohort in terms of sex, race and age on the first AD diagnosis date is listed in Table 2.

Table 2. Demographics of AD patient cohorts for clustering

<table>
<thead>
<tr>
<th>Total</th>
<th>Patients</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>20,951</td>
<td>70.0%</td>
</tr>
<tr>
<td>Male</td>
<td>8,971</td>
<td>30.0%</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Indian or Alaska Native</td>
<td>36</td>
<td>0.1%</td>
</tr>
<tr>
<td>Asian</td>
<td>368</td>
<td>1.2%</td>
</tr>
<tr>
<td>Black or African American</td>
<td>4,599</td>
<td>15.4%</td>
</tr>
<tr>
<td>Native Hawaiian or Other Pacific Islander</td>
<td>7</td>
<td>0.0%</td>
</tr>
<tr>
<td>White</td>
<td>13,065</td>
<td>43.7%</td>
</tr>
<tr>
<td>Multiple Race</td>
<td>208</td>
<td>0.7%</td>
</tr>
<tr>
<td>Refuse to Answer</td>
<td>10</td>
<td>0.0%</td>
</tr>
<tr>
<td>Other</td>
<td>8,462</td>
<td>28.3%</td>
</tr>
<tr>
<td>Unknown</td>
<td>3,167</td>
<td>10.6%</td>
</tr>
<tr>
<td>Age on 1st AD Diagnosis Date (Years)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 65</td>
<td>2,767</td>
<td>9.2%</td>
</tr>
<tr>
<td>65-75</td>
<td>5,201</td>
<td>17.4%</td>
</tr>
<tr>
<td>75-85</td>
<td>10,249</td>
<td>34.3%</td>
</tr>
<tr>
<td>&gt;= 85</td>
<td>11,705</td>
<td>39.1%</td>
</tr>
</tbody>
</table>

Clustering results

We performed clustering analyses using the top 40 conditions prior to the first AD diagnosis in both the aggregate and temporal formats. We labelled the clusters resulting from the aggregate format as Cluster A, B, C, D, and the clusters resulting from the temporal format as Cluster 0, 1, 2, 3. Clustering with aggregate conditions clustered patients into cluster A, B, C, D with 10141, 5910, 6599, 7272 patients, respectively. Clustering with temporal conditions clustered patients into Cluster 0, 1, 2, 3 with 5107, 13620, 6005, 5190 patients, respectively. Clustering with temporal conditions clustered patients into Cluster 0, 1, 2, 3 with 5107, 13620, 6005, 5190 patients, respectively. Cluster A (N=10,141) overlapped mostly with Cluster

Figure 3. Percentage of patients having certain condition (top 20) in total patients for each cluster by clustering with aggregate conditions (n=10141, 5910, 6599, 7272 for cluster A, B, C, D)
To examine the characteristics of clusters with respect to conditions prior to the first AD diagnosis, we compared the percentage of patients with a certain condition in each cluster. Regarding the clustering result with aggregate conditions, as shown in Figure 3, Cluster A can be characterized as more patients diagnosed with essential hypertension and dementia but fewer patients with other conditions. Compared to Cluster A, Cluster B, C, and D had more patients with multiple conditions. Although Cluster B, C and D had more patients diagnosed with essential hypertension, they differed in other conditions with more patients. For example, Cluster B and C had more patients with dementia compared to Cluster D, while Cluster B and D had more patients with hyperlipidemia compared to cluster C. Cluster B and C had more patients with malaise, urinary tract infection, anemias, altered mental status than other two clusters.

For clustering with temporal conditions, we compared the percentage of patients with a certain condition in each timeslot for each cluster. As shown in Figure 4, the patients in Cluster 0 had a long history of essential hypertension and type 2 diabetes but only had dementia close to the first AD diagnosis. Cluster 1 and 2 had more patients with dementia only in the past 6 months prior to the first AD diagnosis. Cluster 2 and 3 had more patients with other conditions close to first AD diagnosis compared to other clusters. Cluster 3 had more patients with essential hypertension in the past 3 years and more patients with dementia in the past 2 years prior to the first AD diagnosis. A higher percentage of patients in Cluster 3 has major depressive disorder.

Figure 4. Percent of patients with certain condition (top 20) in total patients having diagnoses in each timeslot with temporal conditions in (a) Cluster 0 (N=5,107), (b) Cluster 1 (N=13,620), (c) Cluster 2 (N=6,005), and (d) Cluster 3 (N=5,190).

Comparing the results of clustering with aggregate vs. temporal conditions, we found that clustering with temporal conditions was more capable of creating subgroups that are clinically interpretable. Hence, we examined the distinct characteristics of each cluster by clustering with temporal conditions in the following subsections.

Demographics characteristics and mortality outcome

Figure 5 shows the characteristics of the clusters with respect to sex, race and age group on the first AD diagnosis date, as well as the mortality outcome for each cluster by clustering with temporal conditions. Cluster 3 had the highest percentage of female (N=3,815, 73.5%) and lowest percentage of male (N=1,375, 26.5%) while Cluster 2 had the lowest percentage of female (N=4,012, 66.8%) and highest percentage of male (N=1,993, 33.2%). Regarding race, Cluster 0 had the highest percentage of Black or African American (N=100, 20.0%) and lowest percentage of White (N=1,623, 31.8%), while Cluster 3 had the highest percentage of White (N=2,851, 54.9%) and lowest percentage of
Black or African American (N=687, 13.2%). Regarding age groups, cluster 0 had the highest percentage of patients in group <65 (N=605, 11.9%) and lowest percentage of patients in group >=85 (N=1,420, 27.8%) while Cluster 2 had the lowest percentage of patients in group <65 (N=388, 6.5%) and highest percentage of patients in group >=85 (N=2,553, 42.5%). According to Chi-square tests, clusters were statistically significantly different by sex (P=3.92E-15), race (P=8.76E-166) and age groups (P=1.14E-99). Regarding mortality outcome, Cluster 2 had the highest mortality rate (24.7%), followed by Cluster 3 (23.0%), Cluster 0 (17.7%), and Cluster 1(17.6%).

![Figure 5](image-url)

**Figure 5.** Number of patients in each cluster resulting from clustering with temporal conditions stratified by (a) sex: F-female, M-male; (b) race: 01-American Indian or Alaska Native, 02-Asian, 03-Black or African American, 04-Native Hawaiian or Other Pacific Islander, 05-White, 06-Multiple Race, 07-Refuse to Answer, OT-Other, UN-Unknown; (c) age groups on 1st AD diagnosis date; (d) mortality: No-did not die, Yes-died. (n=5107, 13620, 6005, 5190 for cluster 0, 1, 2, 3)

To examine the difference in each pair of clusters, Chi-square tests were conducted for each demographic variable and the mortality outcome. Table 3 shows the p-values of these tests. At the significance level of 0.05, all the pairs of clusters were statistically significantly different in sex except for the pair of Cluster 0 and 2 (P=0.099). All the pairs of clusters were statistically significantly different in race. The percentages of Asian and White in each pair of clusters were significantly different except for the pair of Cluster 1 and 2 (P=0.530 and 1.000), while the percentage of Black or African American in each pair of clusters were significantly different except for the pair of Cluster 1 and 3 (P=0.944). With respect to age, all the pairs of clusters were significantly different, while certain age groups were not significantly different among some pairs. For mortality, all the pairs of clusters with patients who died vs. did not die were significantly different except for the pair of Cluster 0 and 1. To ascertain the statistical importance of the subtypes, we also applied a multiple hypothesis correction method Bonferroni correction to adjust the significance...
level by dividing 0.05 by 15 (number of hypotheses for each pair of clusters). As such, the differences pertaining to black or African American for Cluster 0 vs. Cluster 2, multiple races for Cluster 1 vs. Cluster 3, and age group 75-85 for Cluster 0 vs. Cluster 2 (the p-values in italic) became insignificant with the new significance level of 0.003. The other conclusions made with the significance level of 0.05 remained the same.

Table 3. P-values of pairwise and all-clusters Chi-square tests for sex, race, age on 1st AD diagnosis date, and mortality outcome by clusters with temporal conditions.

<table>
<thead>
<tr>
<th></th>
<th>All-clusters</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 vs. 1</td>
<td>0 vs. 2</td>
<td>0 vs. 3</td>
<td>1 vs. 2</td>
<td>1 vs. 3</td>
<td>2 vs. 3</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>#</td>
<td>#</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Indian or Alaska Native</td>
<td>0.536</td>
<td>0.053</td>
<td>0.520</td>
<td>0.077</td>
<td>1.000</td>
<td>0.288</td>
<td>0.086</td>
</tr>
<tr>
<td>Asian</td>
<td>#</td>
<td>#</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black or African American</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native Hawaiian or Other Pacific Islander</td>
<td>0.901</td>
<td>0.935</td>
<td>1.000</td>
<td>0.317</td>
<td>0.887</td>
<td>0.942</td>
<td>0.474</td>
</tr>
<tr>
<td>White</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Races</td>
<td>0.428</td>
<td>0.082</td>
<td>0.003</td>
<td>0.223</td>
<td>0.008</td>
<td>0.190</td>
<td>0.010</td>
</tr>
<tr>
<td>Refuse to Answer</td>
<td>0.728</td>
<td>0.935</td>
<td>0.628</td>
<td>0.237</td>
<td>0.990</td>
<td>0.199</td>
<td>0.298</td>
</tr>
<tr>
<td>Other</td>
<td>0.114</td>
<td>0.895</td>
<td>0.137</td>
<td>#</td>
<td>#</td>
<td>#</td>
<td>#</td>
</tr>
<tr>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age groups (age on 1st AD diagnosis date, years)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 65</td>
<td>#</td>
<td></td>
<td>#</td>
<td>#</td>
<td>#</td>
<td>#</td>
<td>#</td>
</tr>
<tr>
<td>65-75</td>
<td></td>
<td>#</td>
<td></td>
<td>#</td>
<td>0.305</td>
<td>0.002</td>
<td>0.071</td>
</tr>
<tr>
<td>75-85</td>
<td></td>
<td></td>
<td></td>
<td>#</td>
<td>0.009</td>
<td>0.911</td>
<td>#</td>
</tr>
<tr>
<td>&gt;= 85</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>#</td>
<td>0.039</td>
<td>0.432</td>
</tr>
<tr>
<td>Mortality</td>
<td>0.802</td>
<td>#</td>
<td></td>
<td>#</td>
<td>#</td>
<td>#</td>
<td>0.034</td>
</tr>
</tbody>
</table>

Note: #: p-values <= 0.001. n=5107, 13620, 6005, 5190 for cluster 0, 1, 2, 3.

Prescription

Prescription information is only available for 2,012 patients in this cohort. Figure 6 shows the percent of patients prescribed with selected drugs in each cluster by clustering with temporal conditions. All the clusters had a higher percentage of patients prescribed with other analgesics and antipyretics than other medications. Cluster 2 and 3 had higher percentage of patients prescribed with drugs for constipation. Cluster 0 had the highest percentage of patients prescribed with drugs for peptic ulcer and GORD whereas Cluster 2 and 2 had higher percentage of patients prescribed with insulin and analogues. Note that some medications (e.g., drugs for constipation, GORD) could have been prescribed due to side effects of medications for other conditions such as diabetes and hypertension.

Figure 6. Percent of patients having prescription of selected drugs in total patients having prescription in each cluster by clustering with temporal conditions (n= 289, 1094, 363, 266 for cluster 0, 1, 2, 3)

Multinomial logistic regression

To further evaluate the clustering results and identifying the factors that cause patients to be more likely in a certain cluster, we conducted multinomial logistic regression for assigning each patient into different clusters using demographic features as independent variables (predictors). We used Cluster 0 as the reference group. According to
the results in Table 4, male patients were more likely to be in Cluster 2 (Relative Risk Ratio [RRR]=1.237, p<0.01) but less likely to be in Cluster 3 (RRR=0.858, p<0.01). In terms of race, Asian patients had lower probability of being in Cluster 1 (RRR=0.326, p<0.1), Cluster 2 (RRR=0.137, p<0.01), and Cluster 3 (RRR=0.152, p<0.05). Black or African Americans and Native Hawaiian or Other Pacific Islanders were less likely to be in Cluster 2 (RRR=0.240, p<0.05; RRR=0.00002, p<0.01), whereas patients of multiple races had lower probability of being assigned into Cluster 2 (RRR=0.162, p<0.01) and Cluster 3 (RRR=0.210, p<0.05). With respect to age, patients of age 65-75 had higher chance of being in Cluster 2 (RRR=1.403, p<0.01) but had lower chance of being in Cluster 3 (RRR=0.864, p<0.05), patients of age 75-85 had higher chance of being in Cluster 2 (RRR=1.319, p<0.01) and Cluster 3 (RRR=1.632, p<0.01), and Cluster 2 (RRR=3.059, p<0.01) in particular.

### Table 4. Results of multinomial logistic regression models for predicting patients to be in different clusters by clustering with temporal conditions

<table>
<thead>
<tr>
<th>Demographic Characteristics</th>
<th>Cluster 1 n=13,620</th>
<th>Cluster 2 n=6,005</th>
<th>Cluster 3 n=5,190</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Relative Risk Ratio (RRR) (Robust SE)</td>
<td>Relative Risk Ratio (RRR) (Robust SE)</td>
<td>Relative Risk Ratio (RRR) (Robust SE)</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1.003 (0.036)</td>
<td>1.237*** (0.042)</td>
<td>0.858*** (0.045)</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>0.326* (0.651)</td>
<td>0.137*** (0.663)</td>
<td>0.152** (0.737)</td>
</tr>
<tr>
<td>Black or African American</td>
<td>0.369 (0.639)</td>
<td>0.240*** (0.645)</td>
<td>0.321 (0.710)</td>
</tr>
<tr>
<td>Native Hawaiian or Other Pacific Islander</td>
<td>1.048 (1.270)</td>
<td>0.00002*** (0.001)</td>
<td>0.526 (1.584)</td>
</tr>
<tr>
<td>White</td>
<td>0.738 (0.639)</td>
<td>0.354 (0.644)</td>
<td>0.809 (0.709)</td>
</tr>
<tr>
<td>Multiple Race</td>
<td>0.445 (0.663)</td>
<td>0.162*** (0.681)</td>
<td>0.210** (0.756)</td>
</tr>
<tr>
<td>Refuse to Answer</td>
<td>1.384 (1.257)</td>
<td>0.00004*** (0.001)</td>
<td>1.476 (1.356)</td>
</tr>
<tr>
<td>Other</td>
<td>0.49 (0.639)</td>
<td>0.239*** (0.644)</td>
<td>0.331 (0.710)</td>
</tr>
<tr>
<td>Unknown</td>
<td>0.438 (0.640)</td>
<td>0.110*** (0.646)</td>
<td>0.286* (0.711)</td>
</tr>
<tr>
<td><strong>Age Group</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65-75</td>
<td>0.952 (0.062)</td>
<td>1.403*** (0.079)</td>
<td>0.864*** (0.074)</td>
</tr>
<tr>
<td>75-85</td>
<td>1.319*** (0.058)</td>
<td>1.912*** (0.073)</td>
<td>0.926 (0.069)</td>
</tr>
<tr>
<td>85+</td>
<td>2.027*** (0.059)</td>
<td>3.059*** (0.075)</td>
<td>1.632*** (0.069)</td>
</tr>
<tr>
<td>Constant</td>
<td>3.578** (0.640)</td>
<td>2.159 (0.647)</td>
<td>2.014 (0.711)</td>
</tr>
<tr>
<td>AIC</td>
<td>75,632.55</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: *p<0.1, **p<0.05, ***p<0.01. Cluster #0 serves as the reference group (n=5,107). ‘Reference is Female. ‘Reference is American Indian or Alaska Native. ‘Reference is Age Group <65.

### Discussion

Identifying meaningful subtypes of AD can support the development of prevention and treatment strategies for AD by informing the design of clinical trials as well as the analysis of their generalizability, which is often overlooked. An intervention can be tested in different subtypes, for which efficacy and safety of the intervention can be evaluated separately. With state-of-the-art trial generalizability analysis methods, one can also assess the generalizability of a trial using different target populations based on the subtypes, making it easier to identify the subgroups to which the results are mostly applicable.
Based on our hypothesis that medical conditions prior to AD diagnosis are associated with different subtypes of AD, we performed spectral clustering to cluster the AD cohort of 29,922 patients using their physician-diagnosed or self-reported conditions in 6 consecutive half-year periods (timeslots) before the first diagnosis of AD from EHRs and identified four subtypes of AD (subtypes 0, 1, 2, and 3). Chi-square tests indicated that the subtypes found in this study were statistically significantly different with respect to demographics, mortality, and prescriptions after AD diagnosis.

Among the 4 derived subtypes of AD, Subtype 0 (Cluster 0) can be characterized as having a long history of essential hypertension, type 2 diabetes and hyperlipidemia, short history of dementia, and fewer other conditions before the diagnosis of AD. Subtype 0 has a higher percentage of Black or African Americans and younger patients (age<75), lower mortality rate, and a higher percentage of patients prescribed with drugs for peptic ulcer and GORD and insulins. Subtype 1 can be characterized as having long history of essential hypertension, short history of dementia, and few other conditions before first AD diagnosis. Subtype 1 (Cluster 1) has a higher percentage of White, lower mortality rate, and higher percentage of patients prescribed with corticosteroids, and anti-dementia drugs. Subtype 2 (Cluster 2) can be described as having a long history of essential hypertension, short history of dementia, and multiple other conditions close to the diagnosis of AD, a lower percentage of female, a higher percentage of White and older patients (age>=75), higher mortality rate, and a higher percentage of patients prescribed with drugs for constipation, and insulins and analogues. Subtype 3 (Cluster 3) has a long history of essential hypertension and diverse history of multiple other conditions before first AD diagnosis, a higher percentage of female and White, lower mortality rate, and a higher percentage of patients prescribed with drugs for constipation, hypnotics and sedatives, corticosteroids, and antiepileptics. As this is the first study on the temporal subtyping of Alzheimer’s disease using the conditions before the first AD diagnosis in EHR, the results are not directly comparable to other studies that used EHR data for AD subtyping. Nonetheless, some clusters do share similar characteristics with previous studies. For example, patients in Subtype 2 are older (mean age: 81.9) and have more comorbidities, similar to Subphenotype C in Xu et al.

There are a few limitations of this study. First, the cohort from the OneFlorida EHRs has been greatly narrowed down for our analysis (122,669->29,922) due to our focus on temporal changes of one’s medical conditions in the past 3 years prior to the first AD diagnosis. Therefore, it may not be representative of the general AD population. Carrying out these analyses in other AD data sets would speak to generalizability of these results. Nonetheless, as Florida is one of the most populous states for older adults, OneFlorida Data Trust is still one of the largest EHR datasets available for such a study. Second, the information about medical conditions prior to the first AD diagnosis may be incomplete and OneFlorida contributing sites may use different diagnostic criteria and coding conventions, hence the clusters based on time interval analyses may lack sensitivity for identifying other precursors of AD. Third, OneFlorida Data Trust does not have AD progression information such as Mini-Mental State Exam (MMSE), neuroimages, or neuropathological, biochemical biomarkers which can also be used for subtyping. Nonetheless, the results of this study have demonstrated the potential of using longitudinal EHR data for AD subtyping. One could also evaluate the hazard ratio for mortality for different subtypes when recruiting patients into clinical trials. Aside from the information used in this study, EHRs have rich additional information contained in unstructured text of clinical notes. Extracting information from the clinical notes with state-of-the-art natural language processing (NLP) techniques will potentially facilitate the use of EHR data for AD subtyping. In addition, deep learning has achieved great success in many applications in computer vision and NLP with multiple data modalities. Given the diversity of data in the medical domain, promising achievements can be expected for further research of multi-modal subtyping and deep phenotyping using EHR data coupled with neuroimages, neuropsychological data, and neuropathological, clinical and biochemical biomarkers.

**Conclusion**

In this study, we performed spectral clustering of AD patients using the longitudinal medical condition information before the AD diagnosis. Our analysis using OneFlorida Data Trust indicates that subtypes of AD created by this technique are significantly different in terms of the conditions, demographics, mortality, and prescription medications. Future research is warranted to further evaluate the clinically meaningfulness of these clusters. This work could facilitate not only early detection and personalized treatment of AD but also data-driven generalizability assessment of clinical trials for AD.

**Acknowledgements**

This study was partially supported by the National Institute on Aging under Award Number R21AG061431; and in part by Florida State University-University of Florida Clinical and Translational Science Award funded by National Center for Advancing Translational Sciences under Award Number UL1TR001427. The content is solely the
responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Applying FHIR Genomics for Research – From Sequencing to Database

Sean Hernandez, MD1; Karen Fairchild2,3; Mark Pemberton, MS4; Jonathan Dahmer2; Wei Zhang PhD1; Matvey B. Palchuk, MD, MS2; Umit Topaloglu, PhD1

1Wake Forest School of Medicine, Winston-Salem, NC, 2TriNetX, LLC, Cambridge, MA, 3HL7 Clinical Genomics Working Group, Ann Arbor, MI, 4Information Technology Services, Atrium Health Wake Forest Baptist, Winston-Salem, NC

Abstract

The availability of next-generation sequencing (NGS) technologies and their continually declining costs have resulted in the accumulation of large genomic data sets. NGS results have traditionally been delivered in PDF format, and in some cases, structured data, e.g., XML or JSON formats, are also made available, but there is a lack of uniformity around the profiling of external vendor testing platforms. Atrium Health Wake Forest Baptist and TriNetX have harmonized and mapped genomic data to FHIR Genomic standards and imported it into the TriNetX database through a data pipeline. This process is translatable to other sequencing platforms and to other institutions. The addition of genotypic data to the TriNetX database to the reservoir of phenotypic data will promote (i) enhanced industry trial recruitment, (ii) comprehensive intra-institutional genomic benchmarking/quality improvement, and eventually (iii) sweeping inter-institutional genomic research and treatment paradigm shifts.

Introduction

The advancements in next generation sequencing (NGS) technologies and their continually declining costs have resulted in the accumulation of very large sets of genetic data and facilitated the identification of actionable genetic alterations in different tumor types. Correspondingly, the field of cancer genomics has rapidly grown with the availability of NGS and the prompt development of clinical applications in cancer treatment and research. Precision Oncology, which aims to identify targetable alterations, has steadily built the infrastructure for individual variants often dictating treatment options in conjunction with anatomical site-specific treatment1–3.

However, a barrier to utilization of genomic data frequently has been data representation4. Individualized genomic variants have traditionally been delivered by NGS services in a Portable Document Format (PDF). This PDF would then be scanned or uploaded into the electronic health record, ultimately limiting utilization by treating physicians. As structured data (e.g., XML or JSON formats) became part of the package delivered to ordering institutions along with the PDF, this eliminated one of the pain points. Another pain point for interoperability and data collaborations is the differences in data formatting5,6. While PDF formatted results are needed for clinical care, structured data has become a rich resource for research purposes. But, there is a lack of uniformity around the profiling of external vendor testing platforms as well as evolving interoperability standards7. Due to a lack of standardization in structured genomic results, there is an urgent need for developing consistent mechanisms for ingesting, storing, organizing, and accessing the data generated. Ultimately, the issues surrounding data discovery and utilization contribute to a substantial gap in research potential for both hypothesis testing, hypothesis generation, and even patient identification for clinical trials.

Organizations approach this interoperability challenge in varying ways. For example, Rush University worked with its EHR company, Epic, and one of its NGS services, Tempus, to create an interface between the EHR and NGS results. Using Epic’s genomic module, structured data elements were created to directly file the genomic variants in the reports provided by Tempus utilizing Health Level 7 (HL7) v28.

HL7, which had developed Fast Healthcare Interoperability Resources (FHIR)® as a means of standardizing data formats, also recognized the need for shared data formatting in the field of Genomics and commissioned the HL7 Clinical Genomics Workgroup (CG WG) to mature a similar model known as the FHIR Genomics Reporting Implementation Guide®10–11.
Building on this framework, the Electronic Medical Records and Genomics (eMERGE) Network comprising 11 institutions explored standardizing their genomic data to FHIR genomics in order to create a pipeline between sequencing company, data reservoir and data import into the EHR\textsuperscript{11}. They were successfully able to channel genomic data from all participating institutions to harmonize and map the data into an FHIR database and return the information to their respective EHR's\textsuperscript{12}.

Furthermore, the American Association of Cancer Research (AACR) project Genomics Evidence Neoplasia Information Exchange (GENIE), which has successfully implemented an international consortium, pooled clinical-grade outcomes and genomics data from more than a hundred thousand patients from US and international cancer centers in efforts to create a publicly available cancer data reservoir. Project GENIE participating institutions agreed on core data elements and data definitions\textsuperscript{13}. Data on mutations, demographics and outcomes were fed into the “Synapse platform,” where data elements were harmonized and de-identified in a HIPAA compliant manner.

Most of these approaches to the interoperability challenges (i) have been multi-institutional central repositories, (ii) utilize local/single vendor next-generation sequencing services, and (iii) primarily focus on direct patient care. To the authors’ knowledge, no effort has been published documenting the development of a research-directed FHIR Genomics harmonization using a commercial NGS service. Our study demonstrates the successful implementation of a data pipeline with TriNetX built for research purposes with goals of broadening to multiple other NGS platforms.

TriNetX has successfully developed a clinical research collaboration platform built on a federated network of healthcare organizations (HCOs), pharmaceutical firms, and contract research organizations (CROs). Their efforts have been dedicated towards the development of federated data models to facilitate data-driven clinical research with a focus on decreasing clinical trial accrual failure. This international network currently includes more than 74 million patients (mostly within the US) and 55 HCO’s\textsuperscript{14}. The TriNetX business model has successfully relied on sponsorship from pharmaceutical firms and CROs who pay subscription fees to query the de-identified database for aggregate counts grouped by HCO. This financial infrastructure has supported HCO participation at no cost, with HCOs obtaining access to the TriNetX data visualization software and participating in the selection pool for targeted industry trials.

The TriNetX platform supports genotypic data, including genomic variants, integrating it seamlessly with the phenotypic information about patients drawn primarily from EHR’s. This manuscript details the development of a data pipeline using FHIR Genomic standards and connecting commercially derived genomic information at Atrium Health Wake Forest Baptist into the TriNetX database.

Methods

At Atrium Health Wake Forest Baptist (AHWFB) flow of Genomic information from NGS vendors involve multiple files and data types (Figure 1). Our major NGS vendors are Foundation Medicine, Inc. (Cambridge, MA), Guardant Health, Inc. (Redwood City, CA), and Caris Life Sciences (Irving, TX). We have established Data Use Agreements (DUA) with respective vendors that allow us to use the data for research purposes.

The vendors provide (i) PDF result files for clinical use, (ii) Extensible Markup Language (XML) or JavaScript Object Notation (JSON) for structured data including variant details, and (iii) Binary Alignment Mapping (BAM) and FASTQ (a text-based format for storing biological sequence and its corresponding quality score). The PDFs are uploaded to our EHR, EPIC utilizing OnBase (Hyland Software, Westlake, OH) document management system. BAM and FASTQ files are stored in our local Storage Area Network. Finally, JSON and XML files are parsed and stored in a SQL Server database.
Genomic reporting is typically done on the basis of differences between the sequence observed in the tested specimen and the sequence in a reference sequence. Such differences are called variations. They are not usually reported as a complete enumeration of the whole sequence of interest. This approach is also used because the clinical relevance of genomic tests is based on the presence of a divergence from the norm. Note that a sequence might still be relevant even if it is unchanged from the reference.

Given the wide variety of sources of genomic data, the FHIR standard has again been regarded for standardization and interoperability. However, the challenge of exchanging FHIR genomic data is the lack of normative resources handling this type of data. Furthermore, the current FHIR MolecularSequence resource includes many attributes that are under evaluation for inclusion or exclusion in future advancements of the standard. Currently, the HL7 Clinical Genomics Working Group is modeling molecular sequencing to address these obstacles.

Although genomic interoperability standards are yet to mature, the dramatic increase in molecular diagnostic testing involving sequencing is fueling the standards work. The MolecularSequence resource is in its infancy at Maturity Level 1 for Trial Use. The Variant Profile is also currently in its infancy and under review for advancement by the HL7 Clinical Genomics Working Group.

In our use case, we wanted to be able to send select data elements from AHWFB outside labs such as Foundation Medicine to TriNetX. We did not need to send entire sequences, only select information about any variants that were detected by the third-party labs. In evaluating options for using FHIR to move molecular diagnostic test results data between AHWFB and TriNetX for population-based research, our team reviewed some options with current R4 versions of FHIR. For these molecular diagnostic test results, we looked at the MolecularSequence Resource and the Observation Variant Profile. The data elements included in the MolecularSequence Resource (see Figure 2 - Molecular Sequencing Resource model) were very complete. Many of the attributes were not needed, nor were the data available to us. We needed variant descriptions and reference sequence information which is supported in the MolecularSequence Resource. However, the raw sequencing data and other overhead used for encoding the sequence block were not needed. Therefore we considered the Observation Resource with the (genomic) Variant Profile (see Figure 3 – Variant Profile).
The Variant Profile is geared towards reporting only the variant portion of the sequence as opposed to entire sequence blocks. This met the needs of our use case, to be able to extract genomic report data into a FHIR message to send to TriNetX for population-based cohort identification and research. The components in the variant profile contain attributes for single-nucleotide variant (SNV) type variants which can be expressed in Human Genome Variation Society (HGVS) notation, as well as complex structural variants with support for International System for Human Cytogenomic Nomenclature (ISCN) nomenclature support.

Figure 2. HL7 FHIR MolecularSequence Resource model

Figure 3. Variant Profile (http://hl7.org/fhir/uv/genomics-reporting/STU1/variant.html)
Using the Observation Resource with the Variant Profile, our codeable concepts leverage Logical Observation Identifiers Names and Codes (LOINC)® codes from the “Master HL7 genetic variant reporting panel” [81247-9], which is in Trial status. We were able to use LOINC codes for (i) SNV type variants with Human Genome Variation Society (HGVS) notation support; (ii) structural variants with LOINC codes for International Sustainable Campus Network (ISCN) nomenclature and copy numbers; and (iii) structural variant ranges covering the type of molecular diagnostic test results from the three third-party molecular labs that are part of this project.

Since the FHIR server AHWFB utilized did not yet support the FHIR R4 Bulk Data feature, we utilized FHIR Observation resource bundles that are sent to TriNetX when requested for a given patient. The data pipeline is graphically represented in Figure 5.

---

**Figure 4.** LOINC Codes used in WF Codeable Concepts (full list located at https://loinc.org/81247-9/)

**Figure 5.** The data pipeline to the TriNetX database.
Results

More than 5,400 patients in the Wake Forest Baptist Comprehensive Cancer Center have undergone cancer genomic sequencing; 2,600 of these patients used the Foundation Medicine NGS platform. The data from these genomic sequencing reports were successfully formatted to the FHIR Genomics standard and imported into the TriNetX system via the data pipeline. There are currently 36,127 individual genomic facts alongside the previously available phenotypic data.

One use case was successfully conducted in the “Staging” environment of the database (Figure 6). This data query from the AHWFB data reservoir included the following criteria: (i) diagnosis of non-small cell lung cancer and (ii) EGFR positive genomic mutation identified on Foundation Medicine NGS.

![Figure 6. This is a use-case utilizing the data visual software provided by TriNetX with phenotypic and genotypic criteria selected from our institution’s pool of patients.](image)

Discussion / Conclusion

Since molecular diagnostics are utilized as a standard of care for clinical decision making, utilization for research relies on the availability of technological and genomic approaches. In this study, we have successfully compiled genomic data sequenced by a commercial platform, formatted the data to FHIR Genomics standards, and created a pipeline to the TriNetX Database.

The TriNetX database houses phenotypic data for over 2 million patients at Atrium Health Wake Forest Baptist Health. This is the first time that genotypic and phenotypic data will be present in the TriNetX database for AHWFB. The presence of genotypic data, in addition to the phenotypic data, can help pharmaceutical firms and CROs with the recruitment process for industry trials, effectively decreasing accrual failure and study costs.

As noted previously, genomic variants are increasingly part of the selection criteria for clinical trials as well as part of the treatment algorithm for an increasing number of advanced cancers. As more genomic variants are noted to be possible treatment foci or possible prognostic markers, the emphasis on genotypic data as a selection criterion for research and clinical trials will only continue to grow. International, federated data networks such as TriNetX will be in higher demand as the field moves to closer integration of personalized and precision medicine for cancer treatments. As demonstrated through this project, genotypic data integration will be available to meet this demand.
The future for the Atrium Health Wake Forest Baptist Genomic Program will include replication of the process established for Foundation Medicine for the other participating platforms (Guardant, Caris, etc.) There are currently roughly 2800 patients remaining who have had genomic sequencing either through Guardant or Caris. Pooling of all the genotypic data grants Wake Forest a broad array of genomic variants evaluated by the different platforms. With genotypic and phenotypic data compiled through the TriNetX database, internal benchmarking can more easily take place to ensure that, as an institution, patients are receiving the best care possible. Again this mission is also aided by having genotypic data available for industry trial recruitment by pharmaceutical firms and CROs.

Furthermore, the process established by AHWFB can be replicated by other institutions to import genomic data into the TriNetX database. The benefits of compiled data from multiple institutions have been well described. Additionally, compiling genomic data can aid with benchmarking and genomic comparison among institutions to not only serve research initiatives but also as another layer of quality metrics. Benchmarking can also facilitate institutions evaluating local and national genomic diversity while controlling for available phenotypic factors.

In this project, Atrium Health Wake Forest Baptist and TriNetX were able to consume Genomic data from a commercial sequencing service, format it to FHIR Genomics standard, and import this data into the TriNetX database through a data pipeline. This process is translatable to other sequencing platforms and to other institutions. Previous studies have shown the benefits for research and clinical trials from federated data models such as TriNetX. The addition of genotypic data to the TriNetX database to the reservoir of phenotypic data will promote (i) enhanced industry trial recruitment, (ii) comprehensive intra-institutional genomic benchmarking/quality improvement, and eventually (iii) sweeping inter-institutional genomic research and treatment paradigm shifts.

**Acknowledgement**

This work was supported in part by the Bioinformatics Shared Resources under the NCI Cancer Center Support Grant to the Comprehensive Cancer Center of Wake Forest University Health Sciences (P30 CA012197) and by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (UL1TR001420). We would like to thank Michael Horvath and Wendell Futrell for their help during the implementation.

**References**


VitalSeer: The development of a contactless sensing technology based on a user-centric data-driven clinical approach

Kendall Ho, MD, FRCPC 1, David Rivest-Henault, PhD2, Nooshin Jafari, MSc1, Michael Lim, BCom1, Aida Hassani, MSc1, Jennifer L. Cordeiro, BAA(Hons)1, Bruno Falardeau, BSc2, Linda Pecora, BEng2, Catherine Pagiatakis, PhD2, Faten M’hiri, PhD2, Di Jiang, PhD2

1 University of British Columbia, Department of Emergency Medicine, Vancouver, British Columbia, Canada; 2National Research Council of Canada, Medical Devices Research Center, Boucherville, Quebec, Canada

Abstract

The COVID-19 pandemic presented challenges to the healthcare system while catalyzing the adoption of virtual care. The need for remote assessment and real-time monitoring of physiological vital signs has driven towards a need for virtual care solutions. This paper presents the outcome of a multidisciplinary collaboration to ensure clinical usability of a remote contactless sensing technology, VitalSeer, and to help close gaps between emerging technologies and clinical practice. The paper describes the user-centric data-driven clinical approach to address the needs as identified by clinical experts through the iterative and agile development cycle. It highlights findings from preliminary studies to validate proof-of-concept VitalSeer’s adoptability, accessibility and usability. The studies on volunteers demonstrated the accuracy of VitalSeer’s heart rate model at a low MAE of 0.74 (bpm) and a RMSE of 1.2 bpm, below the threshold of clinical grade contact-based sensors. The paper concludes with a discussion on the technology implications in emergency medicine and community care.

Introduction

COVID-19 has propelled the health system to change from predominantly in-person services to virtual care, where healthcare providers use technology, like video-calls, to examine and communicate with the patients. Virtual care has been used for decades; however, its adoption rate has only significantly increased since the beginning of the COVID-19 pandemic1. Virtual platforms with common and affordable information technologies such as computers, smart phones and tablets are becoming well accepted by physicians and patients, and can be helpful during emergencies and outbreaks2-4. Such remote healthcare delivery approach can also support social distancing when accessing needed medical consultations, and reduce unnecessary emergency department visits5-7.

Knowing patients’ vital signs in the course of virtual visits is critical to assess the severity of illnesses and determine whether their conditions warrant in person visits8. Yet, vital sign measurement has not been integrated into the most current routine virtual care workflows. This can limit the scope of virtual care and, ultimately, the quality of care delivered9. For example, in COVID-19 patients with silent hypoxia10, their oxygen saturation (SpO2) can be dangerously low and require hospitalization, yet they may not experience dyspnea as a subjective symptom, or exhibit obvious shortness of breath on video11. Furthermore, a recent study found that SpO2 and Respiratory Rate (RR) at the time of presentation to the hospital are predictive of the risk of patient mortality12. Therefore, during virtual care encounters with patients, clinical questioning coupled with heart rate (HR), RR, and SpO2 assessment are critical to accurately evaluate their need for in person assessment or hospitalization.

While patients can acquire home-based devices and wearables to assess their vital signs, only a small portion may have the means and self-management knowledge to use the devices. To optimize equity to access, accurate and economical methods to assess vital signs remotely need to be developed for patients who have no or limited access to home-based devices. Contactless sensing systems to measure vital signs is a promising approach13,14 that might be more convenient to deploy and use compared to existing wired or wireless alternatives.

Proposed Contactless Sensing Solution: This paper presents the employment of the user-centric data-driven clinical approach for the development, from ideation to pre-clinical proof-of-concept, of VitalSeer, a contactless sensing system solution. In virtual care, “contactless sensing” refers to processes whereby physiological signals are obtained from contactless hardware (for example: microphone, video cameras, etc.) and processed to extract vital signs or other biofeedback information. VitalSeer is based on remote photoplethysmography (rPPG), following similar fundamental principles as the contact-based PPG used in pulse oximeters. However, instead of using diodes and photosensors in
an enclosed device, rPPG is obtained using ambient light and regular, external color cameras as illustrated in Figure 1. VitalSeer software aims to conduct remote monitoring of critical vital signs including HR, RR and SpO2.

Figure 1. Illustration of the rPPG principle. a) A capture device equipped with a camera is used to capture light that is back scattered from a subject’s exposed skin subjected to an external light source, b) Details of the light scattering and reflection processes.

To support patient self-monitoring and enable decision support for clinicians, the developed software needs to be clinically relevant, adoptable, accessible, easy-to-use and low-cost. Such a system would be important in the context of delivering virtual care, and particularly in the context of the current COVID-19 pandemic. The iterative, clinically driven, user-centric development process taken to build the solution was grounded in a close collaboration between multidisciplinary teams to achieve usability and clinical pertinence, helping to close the gap between rapidly emerging technologies and clinical practice. As a demonstration of this work, the development, testing and findings reported in this paper are based on VitalSeer’s HR model.

Methods

Overview: An iterative technology design and development approach was undertaken for the development of the VitalSeer contactless sensing solution. An intertwined approach ensures that the technology being developed is well aligned with the clinical context. The approach comprised of three main steps: 1) healthcare provider interviews, solution requirements definition and design, 2) solution development and testing, and technology benchmarking, and 3) technology validation through clinical research studies. Each of these steps is described in more detail below. Feedback loops allow for iterative development, which in turn allows for the gradual refinement of the quality of the solution. Figure 2 illustrates this development process. The diagram provides a schematic view of the key processes led by both the clinical team (toward the left-hand side of the diagram) and/or the science and engineering team (toward the right-hand side). The overall methodology consists of a main development loop leading from a proof-of-concept implementation to an updated release version of the solution. Two sub-loops (steps 1 and 2 above) are also included to represent shorter-term, fast-paced iterations. To support the development, two rounds of in-lab studies were conducted. These two rounds are designated as “VitalSeer 1.0” and “VitalSeer 1.2” clinical research study through the rest of this paper. After each study, the results were fed back to the design stage to inform the development of the next version.
Healthcare provider interviews, solution requirements definition and design: Exploratory interviews with healthcare providers were conducted to better understand current practices, utility, and challenges with vital sign measurement in virtual care. The interview guide is available upon request to the corresponding author. Healthcare providers with experience providing virtual care (defined as providing care, using technology and tools remotely, or at a distance from the patient) in 2020 or before were recruited. Interviews were between 30 and 60 minutes and conducted over videoconferencing. The protocol of questions used to facilitate the interviews aimed to collect insights and challenges in providing virtual and COVID-related care, and the potential impacts, opportunities, and clinical applications of contactless sensing. Meanwhile, the multidisciplinary team initiated the solution design by constructing use cases that outlined the detailed workflows, scenarios, and conditions under which the proposed contactless sensing technology would operate. The use cases were refined with feedback from both the clinical and technology teams and the feedback collected through the interviews, with particular attention paid to the user experience. Next, the use cases were translated into a technology development document that defined and presented the specific requirements, features and modules of the solution. Table 1 in the Results section documents the outcome of this process.

Solution development, testing, and technology benchmarking: Guided by the results in Table 1, the core technical development commenced which allowed for features and components of the solution to be implemented within the framework architecture. As beta software versions were iteratively released by the technical team, the clinical and technology teams conducted testing (see clinical lab testing loop in Figure 2). These tests systematically identified issues/bugs, test system performance, stability, accuracy, robustness and usability prior to the release candidate version for clinical evaluation with study participants. At the end of each testing loop, the technology and clinical teams jointly reviewed and interpreted the Masimo Root reference sensor and VitalSeer testing results to inform areas for refinement. Additional recommendations, observations, and clinical/user-centric insights into the data values and trends contributed to continual improvements to VitalSeer. These improvements are related to the graphical user interface (GUI), data acquisition process, system configuration and setup, calibration of key parameters and test conditions including, but not limited to, lighting configuration, subject’s motion and skin tone.

In consultation with Biomedical technology experts at Vancouver General Hospital (VGH) in Vancouver, BC, the FDA-approved Masimo Root patient monitoring and connectivity platform with Radius-7 wearable patient monitor, used by VGH for continuous patient monitoring, was identified as the reliable ground truth equipment. Due diligence on the Masimo Root system was performed by testing and verification of equipment specifications, including performance, usage, safety, vital sign data acquisition, instrumentation maintenance, sterilization/sanitation, and calibration. The clinical research team was trained on the optimal use of the Masimo system through iterative dialogues and an in-depth product education session with Masimo product and technical experts. This became the benchmark against which the veracity of datasets of HR, RR and SpO2 vital signs acquired by VitalSeer were compared and verified. The test setup, shown in Figure 3, was composed of the Masimo Root equipment as the ground truth sensor, and VitalSeer installed on a Lenovo T470s laptop (Windows 10), connected to Ethernet for optimal bandwidth and with laptop webcam and FLIR cameras as input imaging devices.
Figure 3. Test and validation setup in a simulated clinical setting, composed of a laptop running VitalSeer application with integrated and FLIR cameras and the Masimo Root reference equipment.

**Technology validation through clinical research studies:** Following the clinical lab testing loops, candidate versions of the VitalSeer software were released and used in a clinical research study, for which details are provided below. The study was approved by UBC clinical research ethics board and received institutional approval by Vancouver Coastal Health. All protocols were designed to follow the relevant COVID-related research restrictions.

Prospective participants were recruited through a provincial online research platform (https://www.reachbc.ca/). Interested volunteers who were 19 years old or older and able to attend a one-hour test session in the research lab located in Vancouver were eligible to participate. Volunteers were excluded if they were unable to read and understand English, were unable to display a clean face (i.e. no facial hair; no facial tattoos; no makeup; no glasses) during the session, were in exacerbation (i.e. experiencing worsening symptoms of an existing illness) or displaying abnormal physiological parameters, or if they were unable to sit still for two minutes. All participants were also required to pass a COVID screening questionnaire upon their arrival to the research lab.

During each test session, the physiological parameters (HR for VitalSeer 1.0; HR, RR and SpO2 for VitalSeer 1.2) were taken by the contactless sensing system via a 2-minute video recording of the face of the seated participant, along with the same measurements from the medical grade Masimo Root contact-type system. For consistency purposes, 3 to 5 2-minute recordings were taken per participant. At the end of the test session, each participant was invited to complete a usability feedback survey about their overall VitalSeer experience.

The raw input data (VitalSeer video footage and the Masimo Root log) was collected through the VitalSeer software. This allowed for reprocessing the data at any selected time with different parameters and algorithms, using the same processing pipeline as the live data collection. All electronic data was saved on a password protected, encrypted and firewall protected network server shared drive on the test laptop.

Meta-information associated with the recording sessions (e.g. participant ID, hardware information, lighting conditions) were stored within a data repository, REDCap (Research Electronic Data Capture), which allows the reuse of any recorded data offline and grouping data according to certain conditions. All other study participant data (ex. demographics) was de-identified and stored in REDCap. A secure data sharing process using MS OneDrive was established for teams to share and access data extracts from VitalSeer, Masimo, and the REDCap repository.

**Results**

**Healthcare provider interview findings:** Five physicians – 2 family physicians and 3 emergency physicians – participated in interviews. Two practiced in rural/remote locations, and 3 practiced in urban areas. The main gaps identified with remote vital sign measurement in virtual care included patient inaccessibility to available and affordable equipment; inability to collect objective and reliable measures of vital signs remotely; and inconsistent methods to follow patients over time. The potential benefits and opportunities of a contactless sensing technology, like VitalSeer, were described as increased engagement and motivation for patients to monitor their own health; improved patient confidence in accessing virtual services with more complete, accurate and objective information; increased physician confidence in decision making; and delivery of effective virtual care to the patients in their own 24/7.
homes. The participating physicians also highlighted the importance of establishing usability, reliability, affordability, trustworthiness, privacy, and data integration with other digital patient information platforms as required to establish VitalSeer as an innovative and scalable technology for virtual care for both patients and clinicians.

Responding to the findings from the interviews, and in conjunction with expert feedback from the clinical team, two use cases for the VitalSeer technology were proposed and refined: (1) short-term scenario: a stand-alone downloadable application and (2) long-term scenario: integration into existing virtual care platform(s).

VitalSeer design, development and testing: A set of clinical and user-centric objectives and requirements, based on the healthcare provider interview findings, were defined to drive the development of VitalSeer. Table 1 outlines the objectives and requirements, their associated performance indicators and the VitalSeer solution features/components.

**Table 1. Summary of clinical requirements linked to corresponding technology features/components**

<table>
<thead>
<tr>
<th>Objective and requirement</th>
<th>Performance indicator</th>
<th>VitalSeer feature/component implementation</th>
</tr>
</thead>
</table>
| **Clinical adoptability** by both healthcare professionals and patients | • Continuous, high-accuracy vital sign estimations
• Confidence level information on assessment results | • Automated offline pipeline to monitor model performance and benchmark assessment results
• Continuous confidence information display for the user on the application user interface (UI). |
| **Clinical accessibility** for the users | • Operating System (OS) agnostic: compatible with the most common OS running on home devices (Windows, OSX, Linux, iOS and Android).
• Hardware/device agnostic: provides accurate assessment on consumer grade devices | • OS Agnostic: avoids the use of any OS specific software component
• Hardware/device Agnostic: (1) outsource computationally-expensive processes to the server to maintain low requirements on users’ devices, (2) assess video quality and present assessment result with its associated confidence indicator (i.e. only displaying ultimate vital sign estimates to the patient or clinician on the UI), and (3) validate the inputs prior to performing a vital sign estimation using a quality metric. |
| **User-friendliness/ease of use and functional usability** | • User guidance for valid vitals assessment for optimal usability | • Provide UI components that: (1) guide the user into properly positioning their face in the target area, and (2) automatically warn of any issue (e.g. getting off the target area), enabling optimal use of the system for the user without technical help. |

The final proposed solution is composed of three main components:

1. VitalSeer Library: a portable Application Programming Interface (APIs) to interface with the input devices (e.g. cameras) and the backend server, and perform image processing and feature extraction;

2. VitalSeer Application: user-facing application (currently PC based, Linux or Windows) that guides users through the vital sign estimation sessions and displays near real-time vital signs measures;

3. VitalSeer Backend: the remote server that converts a time series of processed video features into vital signs by hosting the vital sign assessment models.

Responding to the defined use cases, VitalSeer can be launched either from a web browser (in a workflow that is compatible with virtual care) or through a downloadable desktop application. It includes different configurations to estimate different vital signs, each including the required hardware to be activated (e.g. camera, microphone) and the required processing pipeline, as well as information on how to display the results.

Figure 4 shows the VitalSeer application being used to display a subject’s vital sign estimations computed using only video recording of his face in near real-time. As examples of user-friendliness, the green oval is displayed to help the user adjust distance from the camera and thus guide the user for valid vitals assessment. The message box, above the
video, is included in order to display the status of the acquisition and the presence of any issues. The panel on the left displays the three vitals, HR, RR and SpO2 (in this paper the evaluation focusses on HR only).

![VitalSeer GUI](image)

**Figure 4.** VitalSeer GUI used for display of real-time video-based vital sign estimations.

In addition to the feature implementation presented in Table 1, the proposed software architecture is extensible to support additional operating systems by integrating the required APIs to access the input devices. Multiple level devices such as integrated laptop cameras, external webcams as well as industrial grade cameras (e.g. FLIR BlackFly) were interfaced and tested on various platforms such as desktop computers and laptops of different specifications.

**VitalSeer validation through clinical research studies – product validation findings and participant feedback:** A total of 35 healthy volunteers were enrolled into the study. Of the 35, 22 participated in the VitalSeer 1.0 clinical study sessions, and 9 of the 22 returned to participate in VitalSeer 1.2 study sessions as well. In addition, 13 new volunteers were also recruited to participate in VitalSeer 1.2 study sessions.

Data used for the vital sign estimate analysis came from different sources. A data processing step took place prior to the analysis step, which included the proper synchronization of the time reference of each device, e.g. sampling between the camera and Masimo Root. This was realized by compensating for the observed offset. The offset information was also used to ensure the clocks on all devices were synchronized in subsequent beta testing loops and clinic research experiments. For each 2-minute recording, 10-second running windows with 2 second overlaps were used to compute the heart rate estimations. To evaluate the accuracy of the HR model, a linear interpolation of all Masimo Root pulse rates within the corresponding VitalSeer analysis windows was performed. The error was calculated as the difference between the interpolated Masimo Root readings and the Vital Seer estimations.

There was a total of 165 recordings collected from the 44 test sessions across the VitalSeer 1.0 and 1.2 studies. Considering factors such as the stability in the pulse rate readings from the reference Masimo Root device, as well as the unexpected excessive subject motion in the acquired data, a total of 17757 pulse rate estimations (approximately 110 pulse rate estimations per 2-minute recording) was confirmed by the data processing step. VitalSeer also provides a confidence level on the HR estimation. This confidence level decreases as the signal-to-noise ratio (SNR) decreases. Signal quality can be impacted by various factors such as low lighting conditions, poor camera quality or unexpected motions of the face. When the confidence level is below a certain threshold, the HR estimation is filtered out. Fine tuning this threshold is essential to balance usability and accuracy of the vital sign estimation; in this study, this threshold was set to 90%, see details in\(^\text{15}\). Accuracy metrics were only computed for values which meet this confidence threshold. Of the 17757 pulse rate estimations performed, 86.1% of the HR estimations (15285) were above the confidence threshold.

The first row in Table 2 presents the model accuracy results and the prediction error as a function of the Masimo HR values. The results showed that the computed bias was negligible, Mean Absolute Error (MAE) was below 1 beat per minute, and the Root Mean Square Error (RMSE) of 1.2 beats per minute was below the claimed RMSE of FDA-approved clinical grade contact-based sensors utilized for heart rate measurements (on the order of 3 beats per minute, under ideal conditions)\(^\text{16}\).

It was observed that most low confidence recordings were observed with participants who had darker skin tones (IV and above on the Fitzpatrick scale). We hypothesized that this might be related to camera settings, and a pilot run
including the same subjects that was conducted using an industrial grade camera (i.e. FLIR Blackfly) to evaluate if the use of higher quality video capture devices could improve the quality of the signal. This pilot included 92 2-minutes videos that were recorded concurrently to those used in the main study, so subjects, lighting and reference pulse rate were identical. The results of the pilot (n=10127 pulse rate estimations) are presented in the second data row of Table 2. Using an industrial grade camera significantly increases the signal to noise ratio, leading to less readings rejected due to low confidence levels. However, the accuracy of the HR measurements of the FLIR camera versus the laptop integrated camera well confirmed the proposed low-cost consumer-grade based solution. Of note, VitalSeer using the FLIR Blackfly video data yielded accurate HR estimations for subjects with darker skin.

Table 2. VitalSeer 1.0 and 1.2 clinical studies – HR data analysis summary.

<table>
<thead>
<tr>
<th>Studies</th>
<th>Readings with confidence level &gt;90%</th>
<th>Mean Error (bpm)</th>
<th>RMSE (bpm)</th>
<th>MAE (bpm)</th>
<th>% of values within 2 bpm</th>
<th>% of values within 5 bpm</th>
<th>Outlier ratio (error &gt;10 bpm) %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregated VitalSeer 1.0 and 1.2 studies with integrated webcam</td>
<td>86.1% (n=15285 predictions)</td>
<td>-0.07</td>
<td>1.20</td>
<td>0.74</td>
<td>93.1</td>
<td>99.3</td>
<td>0.05</td>
</tr>
<tr>
<td>Pilot with FLIR Blackfly camera (concurrent to VitalSeer 1.2 study)</td>
<td>97.2% (n=9843)</td>
<td>-0.06</td>
<td>1.09</td>
<td>0.68</td>
<td>94.6</td>
<td>99.3</td>
<td>0.01</td>
</tr>
</tbody>
</table>

During the post-testing interviews with the participants on the various aspects of the experience and the technology, over 75% of participants agreed (37%) or strongly agreed (42%) it was easy to meet the requirement of having a clean face (i.e. no makeup, no facial hair, etc.) in order to have vital signs estimated by the technology; 95% of participants in VitalSeer 1.0 study, who took part in an average of 3.32 2-minute recordings per study session, reported it was comfortable to sit steady for about two minutes. However, in VitalSeer 1.2 study, where, on average, participants were asked to sit for 4.18 recordings per study sessions, 78% reported it was comfortable to sit steady for two minutes at a time.

At the end of each test session, participants in each study were asked if they felt the contactless sensing technology would be valuable for their own health management and care. Almost all, 93%, agreed or strongly agreed contactless sensing would be useful for self-monitoring. All also agreed or strongly agreed that they would be willing to share their vital sign readings with their healthcare provider online.

A majority, 86%, felt this method of vital signs measurement would be helpful for health care providers to monitor their patients remotely.

Discussion

In this work, a clinical research study framework and infrastructure was designed, built and tested, for the development of an innovative technology solution, between a multidisciplinary team with clinical and technical backgrounds. This partnership and the corresponding iterative development methodology fostered the lockstep development of a contactless sensing platform whose features and components were shaped to meet clinical and user-centric objectives, and to address the needs of virtual care. This collaboration needs to be grounded in a solid framework to guide the development of digital health applications that consider continuous and rigorous clinical evaluation and include end-users in the development loop.

From the clinical adoptability perspective, analyses of study recordings above the 90% confidence threshold (i.e. with high SNR) suggest the VitalSeer HR model attains the clinical adoptability performance goal of continuous, high-accuracy vital sign estimations by producing results with negligible computed bias, low MAE and an RMSE that is favorable compared to FDA-approved clinical grade contact-based sensors.

From the clinical accessibility perspective, accessibility of VitalSeer was addressed by the design and implementation of a hardware and operating system agnostic architecture. Through the testing, it was shown that the system could accommodate consumer-grade hardware while maintaining prediction accuracy. Hardware agnosticism is also further
supported by the fact that model evaluation metrics were comparable between the consumer-grade (e.g. laptop built-in cameras) and industrial-grade cameras. With a relatively high confidence threshold of 90% utilized in this analysis with the consumer-grade hardware, 14% of predictions were rejected in the analysis, which could be considered significant. However, in practice most rejected values are dispersed in time, so this rejection rate often translates into less frequent updates of the computed value (e.g. one update every 1-4s instead of every 1s), which might be suitable in many cases. Also, the confidence threshold is a design parameter, and it is possible to deduce the rejection rate to about 5% at the cost of a slightly increase accuracy\textsuperscript{15} to better fit a specific application. Future work should consider the integration and validation of the technology into and with other platforms such as mobile and tablet devices, while expanding the model capabilities to accommodate more variable environmental (in-the-wild) conditions. Such an endeavor will also allow the idea of operating system agnosticism to be evaluated.

From the ease-of-use perspective, the application user interface was developed and iterated over time with continual feedback from clinicians and other end-users. The resulting user interface is now a version that, among other features, provides real-time guidance and feedback (e.g. displays alerts/messages) about the position of the face and illumination management, reporting and automatic recovery when signal and vitals acquisition was challenging. Future work should continue to seek end-user feedback on operating VitalSeer in various environments so as to continuously adapt and improve the ease-of-use and usability of the application. This type of future endeavor should constitute studies that are inclusive to ensure VitalSeer’s ability to meet a wide range of needs.

From the affordability perspective, VitalSeer currently leverages generally available and relatively low-cost technologies to test and validate the viability of conducting virtual vital sign measurement. With further development and refinement, VitalSeer can operate under lower-bandwidth conditions, with less computing resource requirements and lower-resolution video cameras, thereby further reducing overall cost and increasing overall affordability for users of VitalSeer.

**Limitations:** First, the study was conducted under controlled lighting conditions to compute vital signs. An extensive validation would be needed to confirm that different lighting situations will still result in accurate data acquisition.

Second, most of study participants had relatively light skin tones. Based on the standardized Fitzpatrick skin tone scale\textsuperscript{18} from grades I to VI (the higher grade, the darker), they fall within grades I to III. Preliminary results from two subjects in the cohort suggest that the extraction of viable physiological signals is more challenging with regards to darker skin tones, which is consistent with reports in the literature\textsuperscript{19}, not only for rPPG but also for regular contact PPG\textsuperscript{20}, and thus, merits further investigation and work. Our findings must be validated on darker skin tones since less light gets reflected by darker skin tones and this may impact vital signs computations.

Third, in VitalSeer versions tested, the participants needed to sit motionless in front of the camera for two minutes for the vital signs to be computed and tracked across time. The VitalSeer sensing session had to be restarted if there was an object obscuring the patient’s face (e.g. a hand moving in front of the face), or if there was a significant movement within the region of interest such as a sneeze or a cough.

Fourth, our subjects were healthy volunteers with no severe illnesses during measurements. The platform will need to be tested with subjects experiencing a larger range of HR, RR, and SpO2 illnesses for broader use in all clinical contexts and illness severity.

Another limitation of the study relates to conducting in-person clinical research studies in the midst of the COVID-19 crisis, which limited recruitment and access to clinical settings to carry out the test sessions.

**Conclusion**

Empowering and bringing the technology to users has drawn increasing attention in the past decades, as this approach has been demonstrated to be critical for success of an innovation. The close clinical and technical teamwork on VitalSeer, presented in this paper, promotes excellence in health technology innovation by applying best practices in clinical research in the context of innovative medical device by clinical end-users and virtual care solution development by technology experts. The clinical insights collected from healthcare professionals and user-centric data effectively inform the VitalSeer research team in the development of the platform from a conceptual idea to its laboratory creation as a pre-clinical proof-of-concept. Next steps will include further usability studies to further investigate and validate the technology readiness for formal clinical trial to test its effectiveness in real world setting. Incorporating the experiences and insights of the end users, both healthcare providers and patients, will continue to be vital to inform the development of VitalSeer, ultimately to achieve successful implementation in virtual care settings to transform healthcare.
Our study also demonstrates the importance of the multidisciplinary nature of the collaboration to create this dynamic environment for science, engineering, medical, research and technology teams to iteratively exchange and realize innovative ideas in a rapid and disciplined fashion. The ongoing knowledge translation and dissemination activities will further contribute to socializing the technology with healthcare providers, patients and policy-makers for solution deployment, scale up, and sustainability.

References

Investigating the impact of weakly supervised data on text mining models of publication transparency: a case study on randomized controlled trials

Linh Hoang\textsuperscript{a} MS, Lan Jiang\textsuperscript{a}, MS, Halil Kilicoglu, PhD
School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, IL, USA

Abstract

Lack of large quantities of annotated data is a major barrier in developing effective text mining models of biomedical literature. In this study, we explored weak supervision to improve the accuracy of text classification models for assessing methodological transparency of randomized controlled trial (RCT) publications. Specifically, we used Snorkel, a framework to programmatically build training sets, and UMLS-EDA, a data augmentation method that leverages a small number of labeled examples to generate new training instances, and assessed their effect on a BioBERT-based text classification model proposed for the task in previous work. Performance improvements due to weak supervision were limited and were surpassed by gains from hyperparameter tuning. Our analysis suggests that refinements to the weak supervision strategies to better deal with multi-label case could be beneficial. Our code and data are available at https://github.com/kilicogluh/CONSORT-TM/tree/master/weakSupervision.

Introduction

Incomplete reporting and lack of transparency are common problems in biomedical publications and may reduce the credibility of the findings of a study. These problems can have serious consequences, particularly in clinical research publications, since the evidence from these studies inform patient care and healthcare policy. In clinical research, randomized controlled trials (RCTs) are the most robust kind of primary research evidence regarding the effectiveness of therapeutic interventions\textsuperscript{1} and are a cornerstone of evidence-based medicine\textsuperscript{2}. RCTs are expensive, and if inadequately designed, conducted, or reported, they lead to poor health outcomes and significant research waste\textsuperscript{3}.

Reporting guidelines have been proposed to improve transparency and completeness of reporting for various types of biomedical studies. For example, the CONSORT statement focuses on RCT reporting\textsuperscript{1,4}, and consists of a 25-item checklist and a flow diagram. While endorsed by many high-impact medical journals, adherence to CONSORT remains inadequate\textsuperscript{1,5} and difficult to enforce in practice, due to substantial workload it involves for journals. Manual CONSORT compliance checks before peer review have been shown to improve reporting quality\textsuperscript{6}; however, they are difficult to scale and require significant domain expertise.

In previous work, we presented a corpus of 50 RCT publications manually annotated at the sentence level with fine-grained CONSORT checklist items and proposed a text mining approach to automate the task of transparency (reporting quality) assessment\textsuperscript{7}. As a first step toward full transparency assessment, we developed sentence classification models to categorize sentences in the Methods sections of RCT publications into 17 methodology-related checklist items (e.g., Eligibility Criteria, Outcomes, Sequence Generation, Allocation Concealment). The best-performing model, based on BioBERT pretrained language model\textsuperscript{8}, yielded reasonable performance on some items, particularly those that are commonly discussed in RCT Methods sections and thus are well-represented in the dataset. However, the results overall suffered from the relatively small size of the dataset and largely failed on the checklist items that are infrequently reported in RCT publications (e.g., Changes to Outcomes).

Annotated data is critical in training modern natural language processing and text mining (NLP) algorithms. In particular, deep neural network architectures heavily depend on large quantities of training data for learning model parameters. While recent pretrained language models, such as BERT\textsuperscript{9} and its variants, exhibit better sample efficiency and often work well even with relatively small datasets, the importance of annotated data has not diminished. High performance of BERT-based models in NLP tasks and the resulting standardization of architectures arguably underlines data scarcity as the primary bottleneck in NLP\textsuperscript{10}. In response, weak supervision techniques have become increasingly popular, as they offer cheaper or more efficient ways for generating training data\textsuperscript{10}.

\textsuperscript{a}Equal contribution.
In this study, we investigated whether weak supervision techniques can be used to effectively label additional data and improve our sentence classification models for transparency assessment of RCT publications. More specifically, we focused on weak supervision using the Snorkel framework\(^\text{10}\) and data augmentation based on the UMLS-EDA algorithm\(^\text{11}\) and used the labels that they generated as additional data for our previously reported BioBERT-based model\(^\text{7}\). The results show that weak supervision has limited effectiveness on our dataset, while at the same time indicating that hyperparameter tuning can have a more significant impact on model performance.

**Related Work**

**Weak supervision**

Weak supervision seeks to use domain knowledge and subject matter expertise in opportunistic ways to assign (somewhat noisy) labels to unlabeled data or generate synthetic data. Several general approaches to weak supervision exist. One well-known technique is *distant supervision*\(^\text{12}\), based on using domain knowledge in external knowledge bases. While often used for relation extraction\(^\text{12,13}\), it has also been used for classification tasks applied to RCT publications\(^\text{14,15}\). For example, risk of bias judgements in the Cochrane database of systematic reviews were used to automatically label sentences in RCT publications and train models for assessing risk of bias in the publications\(^\text{14}\).

Another related approach is *data augmentation*, the goal of which is to increase a model’s generalizability by generating realistic data from a limited number of existing examples. First proposed in computer vision research\(^\text{16}\), it has more recently been adopted in NLP research as well\(^\text{11,17}\). For example, simple transformations of individual sentences (e.g., synonym replacement, random insertion/deletion) were used to generate additional data and improve modeling accuracy with small datasets\(^\text{11}\). Similar approaches have been adapted to biomedical domain, for tasks ranging from medical abbreviation recognition\(^\text{18}\) to named entity recognition\(^\text{19}\).

Snorkel has been proposed as a general weak supervision framework\(^\text{10}\). Based on *data programming* paradigm, Snorkel relies on user-defined labeling functions (LFs), which are heuristic methods that can noisily label large quantities of unlabeled data, learns a generative model over the labeling functions to estimate their accuracy and correlations, and generates probabilistic labels that can be used to train machine learning models. Snorkel has been applied to several biomedical text mining tasks, outperforming distant supervision baselines and approaching manual supervision\(^\text{10}\). Other weak supervision approaches have also been developed for biomedical NLP tasks, including smoking status classification from clinical notes\(^\text{20}\), semantic indexing\(^\text{21}\), and clinical entity classification\(^\text{22}\).

**Text mining on RCT publications**

Text mining on RCT literature has mostly focused on annotating and extracting study characteristics relevant for systematic reviews and evidence synthesis\(^\text{23,24}\). PICO elements received most attention; several corpora have been developed at the sentence and span levels\(^\text{15,25,26}\), and a variety of traditional and deep machine learning models have been developed to extract these elements from abstracts or full text\(^\text{15,26–28}\). There is less research on non-PICO elements. Most notably, RobotReviewer\(^\text{14}\) focuses on risk of bias assessment and classifies RCT publications as high or low risk on several risk categories, including sequence generation and allocation concealment. ExaCT\(^\text{29}\) identifies 21 elements in clinical trial publications including sample size and drug dosage. Recently, we constructed a corpus of 50 RCT publications (named CONSORT-TM) annotated at the sentence level with 37 fine-grained CONSORT checklist items to assist with transparency assessment\(^\text{7}\). We also developed baseline NLP models to recognize 17 methodology-specific CONSORT items: two rule-based methods (one keyword-based and another section header-based) as well a linear SVM classifier and a BioBERT-based model. The BioBERT model performed best overall (micro precision: 0.82, recall: 0.63, and F\(_1\): 0.72), although it failed to recognize infrequent items, which partly motivated this study.

**Materials and Methods**

We explored weak supervision to improve the classification performance of our best-performing BioBERT model\(^\text{7}\). In this section, we first describe the collection and pre-processing of unlabeled RCT data from PubMed Central (PMC) for weak supervision. Second, we provide a brief description of the baseline BioBERT models. Third, we discuss our methodology for generating labels using Snorkel framework as well as the UMLS-EDA algorithm. Lastly, we provide evaluation details. The overall procedure is illustrated in Figure 1.

---

\(^{10}\text{Snorkel} = \text{Snorkel} Framework, \text{UMLS-EDA} = \text{Uniform Medical Language System - Expanded Data Augmentation, F\(_1\) = F-Measure.}\)
Data collection and pre-processing

We followed the data collection strategy used in previous work\(^7\) to obtain a large set of RCT articles. Cochrane precision-maximizing search query\(^b\) was used on 1/15/2021 to search PMC Open Access subset (PMC-OA) for RCT articles published between 1/1/2011 and 12/31/2020\(^c\). The results were further limited to articles that have full-text XML in PMC-OA. To get a more reliable RCT subset (since publication types in PubMed can be inaccurate), we filtered the results through RCT Tagger\(^30\), a machine learning model that determines whether a publication is a RCT or not. Its accuracy was found to be 99.7% in predicting RCT studies included in Cochrane systematic reviews\(^31\). Lastly, we eliminated publications with the word protocol in their title (generally study protocol publications).

We used NCBI e-utils\(^d\) to retrieve publications in XML format, and split them into sentences using our in-house sentence splitter\(^32\). Only sentences that belong to Methods section of the publications were taken into account. Stanford CoreNLP\(^33\) package was used for tokenization and part-of-speech tagging. We eliminated the sentences meeting the following criteria from further consideration, since they are unlikely to indicate CONSORT methodology items: a) contains fewer than five tokens; b) contains numbers only; and c) is a section header or a table/figure caption.

Baseline models

Our best-performing classifier in previous work\(^7\) was a BioBERT-based sentence classification model, which uses the BioBERT pretrained language model\(^8\) as a sentence encoder, considers the model’s output for the [CLS] token as the sentence representation, and trains a sigmoid layer for multi-label classification of 17 CONSORT methodology items. The input to the model is the raw sentence text prepended with its subsection header. The classifier was implemented using simpletransformers\(^e\). We refer to this model as BASELINE below.

In this study, we used the huggingface\(^f\) BERT implementation. While mostly using the same hyperparameters as BASELINE (batch size: 4, number of epochs: 30, optimizer: Adam, dropout: 0.1), we modified two hyperparameters. First, we used adaptive learning rate instead of a fixed learning rate to optimize the algorithm with different rates based the model performance during training. Second, we set the gradient accumulation steps to 1 (16 for BASELINE), which increases the frequency of model parameter updates. We refer to this optimized model as BASELINE_OPT below.

\(^b\)https://work.cochrane.org/pubmed
\(^c\)The start date is chosen based on the most recent publication of CONSORT guidelines (2010)\(^1\).
\(^d\)https://www.ncbi.nlm.nih.gov/books/NBK25501/
\(^\circ\)https://github.com/ThilinaRajapakse/simpletransformers
\(^f\)https://huggingface.co/
Generating weak labels using Snorkel

Snorkel\textsuperscript{10} generates weak labels in three steps: a) LF construction; b) creation of a generative model to capture label agreements/disagreements; and c) generation of probabilistic labels for sentences. Input for Snorkel pipeline are unlabeled sentences from RCT publications from PMC-OA.

LFs are expert-defined heuristic rules that can be used to label sentences. For NLP tasks, these can be based on text patterns, syntactic structure, or external knowledge bases. In general, LFs that have high coverage and low overlap are desirable. Such LFs apply to many instances in the dataset yet are unique enough to distinguish instances with different labels. In this study, we used three LF approaches to label CONSORT items: keyword-based, section header-based, and sentence similarity-based. 17 individual LFs were created for each approach (one corresponding to each label).

**Keyword-based LFs.** These LFs mimic the keyword-based method used in previous work\textsuperscript{7}. Each CONSORT item is associated with a set of keywords or phrases (e.g., \textit{power to detect} with Sample Size Determination (7a)\textsuperscript{7}). A total of 232 phrases are used. Each LF checks whether an input sentence contains one of its keyphrases, and if so, returns the corresponding label as a weak label (or NO-LABEL, if the sentence does not contain any relevant keyword/phrase).

**Section header-based LFs.** These LFs also mimic a baseline method from earlier work\textsuperscript{7}. In this case, common subsection headers in Methods sections are associated with CONSORT labels. 48 section header keywords/phrases are mapped to CONSORT items (e.g., the word \textit{concealment} to the item Allocation Concealment (9)). These LFs check whether the header of the section to which the sentence belongs matches one of the relevant key phrases.

**Sentence similarity-based LFs.** These LFs assign weak labels to unlabeled sentences based on their similarity to a set of “ground truth” sentences (95 sentences provided as examples for checklist items in the CONSORT Explanation and Elaboration document\textsuperscript{1} and the CONSORT website\textsuperscript{2}). We used BioBERT to generate vector representations of these sentences. Given an unlabeled sentence, we calculate its cosine similarity with every ground truth sentence and consider two labels based on similarity scores: the label of the sentence with the highest similarity and the label that appears most frequently for the top 10 most similar ground truth sentences. If two labels are the same, we use it as the sentence label. Manual checks showed this combination to be more accurate than the most similar sentence label only.

Snorkel applies all LFs to generate a LF matrix that shows the coverage, overlaps, and conflicts between the LFs. Coverage information indicates the fraction of the dataset to which a particular LF is applied. Overlap shows the fraction of dataset where a particular LF and at least one other LF agree. Conflict indicates the fraction of dataset where a particular LF and at least one other LF disagree. Snorkel pools noisy signals from these three features into a generative model to learn the agreements and disagreements of the LFs, thus assessing the weights of accuracy for each LF. The model then takes into account these accuracies to make a final label prediction for each sentence.

Generating synthetic data using UMLS-EDA

Several CONSORT items are infrequently reported, as they are contingent upon changes in the trial, which may or may not occur (e.g., Changes to Trial Design (3b)). In previous work, text mining methods yielded poor results for these classes\textsuperscript{7}, as may be expected. BASELINE model, although it performed best overall in terms of micro-averaging, yielded no predictions for five labels (out of 17) and less than 0.5 \textit{F}$_1$ score for 11 items. We do not expect Snorkel to provide significant number of examples for infrequently reported items, since they are also likely to be rare in the unlabeled dataset and Snorkel’s generative model relies on LF agreement, also likely to be uncommon for such labels.

Therefore, we sought to improve the classification performance for such infrequently reported labels using data augmentation. Specifically, we used UMLS-EDA\textsuperscript{19} and leveraged UMLS\textsuperscript{34} synonyms to generate sentences that are similar to CONSORT-TM training instances. We define a class as rare if the class frequency in the original dataset \((f)\) is under a pre-determined threshold \((t)\). In generating instances, we make up the difference between the frequency in the original dataset and the threshold (i.e., \(t \cdot f\) instances generated) to make the distribution of the training dataset more uniform. If a class is not rare in the original dataset (i.e., \(f \geq t\)), no sentences are generated for that label.

UMLS-EDA uses five operations to augment data. \textit{Synonym replacement using WordNet} randomly chooses \(n\) words

\begin{itemize}
  \item \textit{we use the item numbers used in CONSORT guidelines, as well, hereafter.}
  \item \textit{http://www.consort-statement.org/examples/sample}
\end{itemize}

257
Table 1: Example of data augmentation using UMLS-EDA. Bold words indicate modifications made by UMLS-EDA. The label of the original sentence is Eligibility Criteria (4a).

<table>
<thead>
<tr>
<th>Operation</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>children were excluded if they had impaired fasting glucose, were diabetic, or reported a diagnosed renal, or hepatic disease that might alter body weight.</td>
</tr>
<tr>
<td>Synonym replacement (WordNet)</td>
<td>children were leave off if they had impaired fast glucose, were diabetic, or account a diagnosed renal, or liverwort disease that mightiness alter body weight.</td>
</tr>
<tr>
<td>Random insertion</td>
<td>children mightiness were excluded if they had impaired fasting mightiness leave off glucose, were diabetic, or reported a diagnosed child renal, or hepatic disease that might alter body weight.</td>
</tr>
<tr>
<td>Random swap</td>
<td>children were diagnosed if they had impaired fasting glucose, might excluded or a reported renal, diabetic, hepatic disease that were alter body weight.</td>
</tr>
<tr>
<td>Random deletion</td>
<td>children were excluded if they had impaired fasting glucose, were diabetic, or reported a diagnosed renal, or hepatic disease that might alter body weight.</td>
</tr>
<tr>
<td>Synonym replacement (UMLS)</td>
<td>children were excluded if they had impaired fasting glycaemia, were diabetic, or informing a diagnose nephros gastric, or liver disease that might alter body weight.</td>
</tr>
</tbody>
</table>

from the given sentence that are not stopwords and replaces each with a synonym randomly chosen from WordNet. Random insertion inserts random WordNet synonyms of n words in the sentence in random positions. Random swap randomly swaps the position of two words and repeats this n times. Random deletion samples and deletes n words according to a uniform distribution. Synonym replacement using UMLS identifies all the UMLS concepts in the sentence and randomly replaces n words in the sentence with a UMLS synonym, also randomly selected. Operations of the UMLS-EDA data augmentation are illustrated on an example sentence in Table 1. The parameter n is determined dynamically based on the sentence length (l) and the operation type (n = 0.5*l for synonym replacement with UMLS at most and n = 0.2*l for others). While UMLS-EDA aims to generate t-f instances, in most cases, a larger number of instances are generated using these operations and we subsample from the generated instances to reach the threshold.

Evaluation

To evaluate whether weak supervision generated labels useful for improving sentence classification performance, we compared the results obtained with BASELINE model on the CONSORT-TM dataset using 5-fold cross validation to results obtained when weakly labeled examples from different strategies are added to the training portion of the folds in cross validation (Figure 1). In this setup, data used for validation and testing in each fold remain the same for all the models. As in previous work, we used precision, recall, and their harmonic mean, F₁ score, and calculated 95% confidence intervals. In addition to calculating these measures per CONSORT item, we also report micro- and macro-averaged results and the area under ROC curve (AUC).

Results

Weak supervision using Snorkel

Our search strategy retrieved 608K RCTs from PubMed, 155,183 of which have XML full text in PMC. RCT Tagger predicted 71,948 of these as RCTs. Considering only those predicted with a confidence score over 0.95 reduced the dataset to 14,534 publications. Further eliminating publications with protocol in the title, we obtained a set of 11,988 papers. A total of 721,948 sentences from these publications was reduced to 551,936 sentences after filtering.

We processed 551,936 unlabeled sentences using the Snorkel model, which generated 17 probabilities for each sentence. We empirically set a probability threshold of 0.8 to predict the final weak labels for the unlabeled sentences. If no label was predicted with a probability higher than 0.8, no label was assigned. The distribution of weak labels generated by Snorkel are shown in Table 2. Most weak labels corresponded to items that are already relatively well-represented in the dataset; thus, we limited the number of weakly labeled examples for each CONSORT item to a pre-determined threshold in our classification experiments and randomly sampled these examples. We report the results with the threshold that performed best in our experiments (500).
Weak supervision using UMLS-EDA

We used thresholds 50, 100, and 200 to generate 246, 844, and 2217 additional examples, respectively, using UMLS-EDA. Data augmentation was implemented as part of 5-fold cross-validation; and therefore, number of examples between folds differ. The numbers of instances for each label in the original dataset and the augmented datasets (for one of the folds) are shown in Table 2. A label can be considered rare or not at different threshold values and may or may not be augmented. For example, while the item Trial Design (3a) is not rare when the threshold is 50, it is considered rare for the threshold 100 and, therefore, augmented (Table 2). The number of rare class instances generally exceed to the threshold slightly, because it is possible to label an augmented example with more than one class.

Table 2: The frequency of each methodology item in CONSORT-TM and the augmented data generated by Snorkel and UMLS-EDA. Numbers in bold correspond to the cases when the CONSORT item was considered rare and augmented for the given threshold $t$.

<table>
<thead>
<tr>
<th>CONSORT Item</th>
<th>Snorkel</th>
<th>UMLS-EDA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original</td>
<td>$t=50$</td>
</tr>
<tr>
<td>Trial Design (3a)</td>
<td>3,932</td>
<td>55</td>
</tr>
<tr>
<td>Changes to Trial Design (3b)</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Eligibility Criteria (4a)</td>
<td>17,182</td>
<td>129</td>
</tr>
<tr>
<td>Data Collection Setting (4b)</td>
<td>740</td>
<td>32</td>
</tr>
<tr>
<td>Interventions (5)</td>
<td>11,415</td>
<td>199</td>
</tr>
<tr>
<td>Outcomes (6a)</td>
<td>24,104</td>
<td>535</td>
</tr>
<tr>
<td>Changes to Outcomes (6b)</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Sample Size Determination (7a)</td>
<td>6,674</td>
<td>93</td>
</tr>
<tr>
<td>Interim Analyses / Stopping Guidelines (7b)</td>
<td>124</td>
<td>14</td>
</tr>
<tr>
<td>Sequence Generation (8a)</td>
<td>7</td>
<td>35</td>
</tr>
<tr>
<td>Randomization Type (8b)</td>
<td>2,915</td>
<td>40</td>
</tr>
<tr>
<td>Allocation Concealment (9)</td>
<td>274</td>
<td>13</td>
</tr>
<tr>
<td>Randomization Implementation (10)</td>
<td>1,785</td>
<td>42</td>
</tr>
<tr>
<td>Blinding (11a)</td>
<td>525</td>
<td>47</td>
</tr>
<tr>
<td>Similarity of Interventions (11b)</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>Statistical Methods for Outcomes (12a)</td>
<td>45,353</td>
<td>215</td>
</tr>
<tr>
<td>Statistical Methods for Other Analyses (12b)</td>
<td>49</td>
<td>65</td>
</tr>
<tr>
<td>NO_LABEL</td>
<td>436,854</td>
<td></td>
</tr>
</tbody>
</table>

Classification results

We evaluated BASELINE and BASELINE_OPT models on CONSORT-TM using 5-fold cross-validation. In other experiments, we used various sizes of weakly supervised data from Snorkel and UMLS-EDA for additional training. For brevity, we only report the weak supervision results for the best-performing model-data size combinations. For Snorkel, this is BASELINE_OPT model augmented with maximum 500 examples per label. For UMLS-EDA, it is the same model augmented with UMLS-EDA data with a threshold of 50. The results are provided in Table 3 (due to space constraints, we provide precision and recall as supplementary material on the project GitHub repository). The results show that hyperparameter tuning (BASELINE_OPT) makes a significant difference in performance (7% increase in micro-$F_1$ and 63% in macro-$F_1$), while the impact of weak supervision strategies seems minor; Snorkel data leads to a slight performance degradation, while UMLS-EDA data increases micro-$F_1$ by one percentage point and AUC with 1.6 points, with practically no change in macro-$F_1$.

Discussion

Weak supervision with Snorkel

Approximately 21% of unlabeled sentences were weakly labeled by Snorkel. The number of weak labels reflected to some extent the distribution of labels in the original dataset. Many sentences were weakly labeled with common labels.
Table 3: Classification results using CONSORT-TM and weakly supervised data. SNORKEL uses BASELINE_OPT with additional 500 instances per label from Snorkel data. UMLS-EDA(50) uses BASELINE_OPT with additional instances from UMLS-EDA to add up to at least 50 instances for each label. 3a: Trial Design; 3b: Changes to Trial Design; 4a: Eligibility Criteria; 4b: Data Collection Setting; 5: Interventions; 6a: Outcomes; 6b: Changes to Outcomes; 7a: Sample Size Determination; 7b: Interim Analyses/Stopping Guidelines; 8a: Sequence Generation; 8b: Randomization Type; 9: Allocation Concealment; 10: Randomization Implementation; 11a: Blinding Procedure; 11b: Similarity of Interventions; 12a: Statistical Methods for Outcomes; 12b: Statistical Methods for Other Analyses.

P: precision; R: recall; F: F1 score; CI: confidence interval; AUC: Area Under ROC Curve.

<table>
<thead>
<tr>
<th>CONSORT Item</th>
<th>BASELINE F1 [CI]</th>
<th>BASELINE_OPT F1 [CI]</th>
<th>SNORKEL(500) F1 [CI]</th>
<th>UMLS-EDA(50) F1 [CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>3a</td>
<td>0.63 [0.46, 0.80]</td>
<td>0.82 [0.69, 0.95]</td>
<td>0.75 [0.63, 0.88]</td>
<td>0.78 [0.72, 0.84]</td>
</tr>
<tr>
<td>3b</td>
<td>0.00 [0.00, 0.00]</td>
<td>0.00 [0.00, 0.00]</td>
<td>0.00 [0.00, 0.00]</td>
<td>0.00 [0.00, 0.00]</td>
</tr>
<tr>
<td>4a</td>
<td>0.85 [0.76, 0.95]</td>
<td>0.89 [0.82, 0.96]</td>
<td>0.88 [0.82, 0.94]</td>
<td>0.90 [0.85, 0.95]</td>
</tr>
<tr>
<td>4b</td>
<td>0.36 [0.06, 0.65]</td>
<td>0.87 [0.74, 1.00]</td>
<td>0.79 [0.61, 0.97]</td>
<td>0.81 [0.68, 0.94]</td>
</tr>
<tr>
<td>5</td>
<td>0.72 [0.66, 0.78]</td>
<td>0.75 [0.68, 0.81]</td>
<td>0.73 [0.66, 0.81]</td>
<td>0.75 [0.68, 0.83]</td>
</tr>
<tr>
<td>6a</td>
<td>0.81 [0.74, 0.88]</td>
<td>0.82 [0.75, 0.89]</td>
<td>0.83 [0.72, 0.87]</td>
<td>0.83 [0.74, 0.91]</td>
</tr>
<tr>
<td>6b</td>
<td>0.00 [0.00, 0.00]</td>
<td>0.00 [0.00, 0.00]</td>
<td>0.00 [0.00, 0.00]</td>
<td>0.00 [0.00, 0.00]</td>
</tr>
<tr>
<td>7a</td>
<td>0.84 [0.76, 0.92]</td>
<td>0.88 [0.87, 0.90]</td>
<td>0.90 [0.86, 0.94]</td>
<td>0.90 [0.87, 0.93]</td>
</tr>
<tr>
<td>7b</td>
<td>0.00 [0.00, 0.00]</td>
<td>0.70 [0.47, 0.94]</td>
<td>0.70 [0.17, 1.22]</td>
<td>0.78 [0.51, 1.05]</td>
</tr>
<tr>
<td>8a</td>
<td>0.38 [0.15, 0.60]</td>
<td>0.88 [0.77, 1.00]</td>
<td>0.86 [0.60, 0.91]</td>
<td>0.88 [0.79, 0.98]</td>
</tr>
<tr>
<td>8b</td>
<td>0.38 [0.10, 0.67]</td>
<td>0.73 [0.53, 0.93]</td>
<td>0.67 [0.51, 0.83]</td>
<td>0.75 [0.60, 0.90]</td>
</tr>
<tr>
<td>9</td>
<td>0.00 [0.00, 0.00]</td>
<td>0.45 [0.35, 0.54]</td>
<td>0.40 [0.03, 0.76]</td>
<td>0.43 [0.26, 0.60]</td>
</tr>
<tr>
<td>10</td>
<td>0.24 [0.05, 0.43]</td>
<td>0.53 [0.36, 0.71]</td>
<td>0.50 [0.22, 0.77]</td>
<td>0.52 [0.32, 0.72]</td>
</tr>
<tr>
<td>11a</td>
<td>0.42 [0.12, 0.71]</td>
<td>0.66 [0.59, 0.74]</td>
<td>0.59 [0.46, 0.72]</td>
<td>0.66 [0.54, 0.77]</td>
</tr>
<tr>
<td>11b</td>
<td>0.00 [0.00, 0.00]</td>
<td>0.45 [0.06, 0.85]</td>
<td>0.41 [0.04, 0.77]</td>
<td>0.44 [0.04, 0.84]</td>
</tr>
<tr>
<td>12a</td>
<td>0.75 [0.69, 0.81]</td>
<td>0.77 [0.69, 0.85]</td>
<td>0.78 [0.69, 0.87]</td>
<td>0.78 [0.72, 0.84]</td>
</tr>
<tr>
<td>12b</td>
<td>0.04 [-0.06, 0.14]</td>
<td>0.32 [0.27, 0.38]</td>
<td>0.24 [0.07, 0.40]</td>
<td>0.31 [0.26, 0.37]</td>
</tr>
<tr>
<td>Micro-average</td>
<td>0.72 [0.66, 0.76]</td>
<td>0.77 [0.71, 0.84]</td>
<td>0.76 [0.69, 0.82]</td>
<td>0.78 [0.72, 0.83]</td>
</tr>
<tr>
<td>Macro-average</td>
<td>0.38 [0.34, 0.41]</td>
<td>0.62 [0.55, 0.69]</td>
<td>0.58 [0.48, 0.68]</td>
<td>0.62 [0.55, 0.69]</td>
</tr>
<tr>
<td>AUC</td>
<td>0.812</td>
<td>0.876</td>
<td>0.875</td>
<td>0.892</td>
</tr>
</tbody>
</table>

(e.g., Outcomes (6a)). On the other hand, Snorkel failed to weakly label any sentences with the two least frequent labels (Table 2). The quality of Snorkel labels depends largely on the quality of LFs. We used two LFs based on heuristics explored in previous work. Micro-F1 for both methods were found to be around 0.50 in previous work (0.50 for keyword-based and 0.45 for section header-based). More accurate LFs could improve Snorkel results.

To better understand the quality of Snorkel-generated weak labels, we sampled 318 sentences and two authors of this paper (LH and HK) independently labeled the sentences, without access to Snorkel labels. We calculated the agreement of these annotations with Snorkel-generated labels, using Krippendorff’s $\alpha$ with the distance metric MASI which accounts for partial agreement in the case of multiple labels. $\alpha$ agreements between Snorkel and each annotator were found to be 0.46 and 0.61, respectively. Inter-annotator agreement was 0.59. Interestingly, agreement between Snorkel and simple majority vote was 0.93. These results suggest that Snorkel may converge to this simple heuristic in some cases, and that it behaves more or less like another annotator in the process.

We found that a large percentage of annotator disagreement with Snorkel came from randomization-related labels (items 8a, 8b, 9, and 10). These items often appear in the same sentence and the clues for them can be overlapping, making it a challenge to label them accurately for both humans and automated methods. In previous work, we found inter-annotator agreement for these items to be somewhat low as well ($\alpha=0.62, 0.48, 0.34, 0.35$, respectively$)^7$. Snorkel tends to pick a single label for sentences, and this was especially problematic for randomization-related sentences.
Weak supervision using UMLS-EDA

Data augmentation is expected to reduce overfitting and help with model robustness\cite{16}. While generating data using UMLS-EDA is relatively cheap, the resulting sentences are generally not meaningful, making it difficult to assess the quality of the augmented data (in contrast to Snorkel), aside from the downstream model performance that it produces. We make several observations based on our examination of the augmented data. One of the data augmentation operations (synonym replacement with UMLS) may need to be refined. UMLS synonyms that are used to replace the original words/phrases are sometimes different from the original only in trivial ways (acronyms or swapped tokens), and strategies that only allow more significant replacements could be beneficial. For example, it might be worthwhile to limit the replacement only to terms of particular semantic types or part-of-speech tags. Similar observations were made for synonym replacement with WordNet, as well. Some replacements involved functional words, which may not be as beneficial as replacing content words (nouns, adjectives).

Effect of weak supervision and model hyperparameters on classification performance

We did not observe significant improvements in classification performance due to weakly supervised data. Neither strategy led to any correct predictions for the two least frequent labels (3a, 6a). While this was not unexpected in the case of Snorkel (as no additional examples were labeled with these items), it was more surprising in the case of UMLS-EDA, which seemed to generate sufficient number of examples for these items. We observed AUC improvement with UMLS-EDA (0.892 vs. 0.876 with BASELINE\textsubscript{OPT}), which may indicate that UMLS-EDA does help with robustness and generalizability. As UMLS-EDA approach is cheap, additional refinements may be promising as a future direction.

Somewhat to our surprise, we found that model hyperparameters made a much more significant difference in model performance. BASELINE\textsubscript{OPT} model yielded about 7% improvement in micro-F\textsubscript{1} and 63% improvement in macro-F\textsubscript{1} over the BASELINE model, with improvements in almost all labels. To assess how hyperparameters interacted with weak supervision, we also measured performance when BASELINE model (instead of BASELINE\textsubscript{OPT}) was trained with weakly supervised data. Using Snorkel for weak supervision in this scenario improved micro-F\textsubscript{1} from 0.72 to 0.75, suggesting that hyperparameter optimization may, in some cases, obviate the need for additional (noisy) data.

Limitations

Our investigation was limited to one relatively small corpus. The findings regarding weak supervision (as well as Snorkel and UMLS-EDA specifically) may not be generalizable to other corpora. We used few heuristics with modest performance as LFs and Snorkel label quality is likely to be improved with with additional more accurate LFs; however, this requires significant domain expertise. While we performed some hyperparameter tuning, we did not do an exhaustive search, and it is possible that more optimal hyperparameters can improve results further.

Conclusions and future work

We investigated the impact of two weak supervision strategies on multi-label sentence classification models of RCT publications. We did not observe a clear positive impact of weak supervision on the specific task we studied. More experiments would be needed to determine whether this is a corpus-specific finding or more general. Various forms of weak supervision has been shown to improve classification performance\cite{11,15}, mostly in multi-class cases; therefore, it is possible that our weak supervision strategies need more refinement for the multi-label case.

In future work, we plan to refine our approach. For example, in UMLS-EDA, we can devise methods to generate more contextually appropriate synonyms from WordNet and UMLS. Snorkel would benefit from more accurate LFs. Other semi-supervised learning approaches (e.g., self-training\cite{36}, few-shot learning\cite{37}) can also be investigated as alternatives.

References


A Privacy-Preserved Transfer Learning Concept to Predict Diabetic Kidney Disease at Out-of-Network Siloed Sites Using an In-Network Federated Model on Real-World Data

Humayera Islam, MS1,4, Khuder Alaboud, MS1,4, Tanmoy Paul3,4, Md. Kamruz Zaman Rana, MS1,4, Abu Mosa, PhD, FAMIA*1,2,3,4

1Institute for Data Science and Informatics; 2Department of Health Management and Informatics; 3Department of Electrical Engineering and Computer Science; 4NextGen Biomedical Informatics Center; University of Missouri School of Medicine, Columbia, Missouri; *Corresponding Author. Email: mosaa@health.missouri.edu

Abstract
Successful implementation of data-driven artificial intelligence (AI) applications requires access to large datasets. Healthcare institutions can establish coordinated data-sharing networks to address the complexity of large clinical data accessibility for scientific advancements. However, persisting challenges from controlled access, safe data transferring, license restrictions from regulatory and legal concerns discourage data sharing among the in-network hospitals. In contrast, out-of-network healthcare institutions are deprived of access to any big EHR database; hence, limiting their research scope. The main objective of this study is to design a privacy-preserved transfer learning architecture that can utilize the knowledge from a federated model developed from in-network hospital-site EHR data for predicting diabetic kidney cases at out-of-network siloed hospital sites. In all our experiments, transfer learning showed improved performance compared to models trained with out-of-network site datasets. Thus, we demonstrate the proof-of-concept of transferring knowledge from established networks to aid data-driven AI discoveries at siloed sites.

Introduction
Data-driven artificial intelligence (AI) applications engender the scope of substantial knowledge discoveries in many industries. However, the successful implementation of data-driven approaches requires large and diverse datasets1. Similarly, AI-driven research using large amounts of digital healthcare data in electronic health records (EHR) can help address relevant persisting clinical research questions to facilitate evidence-based treatment2. However, existing policies to protect patient privacy and organizations’ proprietary rights over their data alongside compliance to regulatory requirements from the HIPAA and HITECH Acts impose legal, technical, and financial burdens on data sharing2,3. In recent times, to address the complexity of data sharing among healthcare institutions, national networks have emerged to develop coordinated distributed data networks to advance research for improving healthcare services and outcomes.

A group of healthcare institutions coming together to establish a coordinated data-sharing network to conduct large-scale health research is referred to as a clinical research network (or “network”, in general). The affiliated “partners” or data contributors of the established network often retain its data locally while accessing some linked data sources with regulatory agreements. One example of such an extensive network is PCORnet, a national-scale clinical research network consisting of 11 other networks of healthcare institutions such as Greater Plains Collaborative (GPC)4, engaging EHR data from over 80 million Americans5. Other examples of data networks include the Food and Drug Administration (FDA) Mini-Sentinel6, NIH Distributed Research Network7, and ESPnet8.

Although the EHR data procured through the network undergo extensive anonymization of patient identifiers, a few elements may allow for reidentification of the identifiers or information leakage9. Moreover, challenges arising from controlled access, safe data transferring, license restrictions from regulatory and legal concerns often keep the data sharing process complex and enduring. Federated learning (FL) is a learning mechanism that can potentially address the persisting concerns of data governance, sharing, and privacy1,10. In FL, machine learning algorithms can be trained collaboratively on data stored from independent health systems without exchanging the raw data itself. In this way, FL can be utilized in established networks like GPC and others to build high-performance predictive models on a large-scale, potentially growing, decentralized database, enabling novel research on complex and rare health outcomes.
In contrast, the healthcare institutions that do not participate in any bigger network are deprived of access to any big EHR database or data sharing mechanism unless the database is publicly available or available through license purchasing. This poses major challenges for these siloed healthcare institutions to conduct novel research on healthcare problems since the only available data for them are their own clinical records. In addition, data-driven research conducted on such local siloed datasets can introduce demographics biases or misrepresent other population characteristics, leading to the non-generalizability of their results.

We propose one approach to address this important persisting problem using the concept of knowledge transportation through “transfer learning”. Transfer learning (TL) is a learning mechanism emerging from the concept of “model” sharing that can potentially improve the learner from one domain by transferring information (in terms of a pre-trained model) from a related domain where there is a limited supply of target training data\textsuperscript{11,12}. The key conceptual difference between FL and TL is that FL sequentially trains a set of independent hospital datasets to build a decentralized model, whereas, in TL, any target hospital site can build a model using a transferred pre-trained model and re-train the model using its own data for local adaptation. To the best of our knowledge, this study is the first of its kind to implement transfer learning to address the problem of limited availability of data for out-of-network siloed sites. Furthermore, we choose our target disease to be diabetic kidney disease, a long-term complication of chronic diabetes and can result in increased costs related to hospitalizations, medicines, and treatment procedures.

The main objective of this study is to design a privacy-preserved transfer learning architecture that can utilize the knowledge from a privacy-preserved federated model developed from in-network hospital-site EHR data for predicting diabetic kidney cases at out-of-network siloed hospital sites. In addition, we aim to demonstrate the potential and scope of transferring knowledge from established data-sharing hospital networks to aid data-driven AI discoveries at siloed healthcare entities without sharing of data.

The key contributions of our study include: (i) introducing the concept of knowledge sharing from in-network hospitals to out-of-network hospitals for building predictive AI models, (ii) utilize an extensive database of real-world healthcare data called Health Facts to demonstrate the proof-of-concept of transfer learning, (iii) implementing decentralized privacy-preserved federated learning architecture using in-network hospital data to facilitate a privacy-preserved transfer learning process to predict diabetic kidney disease at out-of-network siloed hospitals, (iv) comparing the proposed transfer learning mechanism to cases where only siloed data at out-of-network hospitals are used to predict the diabetic kidney cases to manifest the significance of our approach.

**Background**

Transfer learning involves improving the target predictive function by using the knowledge from the source domain and the data from the target domain. Transfer learning applications in machine learning include text sentiment classification, image classification, software detect classification, and multi-language text classification\textsuperscript{11–16}. However, limited research exists in transfer machine learning adaptations in the healthcare domain. Recently, a study by Gao et al. (2019)\textsuperscript{17} demonstrated a transfer learning approach on a case study on the MIMIC-III database to predict in-hospital mortality. Our study aims to illustrate the adaptation of transfer learning concepts for predicting diabetic kidney disease at siloed data sites, which is the first of its kind to the best of our knowledge. Moreover, we set up the experimental design using natural partitions of healthcare institutions in the Health Facts database to mimic the real-world setting to our best.

Federated learning involves sharing only the mathematical parameters, not the actual data itself, to build a global model iteratively over independent databases\textsuperscript{18}. The earlier efforts of adapting federated learning models in healthcare include predicting mortality, hospital stay-time for ICU patients, cardiac event hospitalization, dyspnea, adverse drug reactions, diabetes-related complications\textsuperscript{19–23}. Most of the federated learning healthcare applications applied popular machine learning algorithms such as logistic regression, artificial neural network, and random forest\textsuperscript{21–24,26}. Our previous study\textsuperscript{23} demonstrated the utilization of federated learning architecture for binary classification of the incidence of three diabetes-related complications affecting eyes, kidneys, and peripheral nerves, respectively, using logistic regression and simple artificial neural networks. We observed comparable performance for the federated learning models to the gold standard of centralized learning on a central database. This motivated us to implement a federated learning architecture to build a predictive model for diabetic kidney disease using data from independent healthcare systems included in a network. In addition, we prioritized the importance of privacy-preserved knowledge transfer for predicting diabetic kidney disease at the siloed data sites, which encouraged us to use the federated models to demonstrate our transfer learning approach.
Methods

Data Source
In this study, we used Cerner’s “Health Facts EMR Data,” a de-identified electronic health records database consolidated from over 90 healthcare systems across the US between 2000 and 2016. This database contains demographics, encounters, diagnoses, lab results, procedures, prescriptions, and other clinical attributes for about 69 million unique patients.

Diabetic Kidney Disease Cohort Selection
A diabetes population was identified using the Surveillance, PREvention, and ManagEment of Diabetes Mellitus algorithm (SUPREME-DM) based on eight criteria, two of them were based on International Classification of Disease (ICD-9 and ICD-10) diagnosis codes related to inpatient and outpatient encounters and six were based on lab results (Figure 1). Only the patients over 18 years who satisfied at least one or more were selected in the diabetes population. In addition, a patient cohort with diabetic kidney disease (DKD) was identified using ICD-9 and ICD-10 diagnoses codes (250.4x, E10.2x, and E11.2x) from the selected diabetes population. In this study, we considered a binary classification of diabetic kidney disease. The detail of the population selection process is described in our previous study.

Feature Selection
We used the diagnosis table from the Health Facts database to extract the patient-related comorbid features. Next, we mapped the ICD codes to Clinical Classification Software (CSS) tool developed by Healthcare Cost and Utilization Project (HCUP). Finally, by grouping the individual ICD-9 and ICD-10 codes into similar clinical entities, we extracted a total of 283 unique CCS coded features to predict the complication as binary responses for our experiment.

Experimental Data Architecture
To demonstrate the method of transfer learning on siloed healthcare systems from federated models consolidated using in-network hospitals, we utilized the existing hospital identifiers for each patient in our DKD cohort. The data for our DKD cohort were partitioned into the different hospital sites based on the identifiers to facilitate the federated architecture. However, we discarded the hospital sites with less than 100 cases from our analysis. Assuming that hospitals with a larger patient population are part of a bigger network, we considered any hospital with a population over 1900 as in-network sites, while the others as out-of-network siloed sites. In total, 15 sites were included as in-network sites (referred to as vanguard sites), and 16 were considered out-of-network siloed sites (referred to as siloed sites).

Federated Model for In-Network Hospitals
Using a federated learning approach, we utilized the in-network 15 partitioned sites to build a decentralized model to predict the binary DKD cases and non-cases. We call this model our vanguard FL model and the sites as vanguard sites. To develop our vanguard FL model, 70% of each of the 15 partitioned sites were used as training datasets, while 30% were combined to build a common testing dataset. We used Python 3.6 libraries sci-kit learn, NumPy, pandas, and TensorFlow to develop our machine learning approach. In addition, we used 1-hot-encoding for our labels in the training set and transformed the features into TensorFlow data objects. Finally, we compiled a logistic regression and a 3-layer multi-perceptron model using binary cross-entropy as loss function and stochastic gradient descent as the optimizer.
The training module for federated learning was developed using the federated averaging algorithm\textsuperscript{18}, as demonstrated in our previous study including all 31 healthcare entities\textsuperscript{23}. The federated training approach begins with initiating a global model, which serves the initial weights for all the \textit{vanguard} sites. Next, the \textit{vanguard} sites use the initial weights to train their data and obtain the updated weights. The updated weights from each \textit{vanguard} site are sent to the global model without sharing any raw data. A weighted average of the \textit{vanguard} weights is then used to update the global model. Finally, after many rounds of aggregating \textit{vanguard} weights to update the global model, we save the final global model (the \textit{vanguard FL} model) and its weights for further analysis. (Algorithm 1 STEP 1)

The \textit{Transfer Learning} Approach

The \textit{siloed} hospital site that is not a part of a big network of hospitals can utilize the knowledge of the network to build predictive models by adopting a transfer learning approach. As knowledge is transferred from the in-network sites to the out-of-network sites, this process can be referred to as transfer learning. Generally, transfer learning is expressed using a pre-trained model, a model trained on a large dataset to predict a similar outcome. The out-of-network \textit{siloed} sites can utilize the knowledge from the pre-trained model, which is our \textit{vanguard FL} model, to predict DKD cases without data leaving its door. We adopted two mechanisms of transfer learning to build predictive models for the \textit{siloed} sites based on the knowledge of a \textit{vanguard FL} model\textsuperscript{28}. In mechanism A, the \textit{siloed} site trains the \textit{vanguard FL} model using its training weights with the \textit{siloed} dataset, assuming all the model layers as trainable. In mechanism B, the layers of the \textit{vanguard FL} model are assumed as non-trainable with frozen weights, and new additional layers are built on top of that. The required steps of the two mechanisms are presented in \textbf{STEP 2 and 3} of Algorithm 1. Figure 3 presents the entire experimental architecture of our study design.

\textbf{Algorithm 1:} The Algorithm moves from STEP 1 to either STEP 2 or 3

<table>
<thead>
<tr>
<th><strong>STEP 1:</strong> Federated Learning Approach to Build the \textit{vanguard FL} Model (VGM) for Predicting Diabetic Kidney Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Require:</strong> Datasets ($D_{k}$) where $k = {1, ..., K}$ of $K$ \textit{vanguard} sites ($V_{1}, ..., V_{K}$) with $n_{k}$ sample size</td>
</tr>
<tr>
<td>1: Initiate a VGM with known weights: $\omega_{0}$</td>
</tr>
<tr>
<td>2: Each site ($V_{k}$) receives the \textit{vanguard} initial weights $\omega_{0}$, compiles the VGM with local data ($D_{k}$), sends updated local weights ($\omega_{k}$) to VGM</td>
</tr>
<tr>
<td>3: The local weights are combined to compute the weighted average $\omega = \frac{\sum_{k=1}^{K} n_{k}}{N} \ast \omega_{k}$</td>
</tr>
<tr>
<td>4: VGM is updated with new weights $\omega$</td>
</tr>
<tr>
<td>5: This process continues for $T$ rounds and the final VGM with training weights $\omega_{VGM}$ is saved</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>STEP 2:</strong> Transfer Learning on Out-of-Network \textit{siloed} Sites using Mechanism A</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Require:</strong> Datasets ($Q_{l}$) where $l = {1, ..., L}$ of $L$ Out-of-Network \textit{siloed} Sites ($S_{l_{1}}, ..., S_{l_{L}}$) AND VGM with $\omega_{VGM}$ from \textbf{STEP 1.5}</td>
</tr>
<tr>
<td>6: for $l = {1, ..., L}$:</td>
</tr>
<tr>
<td>7: Load VGM with $\omega_{VGM}$ as transferred knowledge</td>
</tr>
<tr>
<td>8: Create training ($Q_{l_{-TRAIN}}$) and testing ($Q_{l_{-TEST}}$) datasets from locally \textit{siloed} ($Q_{l}$)</td>
</tr>
<tr>
<td>9: Using $\omega_{VGM}$ as pre-training weights, train VGM with training dataset ($Q_{l_{-TRAIN}}$)</td>
</tr>
<tr>
<td>10: Test the new trained model on ($Q_{l_{-TEST}}$)</td>
</tr>
<tr>
<td>11: Compute performance metrics: F-1 score, precision, and recall</td>
</tr>
<tr>
<td>12: end for</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>STEP 3:</strong> Transfer Learning on Out-of-Network \textit{siloed} Sites using Mechanism B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Require:</strong> Datasets ($Q_{l}$) where $l = {1, ..., L}$ of $L$ Out-of-Network \textit{siloed} Sites ($S_{l_{1}}, ..., S_{l_{L}}$) AND VGM with $\omega_{VGM}$ from \textbf{STEP 1.5}</td>
</tr>
<tr>
<td>13: for $l = {1, ..., L}$:</td>
</tr>
<tr>
<td>14: Load VGM with $\omega_{VGM}$ as transferred knowledge</td>
</tr>
<tr>
<td>15: Create training ($Q_{l_{-TRAIN}}$) and testing ($Q_{l_{-TEST}}$) datasets from locally \textit{siloed} ($Q_{l}$)</td>
</tr>
<tr>
<td>16: Convert the layers of VGM to non-trainable layers with frozen weights</td>
</tr>
<tr>
<td>17: Add new trainable layers over VGM</td>
</tr>
<tr>
<td>18: Using $\omega_{VGM}$ as pre-training weights, train the new model from \textbf{STEP 3.16} and 3.17 with training dataset ($Q_{l_{-TRAIN}}$)</td>
</tr>
<tr>
<td>19: Test the new trained model on ($Q_{l_{-TEST}}$)</td>
</tr>
<tr>
<td>20: Compute performance metrics: F-1 score, precision, and recall</td>
</tr>
<tr>
<td>21: end for</td>
</tr>
</tbody>
</table>

\textbf{Experimentation for Out-of-Network \textit{siloed} Sites}

To demonstrate the feasibility of the transfer learning approach, we created training (70\%) and testing (30\%) datasets for each of the 16 \textit{siloed} sites and used the same train-test datasets to conduct the following comparative experiments.

\textbf{CASE I (LOCAL): LOCAL MODEL VALIDATION}

We trained individual CCS feature-data silos to build models to predict DKD cases among the patients for 16 \textit{siloed} sites. For this purpose, we implemented both logistic regression and 3-layer multi-perceptron models using the sci-kit
learn module from Python 3.6. The local model validation mimics the case where the siloed sites are deprived of any existing knowledge to build predictive models.

**CASE II (VANGUARD): VANGUARD MODEL VALIDATION**

This situation mimics the case where the vanguard FL models are available only for testing on the local siloed sites. Therefore, the local test datasets are used to evaluate the predictive performance of the vanguard FL models without any training on the local data. For both the vanguard FL models (logistic and MLP), we tested the performance for all the 16 siloed sites.

**CASE III (TRANSFERRED A or B): TRANSFER LEARNING APPROACH**

This case mimics the condition where knowledge from the vanguard network is agreed to be transferred to the local siloed site. For our proposed approach, we assumed that the vanguard FL model and its weights could be transferred from the in-network sites to the local sites for further learning. The local siloed sites will then train the vanguard FL model with its local training data silos and evaluate the performance on its test datasets. We applied mechanisms A and B (as shown in Algorithm 1 and Figure 3) to build the MLP model, while logistic regression only applies mechanism A. For our MLP model, we froze the two layers from the vanguard FL model and added two more dense layers with activation function “relu” alongside the output layer with activation “sigmoid”. The new MLP model is finally compiled using stochastic gradient descent as the optimizer and binary cross-entropy as the loss function and trained with the pre-training weights from the vanguard FL model. For both MLP and logistic regression, we tested the performance across all the 16 siloed sites.

### Class Imbalance Learning and Evaluation Metrics

The federated datasets for both vanguard sites and local siloed sites are subjected to an unequal class balance for DKD cases and non-cases (See Figure 2 for the class-ratio distribution in Islam (2021)\(^2\)). To account for the varying class distribution in our model training, all our experiments were repeated for sampling techniques, such as oversampling, which supplements the minority class, and undersampling, which randomly removes the majority class. Thus, for all our experiments, oversampling, undersampling, and no-sampling were performed before model compilation. We used the module “resample” from sk-learn to apply the class balancing techniques.

For each vanguard FL model, we ran 48 experiments for the logistic regression case and 64 experiments for the multi-perceptron model. With the class-balancing experiment, we built three vanguard FL models using logistic regression and three using multi-perceptron. In total, we ran 336 experiments for our analysis. Since F-1 scores are a more reliable measure than accuracy in the presence of class imbalance, performance metrics, such as F-1 score, recall, and precision, were computed. We weighted the three-performance metrics by the sample size of the local siloed sites to obtain a weighted average with a 95% confidence interval to compare the performance among the different case scenarios of our comparative experiments.

R version 3.4.4 (R Foundation for Statistical Computing, Vienna, Austria) and Python 3.6 were used for data management. All computations were performed on a Mac Book Pro running macOS Catalina version 10.15.2 with 16GB of RAM.

**Results**

In our analysis, we extracted a sample of 17,455 patients as our diabetic kidney cohort from the Health Facts database with 102,876 population size for diabetes patients. In total, there were 31 independent healthcare systems, and we considered a network of 15 hospitals as our vanguard sites and 16 local out-of-network siloed sites. Figure 2 shows the class-imbalance issue in the datasets by the varying sample size with the number of cases for both the in-network and out-of-network sites. The sample size of the vanguard sites which form the in-network hospitals vary from 1974 to 9386, and the number of cases ranges from 133 to 2096. The siloed local sites have a low as 239 patient sample size.

Figure 2: A diagrammatical representation of class imbalance as shown by the varying sample size and the number of cases for vanguard and siloed sites. The vanguard sites form the larger hospital network, while the siloed sites represent the local out-of-network hospitals.
Table 1 shows the performance for the vanguard FL model for the logistic regression and multi-perceptron models. Both oversampling and undersampling showed better recall performance compared to the case of without sampling case. These models served as our knowledge base for demonstrating the transfer learning approach for the siloed sites.

<table>
<thead>
<tr>
<th>Model</th>
<th>Sampling Method</th>
<th>F1-Score</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>Oversample</td>
<td>0.64</td>
<td>0.55</td>
<td>0.85</td>
</tr>
<tr>
<td>LR</td>
<td>Undersample</td>
<td>0.63</td>
<td>0.53</td>
<td>0.85</td>
</tr>
<tr>
<td>LR</td>
<td>No sample</td>
<td>0.58</td>
<td>0.73</td>
<td>0.52</td>
</tr>
<tr>
<td>MLP (10,10,1)</td>
<td>Oversample</td>
<td>0.67</td>
<td>0.59</td>
<td>0.83</td>
</tr>
<tr>
<td>MLP (10,10,1)</td>
<td>Undersample</td>
<td>0.64</td>
<td>0.55</td>
<td>0.86</td>
</tr>
<tr>
<td>MLP (10,10,1)</td>
<td>No sample</td>
<td>0.65</td>
<td>0.68</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Figure 3: An infographic representation of our proposed transfer learning approach for the out-of-network siloed hospital sites. The decentralized vanguard FL model is built based on the federated learning approach shown in Algorithm 1: STEP 1 with no sharing of actual data among the vanguard sites. Thus, the out-of-network siloed site can utilize the knowledge base from the vanguard FL model to predict the DKD cases without being part of the vanguard network and without sharing data outside the hospital site. We refer to this process of conveying knowledge as transfer learning. The two mechanisms of transfer learning are shown in Algorithm 1: STEP 2 and 3, respectively.

The comparative evaluation for all the experiments under multi-perceptron and logistic regression models are presented in Figure 4 and Figure 5, respectively. Appendix A shows the combined results for the weighted means.
from all the experiments. The weighted means show a comprehensive measure of the performance metrics; hence, we use the weighted means to compare the different approaches and experiments. For our multi-perceptron model, both TRANSFERRED A and B approaches consistently showed improved predictive performances compared to LOCAL validation and VANGUARD validation, as shown by the trend in the weighted means in Figure 4. Under oversampling, TRANSFERRED A showed a 17.9% increase in weighted F-1 score, 25.4% in the weighted recall, and 13% in weighted precision compared to LOCAL. TRANSFERRED B showed a 21.4% percent increase in weighted F-1 score, 10% in the weighted recall, and 28% in weighted precision compared to LOCAL. We observe similar improvements in the measures for undersampling with 27.8% percent improvement in weighted F1 (41.9% in weighted precision and 3.3% in weighted recall) for TRANSFERRED A, while 22.2% in weighted F1 (28.2% in weighted precision and 10% in weighted recall) for TRANSFERRED B. However, in the case of no sampling, the percentage improvement in weighted F-1 score was much less compared to the above two cases with only a 14.3% increase for TRANSFERRED B. Moreover, we observe a decrease in percentage in weighted precision value for no sampling case by about 1.2% for TRANSFERRED B.

![Figure 4: A figure showing the comparative evaluation for CASE 1 (LOCAL), CASE II (VANGUARD), and CASE III (TRANSFERRED A and B) under the multi-perceptron model. TRANSFERRED A considers the layers of the federated vanguard model as trainable. In contrast, in TRANSFERRED B, the federated vanguard layers are frozen, and new trainable layers are added to the top of that. Different colors and symbols represent the four cases. The performance metrics F-1 scores, Precision and Recall are presented with their weighted mean (red diamond) and 95% confidence interval (error bar) for three sampling cases: OVERSAMPLING, UNDERSAMPLING, and NO SAMPLING. The performance for the vanguard federated model (knowledge base) is indicated with grey dashed lines. The weighted means show a consistent improvement for either TRANSFERRED cases compared to the LOCAL case across all the experiments. Additionally, the VANGUARD validation on siloed sites showed performance improvement compared to LOCAL. TRANSFERRED B approach showed improvement over VANGUARD validation (weighted F1 ranging from 1.5 to 6.3%, weighted precision from 1.3 to 7.7%, weighted recall from 1.5 to 8.4%); however, TRANSFERRED A showed a percentage decrease in weighted precision values compared to VANGUARD validation (-6.6% for oversampling and -4.7% for undersampling). In contrast, the weighted recall and F-1 scores showed performance improvements for TRANSFERRED A compared to VANGUARD validation. Furthermore, the weighted recall values for TRANSFERRED B for siloed sites are much closer to the vanguard FL recall value for both the over-and-undersampling cases, as shown in Figure 4. This shows TRANSFERRED B performed consistently well in predicting diabetic kidney disease for the siloed sites.]

270
Similarly, the TRANSFERRED A approach showed a consistent improvement in the performance metrics compared to LOCAL and VANGUARD validation cases for the logistic regression model. The weighted means for F-1, precision, and recall show an upward trend for all the sampling cases, as shown in Figure 5. The recall values for siloed sites are much closer to the vanguard FL performance in the oversampling case. In contrast, most of the precision values and weighted values exceeded the vanguard FL performance level in the undersampling case. Overall, we observe transfer learning to produce comparably better and consistent predictive results than the other cases.

![Figure 5](image)

Figure 5: A figure showing the comparative evaluation for CASE 1 (LOCAL), CASE II (VANGUARD), and CASE III (TRANSFERRED A) under the logistic regression model. Different colors and symbols represent the three cases. The performance metrics F-1 scores, Precision and Recall are presented with their weighted mean (red diamond) and 95% confidence interval (error bar) for three sampling cases: OVERSAMPLING, UNDERSAMPLING, and NO SAMPLING. The performance for the vanguard FL model (knowledge base) is indicated with grey dashed lines. The weighted means show consistent improvement for the TRANSFERRED case compared to the LOCAL across all the experiments.

**Discussion**

The implementation of transfer learning can bring numerous opportunities to investigate key healthcare issues and clinical events when only limited knowledge is available. This paper demonstrates the feasibility of creating data-driven AI models for healthcare institutions with limited access to big databases using a knowledge base from a network of hospitals by a transfer learning mechanism. We acknowledge the importance of preserving privacy, data sharing or transportation challenges to build a central repository, and the lack of data accessibility at out-of-network sites. Thus, through this study, we emphasized the significance of “no data leaving the firewalls of the hospitals” by incorporating the concepts of federated learning using a decentralized AI model building network for the in-network hospital sites. We also demonstrate that knowledge transferred from the privacy-preserved federated AI model in terms of “model transfer” can enhance data-driven AI applications at out-of-network hospital sites.

We used Cerner’s Health Facts database containing the independent healthcare system identifiers to facilitate our federated architecture, hence assigning the hospital sites either in-network or out-of-network. We assumed that smaller patient samples are considered out-of-network sites, although the proportion of cases could be higher in them. We demonstrated the process of federated learning and transfer learning using an artificial neural network (multi-layer perceptron) and a simple case of logistic regression. We compared the performance of the transferred approaches with the case where only local data could be trained to build the predictive model. As for performance measures, we computed F-1, precision, and recall. However, since we are predicting DKD cases, maximizing the chances of true positives is more relevant. Thus, the recall values could provide better insights into model performance. Additionally, we scaled the performance metrics with the individual sample size of the out-of-network sites to compute a weighted
Conclusion
In conclusion, our results prove the feasibility of transfer learning in building AI models for out-of-network sites using a pre-trained model built from a decentralized data sharing network architecture. We present this study as a proof-of-concept of transfer learning using privacy-preserved knowledge developed through federated learning. However, our model architecture is limited to only a case of artificial neural network and logistic regression, which can generate reasonably good but not high-performance models. Also, we assumed the feature set to be similar for all our experiments, which do not reflect the reality. We will explore and experiment further on the loss of accuracy from transfer learning and compare the characteristics of the datasets such as class ratio with the model performance for further illustration of our method. Our future work extends to include more classifiers and deep learning models to improve the performance of learning. We also plan to explore heterogeneous feature and distribution space for transfer learning, averaging techniques for federated learning, machine learning methods to address class-imbalance issues, fine-tuning of the models, extract more patient-related information to build a more reliable and robust predictive model for diabetic kidney disease and other complications of diabetes. Our future work will include more extensive research on making the transfer learning concept more feasible and acceptable for healthcare research advancements.

Glossary terms:
Vanguard sites: The hospital sites with population over 1900 forming a network
Siloed sites: The out-of-network hospital sites with population less than 1900

References
(Association for Computational Linguistics, 2010).

**APPENDIX A**

A table showing the combined results of the experiments for the weighted mean of the performance measures (F-1, precision, and recall) with variability measures such as standard deviation (STD) and 95% confidence interval.

<table>
<thead>
<tr>
<th>Models</th>
<th>Metrics</th>
<th>SAMPLING: NO SAMPLING</th>
<th>SAMPLING: OVERSAMPLING</th>
<th>SAMPLING: UNDERSAMPLING</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LOCAL</td>
<td>Vanguard</td>
<td>Transferred A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PREC</td>
<td>REC</td>
<td>PREC</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>WM ± SD</td>
<td>0.48 ± 0.21</td>
<td>0.67 ± 0.2</td>
<td>0.41 ± 0.22</td>
</tr>
<tr>
<td></td>
<td>[LL, UL]</td>
<td>[0.37, 0.58]</td>
<td>[0.57, 0.77]</td>
<td>[0.31, 0.52]</td>
</tr>
<tr>
<td>Multi-layer Perception</td>
<td>WM ± SD</td>
<td>0.56 ± 0.13</td>
<td>0.68 ± 0.11</td>
<td>0.53 ± 0.11</td>
</tr>
<tr>
<td></td>
<td>[LL, UL]</td>
<td>[0.5, 0.63]</td>
<td>[0.62, 0.73]</td>
<td>[0.46, 0.68]</td>
</tr>
</tbody>
</table>

273
Moving from predicting hospital deaths by antibiotic-resistant bloodstream bacteremia toward actionable risk reduction using machine learning on electronic health records

Inyoung Jun, M.S.1, Shannan N. Rich, M.S.1, Simone Marini, Ph.D.1, Zheng Feng, Ph.D.2, Jiang Bian, Ph.D.2, J. Glenn Morris, M.D., M.P.H., T.M.3, Mattia Prosperi, Ph.D.1

1Department of Epidemiology, College of Public Health and Health Professions & College of Medicine, University of Florida, Gainesville, FL, U.S.A;
2Department of Health Outcomes & Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, U.S.A; 3Emerging Pathogens Institute and Department of Environmental & Global Health, College of Public Health and Health Professions, University of Florida, Gainesville, FL, U.S.A.

Abstract

Drug-resistant bacterial infections are a global health concern with high mortality and limited treatment options. Several clinical risk-severity scores are available, e.g. qPitt, but their predictive performance is moderate. Here, we leveraged machine learning and electronic health records (EHRs) to improve prediction of mortality due to bloodstream infection with Klebsiella pneumoniae. We tested the qPitt score and new EHR variables (either expert-chosen or the full set of diagnostic codes), fitting LASSO, boosted logistic regression (BLR), support vector machines, decision trees, and random forests. The qPitt score showed moderate discriminative ability (AUROC=0.63), whilst machine learning models significantly improved its performance (best AUROC by BLR 0.80 for expert-chosen and 0.88 for full code set). Similar results were obtained in critically ill patients, and when excluding potential non-causal variables to evaluate an actionable model. In conclusion, current risk scores for bacteremia mortality can be improved and, with opportune causal modelling, considered for deployment in clinical decision-making.

Introduction

Drug-resistant bacterial infections are a global health concern with high mortality and limited treatment options. Antimicrobial stewardship interventions are purposed to minimize the emergence of antimicrobial resistance in hospital settings, and to ensure that patients receive the right antimicrobial agent, at the right time, with the right dose. Mortality due to bloodstream infections (BSIs) caused by drug-resistant bacteria is a common metric used to evaluate the impact of antimicrobial stewardship interventions. Researchers have been investigating risk factors for BSI and the increasingly important pathogen Klebsiella pneumoniae in particular. Klebsiella pneumoniae (KP) is a gram-negative bacterium of the Enterobacteriaceae family. KP is notorious for producing extended-spectrum beta-lactamases (ESBL), conferring resistance to many commonly used beta lactam antibiotics, including penicillins and cephalosporins, leaving few available treatment options except carbapenem antibiotics. KP can also become resistant to carbapenem antibiotics by acquiring additional genes such as those that encode for Carbapenemases. BSIs with drug-resistant KP are particularly problematic for patients in the intensive care unit (ICU) due to the limited treatment options and the critical conditions.

With the limited treatment options, predicting mortality risk of BSI-KP is important for minimizing unfavorable health outcomes. To date, a number of clinical scores that assess disease severity and mortality risk are available, including the widely used Pitt Bacteremia Score (PBS) which predicts mortality in nonbacteremic infections. A modified version of the PBS, the quick Pitt (qPitt) bacteremia score, is used to assess severity of illness and mortality in patients with gram-negative BSI. Previous studies have shown that qPitt outperforms other clinical scores, such as the quick Sepsis Related Organ Failure Assessment (qSOFA) and the Systemic Inflammatory Response Syndrome (SIRS), in predicting mortality among patients with gram-negative BSI.

In addition to patients’ clinical status used for computing the referenced clinical scores, increasing research suggests that patients’ electronic health records (EHR) can be leveraged to improve prediction of certain health outcomes. A recent study found that certain microbiological characteristics could be risk factors of 30-day mortality for KP bacteremia such as lower platelet count and microbiologic eradication of more than 7 days. In addition, prior
antibiotic exposure has been known to be associated with drug-resistant bacterial infections. In this study, we utilize a large integrated EHR database and apply linear/nonlinear machine learning to develop a new prediction model of BSI-KP mortality for the general hospital wards, and for the ICU. We show that the predictive performance of the qPitt score can be improved by a fair margin through screening and inclusion of additional patient-level EHR variables, combining clinical and laboratory (including antibiogram test results) data, to derive an interpretable model usable in routine clinical settings.

**Materials and Method**

**Data source and study population**

We used the University of Florida (UF) Integrated Data Repository (IDR), which collates EHR data—demographics, clinical, laboratory, and medications—from multiple clinics under the UF Health provider, Florida, USA. IDR adopts the International Classification of Disease ontology (ICD) ver. 9 and 10 for clinical diagnoses and procedures, RxNorm for medications, and Logical Observation Identifiers Names and Codes (LOINC) for laboratory data. The EHRs in UF Health are stored according to the ontology that was in use at the time the data were collected. Therefore, our dataset included ICD9 as well as ICD10 codes. Since the majority of codes were ICD9, we converted ICD10 codes into ICD9 using the Centers for Medicare & Medicaid Services mapping guidelines.

Our study considered UF Health adult patients (18+ years) admitted either inpatient or outpatient clinics and diagnosed with BSI (ICD-9 790.7, 995.91, 995.92, 038.xx). Only patients who had at least two years of medical history and who had an antibiogram (i.e., drug-resistance test) result within 7 days upon BSI-KP diagnosis were included. **Figure 1** shows a flowchart of the study design and inclusion criteria.

<table>
<thead>
<tr>
<th>Bloodstream infection patients in 2011-2018 (n=14,106)</th>
<th>BSI patients with <em>Klebsiella Pneumoniae</em> antibiogram test results (n=1,892)</th>
<th>Patients age ≥18 at the diagnosis date (n=1,888)</th>
<th>Patients having more than two year of medical history before the diagnosis date (n=1,389)</th>
</tr>
</thead>
</table>

**Figure 1.** Flowchart of the study design and inclusion criteria.

**Study outcome and independent predictors**

The outcome variable was 30-day mortality following the date of BSI-KP diagnosis. Covariates were measured prior or at BSI-KP diagnosis. Expert-chosen variables—selected by the co-authors and supported by literature evidence—included: age (years), sex, race/ethnicity, nosocomial infection, ESBL positivity, weighted Charlson’s comorbidity index (CCI), other comorbidities (acute myocardial infarction, congestive heart failure, peripheral vascular disease, cerebrovascular disease, dementia, chronic obstructive pulmonary disease, rheumatoid disease, peptic ulcer disease, mild liver disease, diabetes, hemiplegia/paraplegia, renal disease, cancer, liver disease, metastatic solid tumor, HIV/AIDS), qPitt bacteremia score, prescription record of antibiotics before and upon diagnosis, most frequently prescribed medications before and upon diagnosis, and antibiogram tests results for *Klebsiella* (all species) or other organisms before and upon diagnosis. In addition to these variables, we also extracted all ICD codes (five-digit granularity) present with at least 10% frequency.

**Statistical analyses and model development**

We summarized the characteristics of the study population by stratifying on inpatient/outpatient status. We further estimated prevalence of antibiotic resistance and of multi-drug resistance (MDR) from the antibiogram test performed before or upon the BSI-KP diagnosis stratified by *Klebsiella* species and other organisms. We comprehensively measured resistance not only for KP but also for other members of the Enterobacteriaceae family because antibiotic-resistance genes can be transferred among bacteria, and many bacteria share similar resistance mechanisms. We ran univariate analysis on the 30-day mortality using qPitt, and multivariable analyses combining qPitt with other variables. For developing the prediction model (using either expert-chosen or full variable input sets), we considered both the whole population and the subset of patients admitted to the ICU. We also evaluated the models by excluding putative non-causal variables. The predictive performance of the qPitt score was compared with that of the following machine learning approaches: LASSO and boosted logistic regression (linear, main effects); support vector machine, decision tree, and random forest (nonlinear). We fit the support vector machine with either a linear or a radial basis
kernel, and tuned its parameters within the bootstrapping using a grid search on the cost (between 0.001 and 10) and the gamma parameter (between $2^{-8}$ and $2^4$). Performance measures included area under the receiver operating characteristic (AUROC), sensitivity, and specificity, assessed via out-of-bag bootstrapping (25 samples), and compared with Bengio’s and Nadeau’s corrected t-test. For further evaluating model generalizability, since UF Health includes EHR data from different clinics, we emulated external validation by setting aside data from clinics in one county-wide zip code area (three-digits), training models on all remaining data points, and testing them on the removed set.

**Results**

**Clinical characteristics of the study population**

The total annual number of patients seen at UF Health between 2011 and 2018 ranged from 25,898 to 32,621. Upon cohort filtering, 1,398 adult patients with BSI-KP were included in our study population. **Figure 2** shows the frequency of BSI-KP infections, 30-days deaths, and MDR in relation to the total number of patients. The study population increased from 2011 to 2014 and then showed a small decline between 2015 and 2018. The trends of 30-day mortality and MDR KP were stable over time. The total number of distinct diagnostic codes above the set frequency was 266.

**Figure 2.** Frequency of *K. pneumoniae* bacteremia infections, 30-day mortality, and multi-drug resistance (MDR) in relation to the total number of patients at UF Health during 2011-2018

The inpatient population was in general older than the outpatient population, male, and of non-white race (Table 1). Within the inpatient population, 50.7% were transferred to the ICU. The median CCI was 4.0 in both groups but the maximum was higher in the inpatient group. The outpatient group had a slightly higher prevalence of acute myocardial infarction and a lower percentage of chronic obstructive pulmonary disease and renal disease, as compared to the inpatient group. The mean (Standard Deviation, SD) qPitt score in the inpatient group was 0.427 (0.682), while it was 0.220 (0.509) in the outpatient group.

**Table 1.** Main characteristics of patients with KP-BSI, shown as N (%), Mean (SD), or Median [Min, Max].

<table>
<thead>
<tr>
<th>Variable</th>
<th>Inpatient (N=1,271)</th>
<th>Outpatient (N=118)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years) at diagnosis</td>
<td>59.3 (16.7)</td>
<td>58.4 (16.6)</td>
</tr>
<tr>
<td>Sex – Male</td>
<td>613 (48.2%)</td>
<td>52 (44.1%)</td>
</tr>
<tr>
<td>Race – White</td>
<td>806 (63.4%)</td>
<td>77 (65.3%)</td>
</tr>
<tr>
<td>Race – Black/Other</td>
<td>465 (36.6%)</td>
<td>41 (34.8%)</td>
</tr>
<tr>
<td>Infection source – blood</td>
<td>397 (31.2%)</td>
<td>35 (29.7%)</td>
</tr>
<tr>
<td>Infection source – urine</td>
<td>529 (41.6%)</td>
<td>28 (23.7%)</td>
</tr>
<tr>
<td>Intensive Care Unit stay</td>
<td>645 (50.7%)</td>
<td>N/A</td>
</tr>
</tbody>
</table>

276
Charlson’s comorbidity index

4.00 [0, 19.0] 4.00 [0, 14.0]

Acute myocardial infarction

185 (14.6%) 18 (15.3%)

Chronic obstructive pulmonary disease

511 (40.2%) 40 (33.9%)

Renal disease

405 (31.9%) 32 (27.1%)

qPitt Score

0.427 (0.682) 0.220 (0.509)

30-day mortality

183 (14.4%) 6 (5.1%)

30-day mortality in intensive care unit

143/645 (22.8%) N/A

**History of medication and antimicrobial resistance**

**Table 2** summarizes the most frequent medications used and antimicrobial resistance results in the inpatient and outpatient populations before and upon BSI-KP diagnosis. Both opioids and acetaminophen were prescribed more often to inpatients. Vancomycin was the most frequently used antibiotic in both the inpatient and outpatient populations. Ceftriaxone, Piperacillin, and Cefazolin were frequently used in the inpatient group while the outpatient population showed less usage of these three antibiotics.

The prevalence of antimicrobial resistance in *Klebsiella* species increased significantly upon BSI-KP diagnosis for all categories of antibiotics, especially in the inpatient group (*Table 2*). Overall, 94.7% of inpatients harbored a *Klebsiella* species resistant to Beta-lactamases (34.5% before the infection and 92.5% upon the infection). Resistance to aminoglycosides was slightly higher in the outpatient population before BSI-KP, while more resistance appeared upon BSI-KP in the inpatients.

**Table 2.** Comparison of medication usage and antimicrobial resistance in the study population before/upon BSI-KP diagnosis.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Inpatient (N=1,271)</th>
<th>Outpatient (N=118)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before infection</td>
<td>Upon infection</td>
</tr>
<tr>
<td>Acetaminophen</td>
<td>763 (60.0%)</td>
<td>995 (78.3%)</td>
</tr>
<tr>
<td>Opioids</td>
<td>1,130 (91.7%)</td>
<td>133 (79.2%)</td>
</tr>
<tr>
<td>Vancomycin</td>
<td>1,204 (94.7%)</td>
<td>45 (38.1%)</td>
</tr>
<tr>
<td>Carbapenem</td>
<td>1,260 (99.2%)</td>
<td>18 (15.3%)</td>
</tr>
<tr>
<td>Cefazolin</td>
<td>1,176 (92.5%)</td>
<td>41 (3.2%)</td>
</tr>
<tr>
<td>Ceftriaxone</td>
<td>1,204 (94.7%)</td>
<td>45 (38.1%)</td>
</tr>
<tr>
<td>Piperacillin</td>
<td>1,580 (12.4%)</td>
<td>11 (9.3%)</td>
</tr>
<tr>
<td>Fluoroquinolones</td>
<td>1,580 (12.4%)</td>
<td>11 (9.3%)</td>
</tr>
</tbody>
</table>
Sulfonamides | 120 (9.4%) | 264 (20.8%) | 300 (23.6%) | 13 (11.0%) | 15 (12.7%) | 23 (19.5%)  
---|---|---|---|---|---|---  
Others | 96 (7.6%) | 168 (13.2%) | 217 (17.1%) | 10 (8.5%) | 8 (6.8%) | 16 (13.6%)  

**Prediction of 30-day mortality from BSI-KP onset**

We embedded the qPitt score in a logistic regression, and then fit all the other machine learning models with either the expert-chosen or full set of baseline and medical history variables (regardless of their putative role as confounders, mediators, or colliders). Table 3 shows the out-of-bag AUROC, sensitivity, and specificity for all methods in the whole study population (n=1,389) as well as the ICU sub-population (n=645). The qPitt exhibited moderate AUROC (0.631 in the whole study population, 0.589 in the ICU stays). When fitting the methods on the expert-chosen input variable set, the boosted logistic regression yielded the highest AUROC in both the full study population (0.801) and the ICU subpopulation (0.726). When fitting the methods on the full input variable set, the boosted logistic regression yielded the highest AUROC in the full study population (0.885) and in the ICU subpopulation (0.854). The random forest showed the highest sensitivity in the full study population (0.810), but very close to boosted logistic regression (0.809), which in turn showed the highest sensitivity in the ICU population (0.810). The highest specificity was shown by the boosted logistic regression in both the full study population (0.849) and in the ICU population (0.810). Of note, the lower AUROC for SVM was due to a portion of the bootstrap runs where the SVM models exhibited concave ROC with AUROC<0.5, rather than lower performance of all SVM models overall. In summary, the machine learning models improved prediction performance over the qPitt up to 26.94% in expert chosen model and 40.25% in full variable set model (p<0.0001). Random forests did not improve performance over boosted logistic regression; in particular, the hypothesis of no difference in mean AUROC between the two in the full dataset (0.881 vs. 0.885) as well as in the ICU subset (0.852 vs. 0.854) was not rejected at the 5% level (p-value 0.58 and 0.81, respectively).

**Emulated external validation**

For external validation, we removed data from clinics located in the 322 zip code (Duval County, Jacksonville metropolitan area) from the training procedure, and saved it for testing. The selected area comprised 19.4% of our data points (269/1,389) with an outcome prevalence of 10%, lower than main dataset. In this external dataset, the AUROC performance of boosted logistic regression was 0.888, the sensitivity was 0.741, and the specificity was 0.901 without recalibration. When we tested on the ICU population, the AUROC of the boosted logistic regression was 0.922, with a sensitivity of 0.773 and specificity of 0.969. These results indicate the models generalize well on other scenarios with putative covariate and outcome distribution shifts.

**Table 3.** Performance of the machine learning models in predicting 30-days mortality, by varying covariate set and target populations.
Figure 3 shows the ROC curve for each method in the whole population and in the ICU subpopulation, averaging over all bootstrap runs. The ROC curve of support vector machine is not convex due to a number of ill-conditioned results over single bootstrap runs.

Figure 3. Receiver operating characteristic curves for qPitt score and other linear/nonlinear models for predicting 30-day mortality in the whole study population and in the intensive care unit subpopulation. Curves are drawn and averaged from out-of-bag predictions (25 bootstrap samples).
The covariates with the largest effect sizes selected by the boosted logistic regression using the expert-chosen input set in the ICU population are listed in Table 4. Beta-Lactamase resistance in *Klebsiella* before or upon the infection was the strongest predictor of 30-days mortality (odds ratio OR=10.12, confidence interval CI=3.28-40.39). The qPitt score was also a strong predictor (OR=1.53, CI=1.22-1.95) together with the infection source of throat/lung (OR=1.78, CI=1.14-2.82), and presence of resistance to other antibiotics. Multiple antibiotic usage was also associated to death, likely indicating a more severe disease course; on the contrary, acetaminophen usage showed decreased risk, perhaps indicating a milder condition (as opioids are used for more severe cases).

**Table 4.** Top-15 variables (by effect size) of the expert-chosen boosted logistic regression model for prediction of 30-day mortality in the intensive care unit population.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Odds Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Betalactam resistance in <em>Klebsiella</em> before or upon the infection</td>
<td>10.12 (3.28 - 40.39)</td>
</tr>
<tr>
<td>HIV/AIDS</td>
<td>2.83 (0.71 - 12.95)</td>
</tr>
<tr>
<td>Amikacin use upon the infection</td>
<td>2.63 (1.3 - 5.57)</td>
</tr>
<tr>
<td>Tobramycin use upon the infection</td>
<td>2.34 (1.25 - 4.53)</td>
</tr>
<tr>
<td>Carbapenem resistance in <em>Klebsiella</em> upon the infection</td>
<td>2.23 (0.8 - 6.72)</td>
</tr>
<tr>
<td>Ethnicitygroup - not hispanic</td>
<td>2.06 (0.46 - 10.36)</td>
</tr>
<tr>
<td>Infection source - throat/lung</td>
<td>1.78 (1.14 - 2.82)</td>
</tr>
<tr>
<td>Vancomycin use upon the infection</td>
<td>1.76 (0.9 - 3.49)</td>
</tr>
<tr>
<td>Carbapenem resistance in non-<em>Klebsiella</em> organisms before or upon the infection</td>
<td>1.74 (0.92 - 3.35)</td>
</tr>
<tr>
<td>Fluoroquinolone resistance in <em>Klebsiella</em> upon the infection</td>
<td>1.68 (0.86 - 3.38)</td>
</tr>
<tr>
<td>Carbapenem use before the infection</td>
<td>0.41 (0.22 - 0.76)</td>
</tr>
<tr>
<td>qPitt Score</td>
<td>1.53 (1.22 - 1.95)</td>
</tr>
<tr>
<td>Hemiplegia or paraplegia</td>
<td>0.5 (0.27 - 0.93)</td>
</tr>
<tr>
<td>Acetaminophen upon the infection</td>
<td>0.5 (0.27 - 0.9)</td>
</tr>
<tr>
<td>Acute myocardial infarction</td>
<td>1.49 (0.85 - 2.67)</td>
</tr>
</tbody>
</table>

The covariates with the largest effect sizes selected by the boosted logistic regression using the full ICD code variable set in the ICU population are listed in Table 5. The do not resuscitate status was the strongest predictor of 30-day mortality (odds ratio OR = 17.01, confidence interval CI=8.54-35.72). The qPitt score was also a strong predictor in this model (OR=2.06, CI=1.45-2.94). Multiple ICD-9-CM codes were also associated to 30 day mortality including arthropathy, hypothyroidism, cardiomyopathies, and septic shock.

**Table 5.** Top-15 variables (by effect size) of the full ICD code set boosted logistic regression model for prediction of 30-day mortality in the intensive care unit population.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Odds Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD-9-CM V49.86 (Do not resuscitate status)</td>
<td>17.01 (8.54 - 35.72)</td>
</tr>
<tr>
<td>ICD-9-CM - V66.7 (Encounter for palliative care)</td>
<td>4.8 (2.47 - 9.48)</td>
</tr>
<tr>
<td>ICD-9-CM - 716.90 (Arthropathy, unspecified, site unspecified)</td>
<td>3.63 (1.65 - 8.03)</td>
</tr>
<tr>
<td>Opioids use upon the infection</td>
<td>3.61 (1.23 - 12.21)</td>
</tr>
<tr>
<td>ICD-9-CM - 244.9 (Unspecified acquired hypothyroidism)</td>
<td>3.02 (1.55 - 5.97)</td>
</tr>
<tr>
<td>ICD-9-CM - 425.4 (Other primary cardiomyopathies)</td>
<td>2.98 (1.3 - 6.83)</td>
</tr>
<tr>
<td>qPitt Score</td>
<td>2.06 (1.45 - 2.94)</td>
</tr>
<tr>
<td>ICD-9-CM - 785.52 (Septic shock)</td>
<td>2.01 (1.1 - 3.67)</td>
</tr>
<tr>
<td>Trimethoprim use upon the infection</td>
<td>0.1 (0.02 - 0.39)</td>
</tr>
<tr>
<td>ICD-9-CM - 041.3 (Friedländer's bacillus infection in conditions classified elsewhere and of unspecified site)</td>
<td>0.15 (0.06 - 0.32)</td>
</tr>
<tr>
<td>Infection source - throat/lung</td>
<td>1.83 (0.96 - 3.5)</td>
</tr>
<tr>
<td>Hemiplegia or paraplegia</td>
<td>0.17 (0.04 - 0.6)</td>
</tr>
<tr>
<td>Ceftazidime use upon the infection</td>
<td>0.31 (0.04 - 1.49)</td>
</tr>
<tr>
<td>ICD-9-CM – 780.09 (Other alteration of consciousness)</td>
<td>0.31 (0.13 - 0.69)</td>
</tr>
<tr>
<td>ICD-9-CM – 263.0 (Malnutrition of moderate degree)</td>
<td>0.39 (0.14 - 1)</td>
</tr>
</tbody>
</table>
Figure 4 shows a decision tree for the ICU population fitted with a fixed depth of six levels (for improved interpretability), providing similar performance to a fully grown-pruned tree. Although inferior in overall performance, a decision tree picked the most relevant predictors as the logistic regression. The qPitt score was the first split, and for those with high qPitt score the next splits included previous usage of vancomycin, age at diagnosis, and resistance to fluoroquinolones in non-Klebsiella organisms. For patients with a low qPitt score, resistance to Betalactams in Klebsiella species was the next split, followed by age at diagnosis, drug resistance to other bacteria, CCI, nosocomial infection, infection site, sex, tobramycin use, vancomycin use, and renal disease.

![Decision Tree Image]

**Figure 4.** Decision tree for predicting 30-days mortality in patients transferred to the intensive care unit.

We are aware that the prediction models included a number of variables that could be confounded by indication (e.g., the prescription of a strong antibiotic). For instance, 37% to 60% of ICU patients prescribed with vancomycin showed MDR to either KP or other organisms in their antibiogram test upon admission. We therefore fit another boosted logistic regression by including only variables measured before the admission and BSI-KP diagnosis. The AUROC for this model was 0.686 in the whole population and 0.604 in the ICU population, thus performance was inferior to the main prediction model.

**Discussion**

In this study, we found that prevalence of BSI-KP, 30-day mortality, and MDR were stable in this population throughout 2011-2018. The high rates of mortality and prevalence of MDR are worrisome. The qPitt score has mild performance in predicting mortality. Our expert-driven screening of additional EHR variables –combining demographic, clinical, and laboratory EHR data-- and the development of a new prediction model that included all diagnostic codes sensibly improved performance. As logistic regression and decision trees had comparable performance to more complex methods like random forests, we chose the simpler setup that also allows higher interpretability. The support vector machine exhibited lower performance than other methods only due to a portion of completely reversed models during bootstrapping. There could be multiple reasons, from cost overfitting to low event sampling in a bootstrap sample. We tuned both cost and gamma parameters, but a better tuning design, or a different kernel rather than radial basis, could improve performance.

One limitation of this work is that we did not have availability of data outside Florida. However, from our emulated external validation using the Duval/Jacksonville area, we confirmed that our model generalizes well over different clinics. Another limitation is that we converted all ICD-10 codes into ICD-9; it has been shown that ontology conversion, often ambiguous, can affect modelling, e.g., introducing including measurement errors, besides bringing operational issues.\textsuperscript{12,13}

Finally, we note that our prediction model is not causal by design, and the recalculation of risk by changing variables, e.g., using a different antibiotic, is not guaranteed to be correct. Even though we tested a secondary model with only pre-admission variables, which yielded also lower performance, there can be still presence of colliders and

281
confounding bias, thus the model’s counterfactuals cannot be used seamlessly for interventions\textsuperscript{14}. Although the model developed in this study was highly predictive of 30-day mortality due to BSI-KP and can be used to prescribe last resort antibiotics more judiciously, interventional models that make use of modifiable variables that can reduce mortality risk are needed.

**Conclusion**

In conclusion, this new model designed to predict 30-day mortality due to BSI-KP improves over the widely-used qPitt score and can be used to more accurately determine patient risk. Improved prediction of unfavorable BSI-KP patient outcomes can help clinicians improve antibiotic stewardship, and promote hospital efficiency from a public health standpoint. We look forward to extending the work by attaching a more solid causal structure to the machine learning to develop an interventional model and optimize patient survival.

**Funding**

This work was supported by UF Informatics Institute, National Institutes of Health - National Institute of Allergy and Infectious Diseases (NIH/NIAID) grant #R01AI141810

**Author Contributions**

IY devised the idea, implemented the code, performed analyses, and wrote parts of the article. SR contributed to study design, interpretation, and wrote parts of the article. SM provided statistical review and support for the analysis, and wrote parts of the article. ZF provided statistical review and support for the analysis. JB provided statistical review and support for the analysis, and wrote part of the article. GM contributed to study design and interpretation. MP contributed to the study design, method’s evaluation, statistical review, support for the analysis, interpretation, and wrote parts of the paper.

**Conflict of Interest Statement**

None declared.

**References**


11. Research NB of E. ICD-9-CM to and from ICD-10-CM and ICD-10-PCS crosswalk or general equivalence mappings. 2016;


Improving the Quality of Suggestions for Medical Text Simplification Tools

David Kauchak, PhD\textsuperscript{1}, Jorge Apricio\textsuperscript{1}, Gondy Leroy, PhD\textsuperscript{2}
\textsuperscript{1}Pomona College, Claremont, CA; \textsuperscript{2}University of Arizona, Tucson, AZ

Abstract

Text continues to be an important medium for communicating health-related information. We have built a text simplification tool that gives concrete suggestions on how to simplify health and medical texts. An important component of the tool identifies difficult words and suggests simpler synonyms based on pre-existing resources (WordNet and UMLS). These candidate substitutions are not always appropriate in all contexts. In this paper, we introduce a filtering algorithm that utilizes semantic similarity based on word embeddings to determine if the candidate substitution is appropriate in the context of the text. We provide an analysis of our approach on a new dataset of 788 labeled substitution examples. The filtering algorithm is particularly helpful at removing obvious examples and can improve the precision by 3% at a recall level of 95%.

Introduction

Limited health literacy continues to be a limiting factor that prevents patients making well-informed health decisions and negatively affects an estimated 90 million Americans\textsuperscript{1,12}. The negative effects directly impact patient health\textsuperscript{3,4}, but also have more broad effects, such as added financial costs in the health system\textsuperscript{5}. These negative effects have been shown in a range of diseases and settings, for example in cervical cancer screenings\textsuperscript{6} and medication use in diabetes patients\textsuperscript{7}. Many national initiatives have emphasized improving health literacy including the Affordable Care Act\textsuperscript{8}, Nationals Plan to Improve Health Literacy\textsuperscript{9}, and the Plain Writing Act\textsuperscript{10}.

A number of solutions have been proposed, but a critical component of these plans is the mechanism used for sharing the health information. Text continues to be one of the most effective means for spreading information as it is relatively easy to produce (in contrast to video or interactive media) and can be utilized in a wide range of mediums including both physical and digital variants. Text has been successfully used to disseminate information in many different environments and for many different health literacy purposes\textsuperscript{11–13}. The difficulty is that producing text that is informative and broadly accessible to a range of audiences and readability levels can be challenging\textsuperscript{14}. Readability formulas (e.g., Flesch-Kincaid\textsuperscript{15} and Simple Measure of Gobbledygook\textsuperscript{16}) are the most common “tool” used to help produce readable text, however, they have a number of critical drawbacks: they only provide document-level feedback; they do not provide concrete suggestions about how to improve the text; and, showing positive outcomes when using these formulas to guide simplification has been problematic\textsuperscript{14,17}.

To address these limitations, we have been developing a text simplification tool to assist content creators\textsuperscript{18}. The tool includes a number of features that have been experimentally shown to affect text difficulty and user understanding, including word-level suggestions\textsuperscript{19}, negation analysis\textsuperscript{20}, morphological analysis\textsuperscript{21}, sentence-level analysis\textsuperscript{22}, and document-level analysis\textsuperscript{23}. Critically, the tool both identifies problematic text portions and provides concrete suggestions for how to improve the text.

One of the most useful components of the tool identifies difficult words based on word frequency and suggests alternative, easier, rewordings using existing text resources (i.e., WordNet and UMLS). When guided just by these word suggestions, we have shown that the resulting text is more readable and results in improved learning\textsuperscript{19}. However, when using these static resources, candidate substitutions/synonyms are not always relevant in all contexts. The result is that some times the candidate substitutions are not helpful for the text that the user is simplifying. For example, air passage is one of the suggestions found in WordNet as a candidate substitution for airways, which is a reasonable substitution option for many health-related settings. Flight path is another synonym that is also suggested. It is a reasonable substitution in some settings, but not likely to be a common substitution in health-related texts.

In this paper, we propose a new algorithm for filtering these candidate substitutions based on the words in the document being simplified. We utilize recent advances in word embeddings to compare the similarity between the original, difficult word and the text and the candidate substitution and the text. Candidate substitutions that have similarities
Figure 1: Text simplification tool applied to a text about asthma. Difficult words are highlighted and underlined and can be clicked on to show a menu of candidate substitutions. *Airways* has been clicked on.

that are significantly different than the original word can be filtered and not shown as candidates to the end user. We provide experimental analysis showing the effectiveness of this filtering and show how it can be integrated into the existing tool so that the user still has some control over the candidate suggestions shown. The algorithm is general-purpose and can be applied broadly in the medical domain to any tool or applications that makes lexical suggestions.

**Lexical Substitution in the Tool**

The text simplification tool is available online and only requires a web browser to use. The tool is interactive and utilizes a built-in editor so that the user can get feedback and guidance from the tool while editing the text. Figure 1 shows a screenshot of the tool in use. To simplify a text, a user copies and pastes the text into the editor, then clicks the “Simplify Text” button. The text is analyzed by the different tool components and the text is marked up showing what portions of text could be improved. The red items are sentence-level suggestions and can be clicked on to show guidance about how to refactor the sentence to make the structure simpler (e.g., “This” in the third sentence in the last paragraph). The blue items are difficult words/phrases where potential substitutions are available. In this paper, we focus on these word substitutions.

When the difficult word is identified, we query two databases for candidate synonyms, WordNet and UMLS. Both have been hand-engineered and contain a large collection of words. UMLS has more medical and health-related terms while WordNet has better general coverage. We filter each candidate synonym to ensure that it is simpler than the original word, where simple is defined by term frequency in a large corpus. Synonyms that are simpler are then displayed to the user.

In Figure 1, *airways* has been clicked on and the tool generated three candidate substitutions. All three substitutions come from WordNet (indicated by the blue color), however, only the first two suggestions are reasonable in the context.
of the text. Although the tool is interactive and the writer can choose not to select the last option, even having to read this option can be distracting and cause extra time to be spent in making decisions. Additionally, in cases where the none of the candidate options are viable, then the word should not be highlighted at all allowing the writer to focus on those words with viable options.

Filtering Candidate Substitutions

The tool has been specifically developed and tested with the goal of simplifying medical texts. However, even in this more restricted domain, there are still a broad range of use cases and a broad range of materials that could be simplified. Motivated by this, we developed a filtering approach that was context dependent, i.e., leverages the content of the text being edited to make the decisions about what substitutions are relevant. This allows the filtering to be applied in a broad range of settings and also has the advantage of being targeted even within different health-related topics.

Given a difficult word, \( d \), occurring in text \( W = w_1, w_2, \ldots, w_n \) (where \( w_i \) are the words of the text, i.e., the document being simplified) and a list of candidate substitutions, \( c_1, c_2, \ldots, c_m \), the goal is to decide for each candidate whether or not it should be filtered, i.e., removed from the list of suggestions shown to the user. As the example highlights, reasonable substitutions should be related to the text that is being simplified. To quantify this, we calculate the semantic similarity between the candidate and the text, i.e., \( \text{sim}(c_i, W) \). To calibrate this similarity score, we also calculate the semantic similarity between the original difficult word and the text, i.e., \( \text{sim}(d, W) \). We then score the quality of each candidate as the ratio of these similarities,

\[
\text{score}(c_i) = \frac{\text{sim}(c_i, W)}{\text{sim}(d, W)}
\]

If this score is 1, then the candidate is similarly related to the text as the difficult word. If the score is greater than 1, then the candidate is more similar to the text than the difficult word and if it is less than 1, it is less similar. Words with higher scores should be better candidates for substitution since they relate better to the text. Normalizing the candidate by the similarity of the difficult word takes into account the actual relationship of the difficult word to the rest of the text and also allows for a broader range of similarity approaches to be used. Additionally, it gives some meaning to the score, e.g., if a candidate has a score of 0.5, then that word is half as similar as the original difficult word is to the text.

To calculate the similarity between a word and the text, we leverage recent advances in word embeddings to capture not just lexical overlap, but semantic overlap. Word embeddings have become a staple for many NLP applications as they allow for a word to be represented by a high-dimensional vector that captures meaning\(^{25,26}\). To calculate the similarity between two words, the distance between the word embeddings of the two words is calculated, often using something like cosine distance. To calculate the similarity between a word and a text, a similar high-dimensional embedding is also required. A number of techniques have been suggested for embedding text\(^{27}\), but we take a simple approach for creating a word embedding for a text as the average of the individual word embeddings of the words in the text. This has been shown to be at least as effective as other more complicated techniques\(^{28}\).

For concreteness, we calculate the similarity between a word, \( d \), and a text, \( W \), as:

\[
\text{sim}(d, W) = \cosine(vect(d), \frac{\sum_{i=1}^{n} vect(w_i)}{n})
\]

where \( vect \) is the word embedding for a word and \( \cosine \) is cosine similarity. The second argument to the cosine similarity is the average word vector for the document. For our implementation, we utilized Google’s Word2Vec pre-trained 300-dimension word embeddings\(^{29}\) (often referred to as Word2Vec\(^{2}\)) since they are freely available and have a large vocabulary, though any word embeddings could be used.\(^{2}https://code.google.com/archive/p/word2vec/\)
Figure 2: Example Amazon Mechanical Turk task for annotated the examples. Participants were asked to choose whether the difficult word could be replaced by the replacement word with minor fixing.

Experiments

To test the effectiveness of our filtering algorithm, we generated a new dataset of 788 examples of difficult words in the context of health-related text along with a candidate substitution for that difficult word. For each example, we solicited annotations about whether or not the substitution was valid for the difficult word within the context of the sentence. Using these examples, we then compared the output of our filtering algorithm to the ground truth annotations about whether the substitution was valid.

Data

For our health-related texts, we used English Wikipedia\(^3\). Wikipedia is frequently one of the top search results returned for search engine queries and is a common source for general information\(^23\). We randomly selected 10 articles that were from the category “Disease and Disorder” and selected the first two paragraphs from these articles as example texts. We used our text simplification tool to automatically identify difficult words in the text that had one or more candidate simplifications. For each occurrence, we extracted an example that had three parts: a difficult word, a sentence with an occurrence of the difficult word, a candidate substitution for the difficult word. This resulted in 788 examples.

For each example, we solicited annotations for whether or not the substitution was valid in the context of the sentence using Amazon’s Mechanical Turk\(^4\) (MTurk). MTurk has been used in a wide range of applications\(^29\) and has been shown to be an effective mechanism for data annotation when appropriate precautions are taken\(^30\). We restricted potential participants to those in the US to ensure English literacy and those that had an approval rating of 0.97 or higher. Figure 2 shows an example of the task shown to participants.

For each example, we solicited annotations from three participants and chose the majority label, i.e., yes the candidate replacement was a valid substitution or no it is not. This resulted in a dataset of 788 examples with annotations. The dataset is publicly available online.\(^5\)

Evaluation

To filter candidates we can threshold the candidate scores keeping only those with scores above the threshold (higher scores represent candidates that are more similar to the text). To evaluate how well our filtering is doing, we used four common metrics: precision, recall, F1, and accuracy. Precision is the proportion of those kept that are good substitutions. Recall is the proportion of the total number of good substitutions in the dataset that we kept as good substitutions. F1 is the harmonic mean of these two and is a common way to combine these two numbers. These three metrics are commonly used in tasks where one class is more important, in this case, the good candidates\(^31\). For completeness, we also include accuracy, which is the total number of correct examples divided by the total number of examples. An example is correct if it was not a good candidate and it was filtered and vice versa for good candidates.

\(^3\)https://en.wikipedia.org/
\(^4\)https://www.mturk.com/
\(^5\)To be made available if accepted.
Figure 3: Precision, recall, F1 and accuracy for increasing thresholds on the data set of 788 examples. Candidates with scores below the threshold were filtered.

Results

Figure 3 shows the four evaluation metrics for increasing thresholds. At a threshold of 0, no filtering occurs and all of the candidates are included. ~28% of the examples were rated as good substitutions and so the precision is ~28% when no filtering occurs. The similarity score does provide some information about the quality of the substitutions and as the threshold is increased, the precision also increases. The highest precisions achieved is 34% at a threshold around 1.0, however, the recall is low at this point with over half of the good candidates being removed.

F1 represents a balance between precision and recall and the highest F1 score achieved is 47.8 at a threshold of around 0.587. At this threshold, the precision is 32.6% and the recall is 89.6%: the candidates shown are of higher quality (4.6% absolute) while retaining most of the good candidates. Accuracy is maximized for the highest thresholds by simply filtering all candidates, thereby predicting everything as a bad candidate. One thing to note is that there were approximately 50 examples where there was no word embedding for the difficult word, but there was an embedding for the candidate substitution. These resulted in high scores without any real feedback about the candidate substitutions, which is why the recall does not drop to 0 on the graph.

The algorithm is particularly good at removing obviously wrong options. For example, the 22 lowest scoring examples were all bad substitutions and 59 of the lowest scoring 60 examples were bad substitutions. At a threshold of 0.47, the recall is 95% and the precision has increased by over 3% absolute to 31%. Not surprisingly, there was a clear tradeoff between precision and recall in our algorithm. As the threshold was increased more examples were filtered, some of which were not correct. Early this tradeoff is only marginally impacts the recall, however, the recall quickly started to drop for thresholds above 0.5.

Discussion

We had two main goals for investigating candidate filtering. First, this is a problem that comes up any time lexical substitution algorithms are applied, i.e., determining whether or not a substitution is relevant in the context of the text. Our algorithm is content-agnostic and could be applied in a range of domains and applications. Second, we wanted the filtering algorithm to help improve the quality and usability of our text simplification tool.

We have added the candidate filtering to the text simplification tool. Motivated by our experiments, we set the default cutoff to be 0.5. At this threshold, the recall was still 93%, retaining almost all of the good examples, but some filtering does occur, particularly those that were obviously incorrect. Depending on the application, the user may be more or less tolerant to seeing spurious suggestions. To allow the user some control over the quality of the suggestions shown,
Figure 4: Same example as in Figure 1, but with more stringent candidate filtering. Using the filtering, flight path is no longer shown as a candidate substitution for airways in this text.

we added a slider to adjust the threshold between 0 and 1.0, the range where the recall is most reasonable and the impact was the highest.

Figure 4 shows another screenshot of the tool being used on the asthma text. The slider in the bottom right corner (“Variety level”) varies the threshold and has been adjusted to be more stringent. At this threshold, flight path is removed as a candidate and is not shown.

One of the main advantages of our filtering algorithm is that decisions are made based on the context of the text. Because WordNet and UMLS are both high-quality resources, most of the synonyms are correct in some context. The key problem is determining if they’re correct for the current context being considered. Figure 5 shows a screenshot of another text that also uses the word airways. In this text, the first sentence references airplanes and provides a different context than the asthma text. Even though the same word appears and the same threshold is used, because the context is different, flight path is kept as a potential candidate.

Figure 5: Another example of the tool being used on text with airways. Even though the threshold is the same as in Figure 4, the candidates shown are different since the text context is different.
There is still room for improvement on this problem. Although we did obtain an increase in precision with minimal impact on recall, the overall precision was still relatively low. Lexical simplification algorithms that attempt to do this automatically, i.e., without a human in the loop, often use a much smaller context, e.g., a small context window around the difficult word. Our algorithm is broader, but could possibly benefit from some local information as well. Alternatively, accuracy might also be improved by employing targeted embeddings that have been adjusted to explicitly emphasize semantic similarity.

For this experiment, we used general-purpose word embedding as they provide broad coverage. However, additional research is needed to investigate the impact of more domain-specified embedding, e.g., either via fine-tuning general embeddings or from embeddings trained on a medical corpus.

We hope that the dataset utilized here will allow for future investigations of these options by other researchers.

Acknowledgements

Research reported in this paper was supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM011975. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

7. Powell CK, Hill EG, Clancy DE. The Relationship Between Health Literacy and Diabetes Knowledge and Readiness to Take Health Actions. The Diabetes Educator. 2007.


31. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one. 2015;10(3):e0118432.


Clinical Data Cohort Quality Improvement: The Case of the Medication Data in The University of Minnesota’s Clinical Data Repository

Sebastien D. Kiogou, MS¹, Chih-Lin Chi, PhD¹, Rui Zhang, PhD¹,², Sisi Ma, PhD¹, Terrence J. Adam, MD, PhD¹,²

1. University of Minnesota, Institute for Health Informatics, 2. University of Minnesota, College of Pharmacy

Abstract: Clinical and translational research centers (CTRCs) have emerged as key centers for electronic medical record related research through integrated data repositories (IDRs) and the ‘secondary use’ of clinical data. Researchers accessing and pre-processing ever increasing amounts of electronic medical records for data mining tasks have a growing need for best practice approaches for clinical data quality assessment and improvement. This project focused on a large data extract for 7 statin medication prescriptions for patients with cardiovascular disease. After the initial data extraction, we proceeded to analyze the data for completeness, correctness, currency, and percentage populated using established data quality frameworks. Assessment of the said data was performed through medication possession ratios, medication discontinuation reasons, and drug dosages. When we compared distributions of data elements such as drug dosage before and after changes were introduced by our pre-processing protocols, only a minimal noticeable difference was found as the clinical data cohort quality assessment and pre-processing were completed without substantially altering the original data structure. Our study demonstrated practical steps for clinical data cohort quality improvement using medication data and illustrates a best practice approach in clinical data cohort quality improvement for any data mining tasks.

Introduction

Clinical and translational research centers (CTRCs) provide resources to support researchers in applying scientific efforts to patient care. Integrated data repositories (IDRs) provide key support of clinical and translational research by providing data sources for retrospective analytic research and the identification and recruitment of prospective research subjects. IDRs are often the byproducts of research collaborations between academic medical centers and healthcare delivery organizations; thus, IDRs are integrated data sources fed with millions of heterogeneous electronic health records (EHRs) from clinics and hospitals. The emergence of CTRCs as research resources and the increasing availability of electronic medical records through IDRs have enhanced the opportunity for ‘secondary use’ of such data for research.

The University of Minnesota’s Clinical & Translational Science Institute (CTSI) is a member of the CTSA consortium. Part of the CTSI’s prerogatives is to support the Best Practices Integrated Informatics Core (BPIC) which provides centralized informatics services and collaborative science opportunities for the university. In the same vein, the clinical data repository (CDR) created by the CTSI grants researchers access to millions of electronic health records. The CDR is an integrated medical record repository comprised of patient records from 8 hospitals and more than 40 clinics - from Fairview Health Services and the University of Minnesota Physicians. The latter information-rich resource gives researchers the ability to study medical conditions, analyze patient outcomes, and identify best practices across large populations, all while protecting patients’ privacy. On December 6, 2019 The University of Minnesota’s Institutional Review Board (IRB) under the human research protection program approved our submission for a study that uses machine learning algorithms to address the critical gap related to the lack of a standardized data approach for statin related adverse events (ADEs) detection and surveillance.

Cardiovascular diseases (CVD) are the leading cause of death both in the United States (U.S) and the world; about 655 thousand Americans and 17.9 million individuals worldwide die from heart disease each year. In the US, CVD claim more lives each year than all forms of cancer and chronic lower respiratory disease combined. In developed countries, cholesterol levels more than 3.8 mmol/liter are responsible for more than 50% of CVD events. Statin drugs or HMG-CoA reductase inhibitors are a class of lipid-lowering drugs used to prevent or treat cardiovascular disease for both primary and secondary prevention. According to the American Heart Association, statin drugs are the most common cholesterol-lowering drugs. Statin drugs can lower LDL cholesterol concentration by an average of 1.8 mmol/l thereby reducing the risk of cardiac events (heart attack, sudden cardiac death) by about 60% and that of stroke by 17% after long-term treatment. However, statin therapies have been associated with ADEs such as muscle damage, increased risk of diabetes mellitus, liver damage, neuropathy, pancreatic and liver dysfunction. In addition, these morbidities constitute major causes of statin medication non-adherence; in a survey
by Cohen, Brinton, Ito, & Jacobson⁸, the primary reason for statin therapy discontinuation was ADEs (62%) which according to Bates, Connaughton, & Watts³ is a major challenge for preventive cardiology.

To address this critical gap, our proposed research elaborated three specific aims, one of which is discussed in this article. Big data technology and advanced data mining algorithms including machine learning, deep learning, and natural language processing have proven optimal in handling high dimensional data. However, the success of data-driven algorithms is influenced by both the algorithm and the data used for modeling. For any data mining task, early steps such as data profiling, data extraction, data cleaning, data pre-processing, and data quality improvement are critical to model success⁹. The focus of this article is to develop practical steps for clinical data cohort quality improvement using the CDR medication data. Though this approach was used for a particular clinical data cohort quality improvement, it illustrates a best practice approach in clinical data cohort quality improvement for any data mining task.

Materials and Methods

The University of Minnesota’s Clinical Data Repository (CDR): The CDR is an electronic medical record repository comprised of hospital and clinic patient records from Fairview Health Services and the University of Minnesota Physicians. The data in the CDR come from the electronic health records of 2.5 million patients at 8 hospitals and more than 40 clinics. Available hospital data started to flow in from year 2011 and clinic data from 2000 with access to the data being granted via the Academic Health Center Information Exchange (AHC-IE) Data Shelter which provides a variety of analytic tools for data analysis. The data in the CDR is housed in patient and encounter centered tables within a SQL Server Management Studio (SSMS) inside the AHC-IE data shelter; the latter data shelter is a collaborative agreement between the University of Minnesota Academic Health Center and Fairview Health Services to support the joint mission of improving patient care and supporting healthcare research and education⁷. The data currently available in the CDR can be classified into the broad categories of demographic, administrative, and procedural data detailing patients’ problem list, vitals, medication, diagnosis, etc.

Data Quality Assessment: For ease of researcher use and data delivery, tables in the AHC-IE data shelter are combined and standardized through a variety of quality checks and then made available in the clinical data repository⁷. To better evaluate the quality of the medical data, we assessed the tables and data elements of interest for completeness, correctness, and currency in conformity with Weiskopf & Weng¹⁷. According to the same authors, EHR data were correct if the information they contained were true; whereas completeness referred to whether a truth about a patient was present in the EHR; as well, EHR data were considered current if they were recorded in the EHR within a reasonable period of time or, alternatively, if they were representative of the patient state at a desired time of interest. To evaluate whether the data was complete enough for our specific purpose, we either looked at the presence or absence of certain data elements, sought agreement between data elements from different tables, or compared distributions / occurrences of certain data elements with nationally recorded rates including with the American Heart Association (AHA) and the Center for Disease Control (CDC). Data correctness was assessed by evaluating the agreement between data elements in the data shelter to their expected physiological value ranges. For currency evaluation, we checked if data were entered into the EHR within a set time limit or if the data element was measured recently enough to be considered medically relevant. The core data elements in our investigation came from seven tables; in addition to the medication table, we utilized the patient, patient coverage, social history, basic vital, diagnoses, and reason for visit tables. Among the data quality assessment methods noted above, we used the most appropriate or feasible method when assessing a particular data element. After assessing data quality for all data elements in tables of interest, the data was filtered, and data elements were joined from these seven tables to obtain the analysis dataset. In the medication table, the following data elements were assessed: patient_id and service_id are identification features necessary to join tables; drug_name_orig and ingredient_name specified the medications used such as Simvastatin; order_date, discontinued_date, start_date and end_date were utilized for filtering and matching; dose_amt along with dose_unit were used as predictors of our outcome of interest; medication_order_status as well as active_order_status were used to implement exclusion / inclusion criteria; discontinued_reason served as predictor and categorization factor. For all these data elements, data quality was assessed through the dimensions of correctness, completeness, and currency. As an example, for dose_amt which is defined as the quantity of medication to be administered at time, we started by checking percent populated; then to assess correctness of dose_amt, we compared its values with expected value ranges based on ingredient_name and clinical expertise / experience of authors. Although the presence of dose_amt added to the completeness of the medication table, we further check for completeness by comparing distribution of the latter data element to what is
specified in the Grundy et al. Clinical Practice Guidelines. Additionally, medication active_order status, order date, discontinued date, start_date, and end_date were used as references to ascertain dose_amt currency.

Data Quality Improvement: The Medication table included 109,870 individual patients and 31,264,618 rows; each row of the table represents a medication order for a particular patient though patients may have more than 1 order. We were interested in seven particular statin drugs including Rosuvastatin, Atorvastatin, Pitavastatin, Pravastatin, Simvastatin, Lovastatin, and Fluvastatin; these agents are the most commonly prescribed statin drugs as attested by the literature and the extracted dataset. Study period was defined to be from January 1st, 2000 to December 31st, 2019 reflecting the availability of data in the CDR and the timing of our IRB acceptance. To ensure we were including only orders that were created and medications that were taken, we focused our study on Active or Completed medications whose Order Status were either Completed, Dispensed, Sent, OR Verified. The study was limited to subjects 18 years or older to focus on adult medication dosages and adverse effect reactions. After implementing all study related inclusion/exclusion criteria, we were left with 56,658 individual patients and 422,138 patient drug orders. The quality improvement effort focused on medication dosage, pseudo medication possession ratio, and medication discontinuation.

Statin Drug Dosage Distribution: While identifying the seven statin drugs in the medication data, we encountered five additional statin related combination drugs including ezetimibe-Simvastatin, Amiodipine-atorvastatin, Lovastatin-Niacin, atorvastatin-ezetimibe, and Niacin-Simvastatin. Dosage distributions for the seven statin drugs and the combination drugs showed dosage amounts / units that could not be used in our planned data mining task. To improve the quality of dosage data and make it suitable to our purpose we made the following changes. First, we converted any combination drug prescription into a single statin prescription making sure we kept rosvuvastatin, atorvastatin, pravastatin, simvastatin, lovastatin, fluvastatin, or pitavastatin afterward since statin exposure was a key information component. Most notations for the dosage amount of statin combination drugs consist of two numbers separated by a dash (-); yet, we used expert knowledge to assign single statin dose amount appropriately. For the combination drug Lovastatin-Niacin for instance, we had a dosage of 1000-20 mg; we chose to keep 20 mg as dosage and Lovastatin as drug name. Furthermore, if the dosage unit of any statin combination drug was tablet, we changed the combination drug name to be that of the statin drug to provide consistent statin exposure representation; the name on the other part of the combination drug was not retained as it was not needed for later analyses. This information could have been retained if we were seeking more specific information on drug exposures and interactions. Nonetheless, we created an additional column to indicate that these were combination drugs. Similarly, all dosage units of capsule were renamed to tablet. Finally, if a dosage unit were tablet, we converted the dosage to mg as follows; we would multiply the number of tablet(s) by the dosage amount of the most recent mg prescription of the same drug for the same patient; if there is no mg prescription of the same drug for a particular patient, we used the most frequent dosage amount in mg for that particular drug in the whole table.

Medication Possession Ratio (MPR): To ensure minimum adherence to the statin drugs in the study population, a pseudo medication possession ratio (pMPR) was calculated, as MPR remains one of the most referenced measures of medication adherence in the health care industry. Since we could not compute a true MPR based on our setting due to limits on available prescription fill data, we computed a medication possession ratio based on medication start date and end date while accounting for overlapping dates; thus, if two dates overlapped, we don’t count the overlapping days twice before dividing the days on medication by the period patient was on the stated drug. An upper bound was also set for the pMPR such that any value greater than 1 was assigned a value of 1. Nonetheless, to improve data cohort quality, we sought to understand why some patients had pMPR below 0.8 which is the threshold for an acceptable MPR. In the pool of patients with a pMPR below 0.8, we compared the group of pMPR below .4 to the rest of the group. We looked at different avenues to identify differences between these groups including time spent on medication, number of medications ordered, and medication discontinuation reason with the most evident difference being revealed through discontinuation reason.

Results

Several tables were instrumental in elucidating the methods we set forth for data quality assessment and improvement. Table 1.1 and table 1.2 below contain detailed data quality assessment of all core data elements used in the medication table.
Knowing the percentage populated, correctness, completeness, and currency of data elements early in the process will ultimately guide the planning of data elements to be used in the data mining task. The majority of our core data elements was reasonably populated with discontinuation reason being the lowest populated data element at 62%; yet once we filtered the data down to study related criteria, the data elements were populated enough to be kept in the data mining task with dose_amt and dose_unit being at the lowest percentage of 75%. We eventually used other data elements including drug_name_orig to upgrade the percentage populated of dose_amt and dose_unit in the data quality improvement process. As can be seen below, the correctness, completeness, and currency of all core data elements were evaluated. Although it seemed tedious at time to perform these tasks, the time we spent during this process turned valuable during the analysis phase.

The data quality improvement process assessed if patients were adherent to their medications but also pre-processed the medication table features in ways that make them amenable to model building. Medication dosage was perhaps the most intriguing feature that needed pre-processing since our initial study was aiming at identifying statin users at risk of developing adverse drug events. Following the pre-processing protocol described earlier, we converted all dosage amounts and units in standard milligram (mg) dosage. Furthermore, we discarded all prescriptions that were less than .1% of total prescriptions for a particular statin drug; we thought these prescriptions were not standard and could not be part of a general trend our models would capture or adding these random prescriptions could add noise to models.

| Table 1.1: Data quality assessment: Data element population |
|---------------------------------|----------------|----------------|----------------|
| Column-Description | % Populated (Original table) | % Populated (Studied drugs) |
| Patient_id: System generated patient ID number | 100.00 | 100.00 |
| Service_id: Medical service ID associated with patient problem | 100.00 | 100.00 |
| Drug_name_orig: Medication name assigned by data source (e.g. FUTUREMIDE 10 MG/ML IJ SOLN) | 100.00 | 100.00 |
| Ingredient_name: Medication ingredient name(s) (may contain > 1 ingredient) | 89.31 | 100.00 |
| Order_datetime: Date and time order was placed. | 100.00 | 100.00 |
| Start_date: Date when administration of the medication should begin | 94.28 | 99.97 |
| End_date: Date when the administration of the medication should end | 86.84 | 89.71 |
| Dose_unit: Units used for the dose amount (e.g. Mg, bottle) | 71.77 | 75.00 |
| Medication_order_status: The current status of the medication order (eg. sent, canceled, completed) | 91.28 | 100.00 |
| Active order: Indicates active order (Active Medication, Completed Medication, Discontinued Medication) | 90.86 | 100.00 |
| Discontinued_reason: Reason for medication discontinuation (e.g. Allergic Response, Dose Adjustment) | 62.72 | 84.45 |
| Discontinued_datetime: The date and time the medication was discontinued. | 74.41 | 99.64 |

| Table 1.2: Medication Table Data Quality Assessment |
|----------------------------------------|----------------|----------------|----------------|
| Column-Description | Correctness | Completeness | Currency |
| Patient_id: System generated patient ID number | System generated | System generated | System generated |
| Service_id: Medical service ID associated with patient problem | System generated | System generated | System generated |
| Drug_name_orig: Medication name assigned by data source (e.g. FUTUREMIDE 10 MG/ML IJ SOLN) | Checked agreement between drug_name, drug_name_orig, ingredient_name, generic_name, brand_name | Checked CDR presence and % populated | Assessed date features: order date, start date, end date, and discontinued date |
| Ingredient_name: Medication ingredient name(s) (may contain > 1 ingredient) | Checked association of diagnosis in diagnosis table and drug name in medication table | Checked CDR presence and % populated | Checked date features: order date, start date, and discontinued date |
| Order_datetime: Date and time order was placed. | Checked agreement between order date, start date, end date | Checked CDR presence and % populated | Checked date features: start date, end date |
| Start_date: Date when administration of the medication should begin | Checked agreement between order date, start date, end date | Checked CDR presence and % populated | Checked date features: order date, end date |
| End_date: Date when the administration of the medication should end | Checked agreement between order date, start date, end date | Checked CDR presence and % populated | Checked through date features: order date, start date |
| Dose_unit: Units used for the dose amount (e.g. Mg, bottle) | Checking expected value ranges and agreement of ingredient name, frequency, strength amount, route, form, instruction | Checked CDR presence and % populated | Checked date features or EHR data entry within study period |
| Medication_order_status: The current status of the medication order (eg. sent, canceled, completed) | Checked for agreement between medication order status, order date, active order | Checked CDR presence and % populated | Checked date features or EHR data entry within study period |
| Active_order: Indicates if an order is active (Active Medication, Completed Medication, Discontinued Medication) | Checked for agreement between medication order status, order date, active order | Checked CDR presence and % populated | Checked date features or EHR data entry within study period |
| Discontinued_reason: Reason for medication discontinuation (e.g. Allergic Response, Dose Adjustment) | Checked for agreement between discontinuation reason, service id, diagnosis code, diagnosis date, discontinued date | Checked CDR presence and % populated | Checked date features or EHR data entry within study period |
| Discontinued_datetime: The date and time the medication was discontinued. | Assessed agreement between discontinuation reason, diagnostic datetime, order date, start date, end date | Checked CDR presence and % populated | Checked date features: order date, diagnostic datetime, start date, end date |

296
Initial dosage distributions for the seven statin drugs and the combination drugs can be seen in table 2.1 and table 2.2. These summary tables show distributions of dosage amount / unit prescribed for all the patients. Each cell shows the dosage amount, dosage unit, and the percentage of time it was prescribed for that specific statin drug; for instance, in the very first cell of the table 2.1 we can see that 40 mg was prescribed 39.51% of the time for all Simvastatin drug prescriptions. The latter tables informed us about the changes that needed to be made to the dose amount and dosage unit in order to fit our data mining algorithms. Instances of dosage with whole numbers and mg as unit were the most appropriate to our data mining task; we then proceeded to convert / change all other values and units.

Following the protocol, we described in the method section, we made all necessary changes and were left with amenable dosage amount all in mg. We then proceeded to evaluate the changes we made; we compared distribution...
of drug dosages before and after changes were introduced by our pre-processing protocol; table 8 below represents
the new dosage distribution after implementing suggested pre-processing protocol whereas table 7 shows dosage
distributions from the initial medication table.

<table>
<thead>
<tr>
<th>Table 7: Initial Statin Drug Dosage Distributions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Drugs</strong></td>
</tr>
<tr>
<td>40 mg</td>
</tr>
<tr>
<td>20 mg</td>
</tr>
<tr>
<td>80 mg</td>
</tr>
<tr>
<td>10 mg</td>
</tr>
<tr>
<td>5 mg</td>
</tr>
<tr>
<td>60 mg</td>
</tr>
<tr>
<td>30 mg</td>
</tr>
</tbody>
</table>

The biggest difference in both tables is found in the distribution of statin drug *Fluvastatin* with difference margin no
more than 6%; otherwise, distribution of dosage amounts for every other statin drug is similar. This is an indication
that the pre-processing protocol did not alter the internal distribution of medication dosage even if it helped improve
the quality of the medication dosage for our data mining purpose. We also checked the validity of the same pre-
processing protocol by categorizing medication dosage based on the guideline found in Grundy et al.11. The
guideline categorized statin drug dosages into low, moderate, and high intensity for treatment consideration.
Likewise, we converted statin drug dosages for the initial and new tables into low, moderate, and high intensity
categories - as shown in table 9 below - and compared their distributions. The only noticeable difference was found
with *Fluvastatin* with a difference margin less than 2% thereby confirming no structural change in dosage
distribution after data quality improvement steps.

<table>
<thead>
<tr>
<th>Table 8: Statin drug dosage distribution after pre-processing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Drugs</strong></td>
</tr>
<tr>
<td>40 mg</td>
</tr>
<tr>
<td>20 mg</td>
</tr>
<tr>
<td>80 mg</td>
</tr>
<tr>
<td>10 mg</td>
</tr>
<tr>
<td>5 mg</td>
</tr>
</tbody>
</table>
| 60 mg | 0.31% | 60 mg 0.25% | 60 mg 0.49% | 30 mg 0.1% | 30 mg 0.1% | 30 mg 0.1% | 30 mg 0.1%
| 30 mg | 0.11% | 30 mg 0.1% | 30 mg 0.1% | 30 mg 0.1% | 30 mg 0.1% | 30 mg 0.1% | 30 mg 0.1% |

In the same vein, pMPR were computed for all patients in the study population to assess medication adherence. 87%
of the study population had pMPR above 0.8. These metrics suggested that the study population was adherent to the
statin drugs studied. However, the computed pMPR was likely higher than typical MPR since the only available
data was on medication orders and did not reflect actual prescription fill data typically used for MPR calculations.

<table>
<thead>
<tr>
<th>Table 9: Statin Drug Dose Pre- and Post-Standardization</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Drugs</strong></td>
</tr>
<tr>
<td>Simvastatin</td>
</tr>
<tr>
<td>Simvastatin</td>
</tr>
<tr>
<td>Simvastatin</td>
</tr>
<tr>
<td>Atorvastatin</td>
</tr>
<tr>
<td>Atorvastatin</td>
</tr>
<tr>
<td>Atorvastatin</td>
</tr>
<tr>
<td>Lovastatin</td>
</tr>
<tr>
<td>Lovastatin</td>
</tr>
<tr>
<td>Pravastatin</td>
</tr>
<tr>
<td>Pravastatin</td>
</tr>
<tr>
<td>Rosuvastatin</td>
</tr>
<tr>
<td>Rosuvastatin</td>
</tr>
<tr>
<td>Rosuvastatin</td>
</tr>
<tr>
<td>Fluvastatin</td>
</tr>
<tr>
<td>Fluvastatin</td>
</tr>
<tr>
<td>Pitavastatin</td>
</tr>
</tbody>
</table>
To understand why some patients had pMPR below the threshold, we compared the group of pMPR below .4 to the rest of patients with pMPR below 0.8. We created five (5) different categories of medication discontinuation reasons based on fifty different medication discontinuation reasons in the study population charted data. This categorization which was based on discontinuity similarity as attested by the present authors, was a way to group medication discontinuation reasons into categories that would better explain medication adverse events occurrence. In table 4, we show the percentage of occurrence of each category as well as the percentage of occurrences within each category of medication discontinuation reason. In more than 87% of the cases, the medication was discontinued for reordering reasons; stopped drugs, administrative, and adjustment were the next most frequent.

<table>
<thead>
<tr>
<th>Discontinue Category</th>
<th>Discontinuation Reason % (total)</th>
<th>Discontinue Category</th>
<th>Discontinuation Reason % (total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reorder</td>
<td>reorder 97.61%</td>
<td>Admin</td>
<td>Medication Reconciliation Clean Up 3.29%</td>
</tr>
<tr>
<td></td>
<td>Stopped by Patient 5.99%</td>
<td>Duplicate</td>
<td>23.45%</td>
</tr>
<tr>
<td></td>
<td>Alternate therapy 30.22%</td>
<td>Stopped Pre-Absence or entry error 15.29%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Therapy completed 27.17%</td>
<td>Error</td>
<td>8.74%</td>
</tr>
<tr>
<td></td>
<td>Discontinued by encounter’s provider 16.76%</td>
<td>Formulary Change 9.50%</td>
<td>5.62%</td>
</tr>
<tr>
<td></td>
<td>Discontinued by other Health Provider 8.82%</td>
<td>Cost of medication 5.42%</td>
<td>5.62%</td>
</tr>
<tr>
<td></td>
<td>Not Needed 2.93%</td>
<td>Unauthorized Entry 2.33%</td>
<td>2.00%</td>
</tr>
<tr>
<td></td>
<td>Rx not filled by Patient 1.52%</td>
<td>Not Covered by Insurance 2.00%</td>
<td>1.90%</td>
</tr>
<tr>
<td></td>
<td>Medication Failed 0.91%</td>
<td>Cost Formulary change 1.90%</td>
<td>1.90%</td>
</tr>
<tr>
<td></td>
<td>Not filled taken by Patient 0.35%</td>
<td>Pharmacy Medication Reconciliation 0.90%</td>
<td>0.90%</td>
</tr>
<tr>
<td></td>
<td>Not Effective 0.25%</td>
<td>Appointment needed for refills 0.71%</td>
<td>0.71%</td>
</tr>
<tr>
<td></td>
<td>OTHER 0.1%</td>
<td>Unavailable 0.35%</td>
<td>0.35%</td>
</tr>
<tr>
<td></td>
<td>Med D/Cd 0.2%</td>
<td>Per Nursing Home MAR 0.24%</td>
<td>0.24%</td>
</tr>
<tr>
<td></td>
<td>Pandevaluated response 0.2%</td>
<td>Availability 0.12%</td>
<td>0.12%</td>
</tr>
<tr>
<td></td>
<td>Med Changed 0.2%</td>
<td>Stopped during discharge/readmit 0.09%</td>
<td>0.09%</td>
</tr>
<tr>
<td></td>
<td>Side effect 0.99%</td>
<td>Contact Move - Error -0.09%</td>
<td>-0.09%</td>
</tr>
<tr>
<td></td>
<td>Allergic response 8.85%</td>
<td>Out of medication -0.09%</td>
<td>-0.09%</td>
</tr>
<tr>
<td></td>
<td>Contraindicated 4.23%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Contraindicated/pregnancy 0.97%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Medication Reagent 0.09%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pregnancy 0.09%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The investigation of medication discontinuation reasons was instrumental in showing that low pseudo medication possession ratio occurs as a result of medication non-adherence. First, we computed the percentage of occurrence of discontinuation categories within each pMPR group as illustrated in table 5 (pMPR <= 0.4 Vs. 0.6 < pMPR <= 0.8). we can see that the main differences appear with the discontinuation categories Reorder and StoppedDrug. More than 88% of high pMPR patients reordered their medications as opposed to 60% of low pMPR patients. Additionally, only 5% of high pMPR patients stopped their drugs versus more than 22% among low pMPR patients. Correspondingly, non-adherence appears to be the main difference between these discontinuation category percentages. Furthermore, the StoppedDrug discontinuation category was analyzed to find which items were involved in this discrepancy.

<table>
<thead>
<tr>
<th>Reason for Discontinuation</th>
<th>High MPR subjects</th>
<th>Low MPR Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reorder</td>
<td>88.63%</td>
<td>60.62%</td>
</tr>
<tr>
<td>Stopped Drug</td>
<td>5.37%</td>
<td>22.87%</td>
</tr>
<tr>
<td>Admin</td>
<td>3.15%</td>
<td>6.91%</td>
</tr>
<tr>
<td>Adjust</td>
<td>2.01%</td>
<td>3.58%</td>
</tr>
<tr>
<td>SideEffect</td>
<td>0.75%</td>
<td>5.58%</td>
</tr>
<tr>
<td>Discharge</td>
<td>0.09%</td>
<td>0.44%</td>
</tr>
</tbody>
</table>

As shown in table 6 below, 28% of high pMPR patients used alternative therapy while only 14% of low pMPR patients did; even so, more than 43% of low pMPR patients stopped their medication on their own compared to 28% for high pMPR patients. Thus, high pMPR patients used more alternative therapies whereas low pMPR patients just stopped using their statin medication unilaterally.
Discussion

Integrated data repositories hosted in academic research centers can host millions of patient records from diverse sources providing an information-rich resource for researchers. However, this opportunity has given rise to another challenge, that of extracting and preparing data from IDRs for research purposes. Data cleaning and quality improvement stages during research are often more time consuming than the later analysis and modeling phases. This article uses the medication data in the CDR as a use case to develop practical steps for clinical data cohort quality improvement; we demonstrated ways to perform data cohort quality checks based on the concepts of correctness, completeness, and currency while also improving data quality when pre-processing medication dose amounts and units to fit our modeling scheme. Achieving quality improvement with the medication data was possible through the instrumentalization of key concepts such as the medication possession ratio, medication discontinuation reasons, medication dosage categorization, etc. The quality of each data element useful to our modeling purpose in the medication table was assessed including the percentage of data populated in each data column as well as the correctness, completeness, and currency. The data was also pre-processed for the medication dosage information to ensure dosage format is amenable to modeling paradigms; we also created dosage categories and compared the distribution of dosages before and after modification were made to the dosage amount and units. However, this effort has a number of limitations. It is focused on electronic medical record data for medication related information. As a result, the observations do not necessarily extend into other areas of medical information. In addition, the data from the project is derived from a single health system which has local and regional practice components which may be unique from the perspective of medication and documentation patterns. The focus of the project which was on statin medications also limits the type of medications which were reviewed and may not apply to other therapeutic or drug classes which may not have similar dosage forms, drug combinations, or patterns of use.

Nonetheless, we observe that data quality assessment and data sources integration have emerged as some of the major topics in health care literature. Weiskopf & Weng\textsuperscript{17} performed a review of the clinical research literature discussing data quality assessment methodology for electronic health record (EHR) data reuse for research. The latter authors concluded that if the reuse of EHR data for clinical research is to become accepted, researchers should adopt validated, systematic methods of EHR data quality assessment. Their conclusion speaks to the importance of developing a systematic and accepted methodology for assessing and improving EHR data quality; the current paper incorporates these recommendations. Nonetheless, the topic of clinical data cohort pre-processing or quality improvement is important. Chi et al.\textsuperscript{6} discussed data cohort extraction from IDRs to facilitate machine learning; while the same authors mentioned large public and private databases available to researchers, they elaborated more on emerging models based on the development of partnerships which make available “primary” use corporate data for secondary data analysis. The summary of the overall approach to clinical data pre-processing using the medication use case is summarized in Appendix I which can provide a framework for similar future efforts. The AHC-IE data shelter provides a collaborative agreement between the University of Minnesota Academic Health Center and Fairview Health Services to support the joint mission of improving patient care and supporting health care research and education. This collaborative work confirms the current trend of agglomerating sources of data through system interoperability, integrated data sources, and collaboration between healthcare organizations and
Future work from this study will expand on the clinical data cohort preparation for machine learning tasks for prediction of adverse event outcomes and continue to develop best practice approaches in clinical data cohort quality improvement.

**Conclusion**

Our study demonstrated practical steps for clinical data cohort quality improvement using the medication table in the AHC-IE data shelter within the University of Minnesota’s Clinical Data Repository. In many regards, we sought to illustrate a best practice approach in clinical data cohort quality improvement for any data mining task.

---

### References

1. Aguilar-Salinas C, Zubirán R. Faculty opinions recommendation of interpretation of the evidence for the efficacy and safety of statin therapy. Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature. 2016;


Mining Medication Use Patterns from Clinical Notes for Breast Cancer Patients Through a Two-Stage Topic Modeling Approach

Kimberley ES Kondratieff, MS¹, J Thomas Brown¹, Marily Barron¹, Jeremy L Warner, MD, MS, FAMIA¹,², Zhijun Yin, PhD, MS¹,³
¹Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN USA; ²Department of Medicine, Division of Hematology/Oncology, Vanderbilt University Medical Center, Nashville, TN USA; ³Department of Computer Science, Vanderbilt University, Nashville, TN USA

ABSTRACT

Obtaining medication use and response information is essential for both care providers and researchers to understand patients’ medication use and long-term treatment patterns. While unstructured clinical notes contain such information, they have rarely been analyzed for this purpose on a large scale due to the demands of expensive manual reviews. Here, we aimed to extract and analyze medication use patterns from clinical notes for a population of breast cancer patients at an academic medical center using unsupervised topic modeling techniques. Notably, we proposed a two-stage modeling process that was built upon correlated topic modeling (CTM) and structural topic modeling (STM) to capture nuanced information about medication behavior, including drug-disease relationships as well as medication schedules. The STM-derived topics show longitudinal prevalence patterns that may reflect changing patient needs and behaviors after the diagnosis of a severe disease. The patterns also show promise as a predictor for medication-taking behavior.

INTRODUCTION

With an estimated 70% of Americans taking prescription drugs¹, care providers must understand how, when, and why patients take (or do not take) their medications in the outpatient setting. In addition, secondary researchers can make use of high-quality information regarding patients’ medication use to define windows of medication exposure, detect potential medication side effects, and understand patient-reported barriers to taking medication²–⁴.

Structured electronic health records (EHRs) contain medication lists and refill records, which are valuable for both care providers and researchers seeking to understand medication use. However, a great deal of information on patient behavior also resides in unstructured text such as clinical notes and patient-provider communications². While use of this information previously required extensive manual review of clinical documents, natural language processing (NLP) has made automated information extraction possible for large clinical document corpora⁶–⁷.

For instance, topic modeling, especially its standard implementation, Latent Dirichlet Allocation (LDA), is a classical unsupervised NLP technique used for these information extraction tasks⁸. LDA treats documents in a corpus as realizations drawn from an underlying probability distribution of abstract topics, which themselves are distributions over the words seen in the corpus. Each document is assumed to be composed of one or more topics, while each topic is assumed to be associated with relatively few high-probability words. The topics produced by such models, which are interpreted by their top probable words, yield useful representations of a text corpus and may possess predictive utility when used as features in other statistical models⁹.

The preponderance of existing work on topic modeling over documents from the EHR includes, but is not limited to, 1) prediction of disease progression, staging, or outcome¹⁰–¹²; 2) prediction of hospital readmissions or mortality¹³–¹⁵; and 3) methods to improve topic modeling, topic summarization, or novel information retrieval in the medical domain¹⁶–¹⁹. Particularly, three recent papers suggested that such techniques can be applied to extract medication-related information from unstructured text. Yin et. al. showed, in two papers, that topics derived via hierarchical clustering of word embeddings over patient-provider communications can be used to predict initiation²⁰ and discontinuation²¹ of hormonal adjuvant therapy in breast cancer patients. Beam et. al.²² applied LDA to summarize information from clinical notes that would influence a provider’s prescription of sleep medication. All three studies achieved relatively high predictive accuracy with their primary models, as well as developing a set of topics that include information on side effects, patient-provider communication styles regarding medication, comorbidities, and medical treatments that might provide insights into medication-related behaviors.
Despite these notable findings, little other research exists on using topic modeling methods to assess the outpatient medication-related content of unstructured EHR text—especially for notes from patients with a life-altering disease, such as breast cancer. Identifying determinants and patterns of patient medication-taking behavior are an active research area, but primarily rely on qualitative approaches on limited amounts of data. An automated method could improve the volume and variety of documents reviewed for these studies.

However, LDA has significant inherent limitations related to modeling topics in clinical corpora. One is an assumption of total independence between topics co-occurring in the same documents, which is unrealistic in medical notes, where topics on diseases would be expected to co-occur with their comorbidities, treatments, and symptoms. Correlated topic modeling (CTM) addresses this limitation by treating topics as draws from a logistic normal distribution that can have significant correlations with each other when they co-occur within documents.

CTM, like LDA, cannot model how topic prevalence may change over time, as the content of patients’ clinical notes are expected to change over time with disease progression and treatment resolution. Structural topic modeling (STM) extends CTM to include document-level metadata in topic models to gain more insightful interpretability regarding topic prevalence. As for clinical notes, time—whether an absolute timestamp, or a relative time from a patient’s diagnosis—can be considered a form of metadata to investigate how topic longitudinal patterns evolve.

To understand the characteristics and quality of medication-related information in clinical notes, and examine the viability of using topic modeling to retrieve such information, we propose a two-stage topic modeling process over a set of clinical documents derived from patients with breast cancer treated at Vanderbilt University Medical Center (VUMC). This approach enables the topic models to separate general medication information into more nuanced topics by removing extraneous documents from a large clinical note corpus to focus on notes that are most likely to relate to medication and medication use. We refine our models through STM to identify significant signals in patients’ long-term care. Our work demonstrates how a two-stage topic modeling approach effectively extracts condition-specific medication information from the EHR and reveals longitudinal patterns in note information content over time from diagnosis with breast cancer.

**METHODS**

In this retrospective longitudinal study, we successively implement unsupervised machine learning techniques to reduce the high-dimensional, sparse latent space of unstructured clinical notes and identify medication-related topics. Figure 1 summarizes our research pipeline, which consists of four major components: data extraction, data preprocessing, model training, and model evaluation.
Data extraction

This study used de-identified EHR data from the VUMC Synthetic Derivative (SD). We restricted our initial study cohort to patients who: 1) received at least one breast cancer-related ICD 9/10 code (174.x, 175.x, C50.x); 2) had at least one record including demographic, condition, and medication information; 3) were at least 18 years of age in 2010; 4) were born in 1923 or later and, if deceased, died after 2009; and 5) had at least one available clinical note between 2010 and 2021. The clinical note had to contain at least one of the following case-insensitive keywords/stems: ‘adher’, ‘compliance’, ‘compliant’, ‘comply’, ‘drug’, ‘med’, ‘pharm’, ‘pill’, ‘presc’, ‘rx’, and ‘side effect’. Filtering notes for these keywords narrowed our initial note corpus to texts that likely contained medication-related information.

As we sought to model a broad range of medication-related topics, we set the first breast cancer diagnosis date as year 0 to align each patient’s clinical notes, such that the clinical notes and medical history data before (after) the first breast cancer diagnosis of each patient would have a negative (positive) year value. The date of first breast cancer diagnosis for each patient was defined as the earliest date a breast cancer ICD code was received.

Data preprocessing

Raw note data were cleaned and tokenized using a combination of NLTK (version 3.5) and hand-written regex patterns in Python 3 (version 3.8). Tokens were not stemmed or lemmatized as recent research indicated such practices may harm topic model performance. The term ‘repnumber’ was substituted for all numeric quantities, while ‘meddose’ was used for combinations of a number and mg or ml. Additional substitutions included ‘rttiming’ for common abbreviations for scheduled medication timing such as ‘qid’ and ‘bid’, and ‘admroute’ for the abbreviations ‘po’ and ‘iv’. The term ‘prn’ (“pro re nata” as needed, for medications patients take at their own discretion) was retained, as “prn medications” and “scheduled medications” constitute different classes where patient behavior is concerned. Some partially numeric terms such as ‘o2’ and ‘co2’ were expanded to ‘oxygen and carbondioxide, respectively) and hyphens were replaced with a null character to collapse terms such as ‘x-ray’ or ‘j-tube’. Stop words—all the common English set from NLTK as well as a corpus-derived set defined by the authors—were removed.

To mitigate the negative impact of redundant information in clinical notes, we incorporated redundancy into our research pipeline. Our approach followed the work in Cohen et al., where redundancy is calculated as the fraction of character sequences in a patient's notes identical to those in their previous notes. In our initial experiments, we found that removing fully redundant documents (redundancy score equal to 1) was best for model performance, as it improved topic semantic coherence (high probability words in a topic frequently co-occur in a document) and exclusivity (high probability words in one topic are exclusive to that topic) while also increasing the lower bound of the maximum model likelihood. Therefore, after calculating the redundancy score of each note, we removed all completely redundant notes from our corpus in the analysis. Finally, tokens which had fewer than three characters and which did not correspond to common clinical terms (see Appendix for list of included tokens <3 characters in length), or tokens appearing in fewer than 100 documents, were removed.

Model training and evaluation

After cleaning the data, we relied on topic modeling to capture nuanced descriptions of medication-taking behavior in clinical notes. We applied multiple rounds of model training for corpus refinement and hyperparameter tuning. All the topic modeling models were trained and evaluated using the stm package (version 1.3.6).

Because the original data may contain notes that have little connection with medication-taking behavior, directly tuning the hyperparameters (e.g., a large range of topic candidates) in this dataset would be time-consuming and generate many irrelevant topics. To mitigate this issue, we designed a two-stage topic modeling pipeline. In the first stage, we refined our corpus to documents with a high prevalence of medication-related terms. To do this, we trained an initial “filtering” topic model, from which we selected all medication-related topics to form a composite medication-relatedness score for each document in the corpus. Filtering the corpus on varying thresholds of this score allowed us to create varyingly medication-related corpora for the second round of model training.

We chose a 30-topic CTM model for our filtering model based on preliminary modeling work which showed that this number of topics yielded a model with good semantic coherence and exclusivity (e.g., it developed several medication-specific topics while not splitting medication-related terminology across too many topics). After training this model, we manually reviewed its topics to identify those related to outpatient medication use. Authors KK and TB independently identified topics they believed represented outpatient medication-related content, then compared their lists. Topics on both lists were immediately accepted as being medication-related, while discrepant topics were discussed until consensus on inclusion or exclusion from the score was reached. Once consensus was achieved on the
medication-related topics, we computed the cumulative gamma value $\gamma_M$ of these topics for each document following Equation (1), creating its medication-relatedness score. $\gamma_M$ scores how much each note in the corpus discussed medication and, potentially, medication-related behaviors.

$$
\gamma_M = \sum_{t \in T} \gamma_t
$$

where $T$ represents the set of topics containing medication-related information

In the second stage, we trained four STM models (of 30, 45, 60, and 75 topics) on each of four different corpora: the original corpus, and three corpora filtered by thresholds of $\gamma_M >0.1$, $>0.2$, and $>0.5$, respectively. The metadata that each STM model was adjusted for included: 1) patient gender, race, and ethnicity; 2) the year the note was documented; 3) the patient’s age when the note was documented; and 4) the number of years from the patient’s first breast cancer diagnosis to the note date, which can be either positive or negative. This feature also allowed us to evaluate temporal topic prevalence relative to the patient’s initial breast cancer diagnosis. The note year served as an adjustment for temporal changes in the standard of care.

We evaluated each model’s residuals, lower bound of the maximum model likelihood, and topic semantic coherence and exclusivity to select a best-performing model. Author JW provided expert annotation of the topics from that model, after which we investigated how topic prevalence varied by the STM model’s metadata.

**RESULTS**

**Study cohort**

Our initial and final study cohorts consist of 1,306 and 1,031 patients, respectively. These patients are primarily White, non-Hispanic females with an average age of 60 years when first diagnosed with breast cancer. The cohort additionally includes a nontrivial proportion of male and Black patients. Table 1 contains cohort demographics for the two study cohorts. It should be noted that the statistics for the final patient cohort are calculated based on the note corpus with $\gamma_M >0.5$, as this corpus was determined to produce the best resulting topic models (see below).

**Table 1.** Cohort demographics.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Raw cohort</th>
<th>Final patient cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count (proportion), $n = 1,306$</td>
<td>Count (proportion), $n = 1,031$</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>1,193 (0.91)</td>
<td>931 (0.90)</td>
</tr>
<tr>
<td>Male</td>
<td>113 (0.09  )</td>
<td>100 (0.10)</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>1,054 (0.81)</td>
<td>862 (0.84)</td>
</tr>
<tr>
<td>Black</td>
<td>130 (0.10 )</td>
<td>111 (0.11)</td>
</tr>
<tr>
<td>Asian</td>
<td>6 (0.004)</td>
<td>6 (0.006)</td>
</tr>
<tr>
<td>Other</td>
<td>8 (0.006 )</td>
<td>5 (0.005)</td>
</tr>
<tr>
<td>Unknown/Not reported</td>
<td>108 (0.08)</td>
<td>47 (0.05)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>9 (0.006)</td>
<td>9 (0.009)</td>
</tr>
<tr>
<td>Non-Hispanic</td>
<td>1,182 (0.91)</td>
<td>972 (0.94)</td>
</tr>
<tr>
<td>Unknown/Not reported</td>
<td>115 (0.09 )</td>
<td>50 (0.05)</td>
</tr>
<tr>
<td>Age at first breast cancer diagnosis (years)</td>
<td>Mean (SD)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>59.6 (13.3)</td>
<td>60.1 (13.2)</td>
</tr>
</tbody>
</table>

From the raw cohort, we obtained 145,934 clinical notes that included at least one of the keywords. The number of notes per patient follows a long-tailed distribution, where 90% of the patients have 283 notes or less (Figure 2). This distribution is preserved for the corpus reduced by the redundancy filter and for our final corpus, which was further reduced by the medication-relatedness score filter. In the final corpus, 90% of patients have 66 notes or less.

**CTM model to filter clinical notes**

After applying the redundancy filter to our initial corpus, we were left with 124,347 clinical notes. We identified seven topics (1, 2, 3, 5, 12, 14, and 15, see Appendix: Filtering CTM Topics) as unambiguously containing content related to outpatient medications. Topics 23 and 29 also appeared to also contain medication-related content, but on review, we determined that 23 contained primarily words related to lab results and 29 was a mixed topic containing many terms unrelated to medication.
on the same grounds as topic 18, or conversely, topic 18’s inclusion in calculating $\gamma_M$ as it contained substantial medication-related content. However, subsequent testing demonstrated that including topic 11, but not topic 18, in calculating $\gamma_M$ improved the subjective quality of topics in the second-stage models. Figure 3 depicts the correlation structure (cutoff $p = 0.01$) of topics from the filtering CTM, showing topic 18 to be uncorrelated with all other topics, while topic 11 clusters with many of the other chosen topics, further justifying its inclusion in calculation of $\gamma_M$. We ultimately selected topics 1, 2, 3, 5, 11, 12, 14, and 15 to calculate $\gamma_M$. These topics contained a mix of generic (e.g., ‘capsule’, ‘meddose’, ‘oral’, ‘prn’, ‘tablet’) and condition-specific (e.g., ‘insulin’, topic 5; ‘amiodarone,’ ‘coumadin’, topic 1) terms.

**Hyperparameter selection**

Next, we trained 16 STM models on corpora further refined by $\gamma_M$ values derived from topics of the previous CTM model. The 16 models differed in the training corpus (per the $\gamma_M$ threshold) and the number of topics. Figure 4a, b, and c display diagnostics for the 16 models. The number of notes for each $\gamma_M$ threshold was as follows: 124,347 for 0; 74,447 for 0.1; 54,807 for 0.2; and 26,363 for 0.5. All two-stage models ($\gamma_M$ threshold > 0) demonstrated performance improvements over the single-stage models ($\gamma_M$ threshold = 0). Topics 11 and 18 proved more difficult to classify. We initially excluded topic 18 from $\gamma_M$ as it primarily contained terms related to infusion of chemotherapy medications in a (presumably) monitored setting, while including topic 11 as it contained generic medication-related terms. However, more careful review of topic 11 showed it contained specific terms (e.g., ‘levophed’, ‘norepinephrine’, ‘vancomycin’, ‘electrolytes’) for medications administered intravenously in an ER or inpatient setting. This argued for topic 11’s exclusion
We found models trained on the corpus of notes with $\gamma > 0.5$—the smallest corpus, with 26,363 notes—have the largest model likelihood lower bound, the smallest residuals, and competitive semantic coherence scores with other models, and so focused on these models. Among the models trained on notes where $\gamma > 0.5$, we found the model with 60 topics showed the best balance between exclusivity and semantic coherence (Figure 4d).

**Topics of the selected STM model**

Among the model’s 60 topics, some topics related to condition-specific medications while others seemed to capture aspects of breast cancer progression and treatment. We display the most probable terms and the most frequent and exclusive (FREX) terms for some of these topics in Table 2. The full set of topics is available in the Appendix: Best STM Topics.

**Table 2.** A collection of breast cancer and medication-related topics from the best STM model.

<table>
<thead>
<tr>
<th>Topic</th>
<th>High Probability Terms</th>
<th>High FREX Terms</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>meddose, insulin, units, diabetes, daily, tablet, left, unit, also, mouth</td>
<td>lancets, pen, insulin, ultra, humalog, lantus, strips, onetouch, sugar, novolog</td>
<td>Diabetes management</td>
</tr>
<tr>
<td>20</td>
<td>inr, warfarin, coumadin, visit, diagnosis, anticoagulation, dose, week, lovenox, dosing</td>
<td>inr, warfarin, fri, tues, coumadin, wed, mon, thurs, sun, anticoagulation</td>
<td>Anticoagulation, including weekly timing of coumadin/warfarin doses and INR testing</td>
</tr>
<tr>
<td>22</td>
<td>disease, r, l, breast, since, treatment, metastatic, meddose, history, progression</td>
<td>view, breepnnumber, restaging, scans, life, palliation, initiation, preservation, understood, participation</td>
<td>Breast cancer status</td>
</tr>
<tr>
<td>36</td>
<td>pain, meddose, oxycodone, medication, trttiming, use, continue, pm, medications, history</td>
<td>oxycontin, oxycodone, fentanyl, worst, patches, patch, breakthrough, dilaudid, prescribed, efficacy</td>
<td>Pain management and pain medications; patient descriptions of pain</td>
</tr>
<tr>
<td>37</td>
<td>meddose, day, trttiming, negative, denies, transplant, gvhdl, started, today, adnroute</td>
<td>gvhdl, cmv, engraftment, pbset, donor, csa, fk, antinfective, valtrex, allo</td>
<td>Immunosuppressives and management of transplant patients</td>
</tr>
<tr>
<td>39</td>
<td>er, bone, metastatic, meddose, breast, fslodex, hours, diagnosis, pain, r</td>
<td>xgeva, faslodex, counter, [identifier], ibrance, aromasin, letrozole, afinityor, exemestane, unexpected</td>
<td>Treatment of hormone-positive metastatic breast cancer</td>
</tr>
<tr>
<td>43</td>
<td>inhaler, meg, daily, nasal, actuation, day, spray, meddose, albuterol, puffs</td>
<td>inhalation, inhaler, puffs, hfa, albuterol, aerosol, puff, spiriva, advair, mcgrepnumber</td>
<td>Inhaled medications for respiratory illness</td>
</tr>
<tr>
<td>47</td>
<td>breast, left, right, biopsy, carcinoma, lymph, axillary, node, grade, negative</td>
<td>mammogram, birads, nipple, histologic, intermediate, ultrasound, extent, receptor, proliferative, breasts</td>
<td>Initial diagnosis and management of breast cancer in the pre-treatment period</td>
</tr>
<tr>
<td>53</td>
<td>pleural, effusion, er, meddose, negative, disease, since, metastatic, diagnosis, pr</td>
<td>thoracentesis, pleural, effusion, individual, gdcrapnumber, pericardial, effusions, cfr, tcr, regulation</td>
<td>Metastatic breast cancer; pleural effusions are a common site of metastasis</td>
</tr>
<tr>
<td>60</td>
<td>radiation, brain, treatment, mri, oncology, cgy, lesion, dose, metastatic, resection</td>
<td>cgy, brain, fractions, frontal, simulation, radiation, onc, srs, cerebellar, roncop</td>
<td>Brain metastases from breast cancer</td>
</tr>
</tbody>
</table>

**Figure 5.** Topic proportion over time from breast cancer diagnosis for breast-cancer-related latent topics.

These topics contained common, nonspecific terms, such as dosing (‘meddose’) and timing (‘trttiming’) information alongside more nuanced, disease-specific information. The anticoagulation topic (Topic 20), for instance, contained terms for weekdays, relating to specific timing of anticoagulant medication administration.

The STM also revealed how breast cancer-
specific topics change in proportion in the corpus over time from a breast cancer diagnosis. Figure 5 displays topic proportions for several breast-cancer-related topics plotted against year from first breast cancer diagnosis. The sharpest increase at the time of diagnosis was seen in topic 47 (initial diagnosis and management of breast cancer), while topics 53 and 60, with terms for metastatic cancer, grew in proportion further from patients’ initial diagnoses.

![Figure 5](image)

**Figure 5.** Topic proportion over time from breast cancer diagnosis for latent topics covering medications used for various comorbid diseases (left) and pain (right).

A different longitudinal pattern was seen in topics for patients’ existing comorbidities. The left side of Figure 6 shows topic proportions over time from breast cancer diagnosis for four comorbidity topics (18, 20, 37, 43) captured by our model. This “bow-shaped” curve showed a sharp reduction in topic proportion in the period immediately around patients’ breast-cancer diagnoses, followed by a local maximum in proportion somewhere between 0 and 5 years following diagnosis, a reduction to a local minimum, and then a gradual increase in topic proportion again. The pattern is much more pronounced for the topic relating to post-transplant immunosuppression, compared to the other three topics, but has the same relative shape in all. By contrast, topic 36 (pain medication and management), seen at the right of Figure 6, increases in prevalence just prior to diagnosis and remains relatively stable in proportion afterward, though it shows some of the same bow-shaped pattern as the others.

**DISCUSSION**

In this study, we demonstrated how a two-stage unsupervised topic modeling pipeline can extract specific topics about patient medication use in clinical documents drawn from the EHRs, which substantially improved model performance over a single-stage model trained on a large document corpus. This supports the use of latent topic models to filter documents of interest out of large research corpora used in training other language or machine learning models.

After using two preliminary steps for filtering our corpus to texts with the highest concentration of medication-specific discussion, we obtained a model that captured several latent topics with high specificity for terms around medication use. These included topics for several diseases and their treatments (diabetes in topic 18, anticoagulant therapy in topic 20, post-transplantation immunosuppression in topic 37, and inhaled medications in topic 43), anti-cancer treatments (topics 17, 23, 26, 30, 35, 39, 59), pain control and opioid medications (topic 36), pharmacy refill requests (topics 14 and 31), and nuanced information relating to the timing of medication (topic 20).

While our filtering criteria were highly specific for medication-related terms, the second-stage topics also captured other areas of patient experience such as ability to perform activities of daily living (topic 2), nutrition needs for patients with difficulty eating (topic 7), urgent care consultation in the context of severe symptoms (topic 32), and non-pharmaceutical pain management (topic 38). All of these topics may represent barriers to patients taking their medications, thus forming rich additional context on patient medication taking-behavior.

With the addition of a temporal covariate to the STM, we found that the prevalence of topics could vary by time from a patient’s breast cancer diagnosis. Figure 5 shows that topic prevalence for topics related to breast cancer progression and staging rises sharply at initial diagnosis, then fluctuates over time as treatment milestones or disease progression occur. Topic proportions of medication-related topics might therefore be expected to increase or subside with the presence and severity of an underlying condition in other diseases.

The bow-shaped pattern seen in Figure 6 further suggests that the diagnosis of a new, severe condition causes significant changes in how existing conditions are discussed. The proportion of topics concerning pre-existing comorbidities dips concurrent with a diagnosis of breast cancer, reflecting how medication-related notes during this
time period will have much more breast-cancer-related content and less focus on existing conditions, then rises again in the following years when a patient is under treatment (potentially corresponding to exacerbation of these conditions and/or complications arising from breast cancer treatment), then declines at the five year mark (which is often a milestone for cessation of adjuvant therapy in breast cancer) only to rise again after. If these changes in topic proportions also reflect neglect or exacerbation of the treatment of pre-existing conditions, then it may be reasonable to conclude that they may be predictive of changes in medication behavior for those conditions as well. This could include deliberate cessation of medications that interact poorly with new treatments or unintentional cessation due to patient lifestyle changes while dealing with a new, life-altering disease. Cheng and Levy demonstrate that even patients with stage I breast cancer have a high treatment workload\textsuperscript{18}, with that workload increasing with higher disease staging. Insofar as treatment burden affects patients’ ability to manage their own care\textsuperscript{4}, our results show that signals of this difficulty may be detected in medication-related note text using a wholly automated method.

The emergence of a topic related to transplantation and immunosuppression was surprising, considering that these do not represent standard current treatment approaches for breast cancer. We initially surmised one of two possible explanations for this: the non-trivial rate of secondary leukemias resulting from cytotoxic chemotherapeutics used to treat breast cancer\textsuperscript{19} and subsequent allogenic stem cell transplant to treat them; or the use of autologous stem-cell transplant following high-dose chemotherapy, a regrettable approach popularized in the 1990s that took many years to debunk\textsuperscript{38}. Subsequent examination of notes with large $\gamma$ values for this topic showed that many of the highest-scoring notes were for patients with a leukemia presumed secondary to prior breast cancer treatment. Meanwhile, we further discovered that patients’ first billing code for breast cancer often corresponded with a note date for initial work-up of the leukemia, indicating that this was the first mention of a prior, resolved breast cancer in medical notes at our institution. This may explain the sharp spike in this topic’s prevalence at year 0 from breast cancer “diagnosis” (Figure 6) and argues for using more precise techniques to determine patients’ initial diagnosis dates for a disease of interest.

Our results suggest that two-stage topic modeling produces nuanced, human-comprehensible latent topics that may be suitable to use as features for studying and predicting patient medication behavior. In addition, when examined in the context of a disease-specific time measure, these topics also reveal how mentions of patient medication behavior are affected by changes in a patient’s condition. Longitudinal changes in medication-related topic proportions might thus be applied to predict changes in patient medication behavior.

### Limitations and future directions

Despite our notable findings, there are several limitations we acknowledge here. Although we worked on a large patient cohort, this is a single-institutional study that may be biased to geographic, temporal, and care delivery particularities. Additionally, we used a very liberal criterion of one breast cancer code to identify the study cohort. While this increased the size and variety of the note corpus, it included notes from patients who did not have breast cancer, which weakened our conclusions about breast cancer patients. However, our final model showed a variety of high-quality topics concerning breast cancer even with the presence of these false positives. Future research should include a more robust phenotype definition as well as more temporal precision in resolving a patient’s actual date of diagnosis.

Technically, we noted that several poorly processed words entered our final topics, which might have been avoided with better preprocessing to remove note metadata. A lack of standardization of terms or expansion/replacement of common medical acronyms (such as those used for timing or route of medications) also rendered our final topics less interpretable than they could have been. However, these topics did conclusively demonstrate the proposed two-stage pipeline’s ability to capture medication-specific behaviors and outcomes, supporting our overall conclusion. Future research in this area is recommended to explore other ways of excluding metadata terms from our text corpora.

Another limitation that could be addressed in future research is that we did not consider multi-word phrases (e.g., bigrams, trigrams) in our model. Inclusion of bigrams in our models would have made topics dealing with, say, side effects easier to detect because they would have included the term “side effect” versus simply “effect”. However, including n-grams becomes very computationally expensive as well.

Finally, we found the subjective quality of the second-stage models to be highly sensitive to the specific topics selected from the first-stage filtering model for generating the medication-relatedness score, and therefore the particular subset of the corpus selected to train on. Inclusion of an uncorrelated topic with a group of correlated topics in calculating $\gamma_M$ greatly reduced subjective topic quality, suggesting future sensitivity analyses on the extent to which the topic correlation affects the filtering model performance.
CONCLUSIONS
In this paper, we presented a two-stage topic modeling pipeline for extracting latent topics related to medication use in clinical notes for a cohort of breast cancer patients. We found relevant, human-interpretable latent topics capturing medication use and patient experience with several medical conditions. In addition to demonstrating the value of the two-stage approach to modeling, the inferred topics showed demographic and temporal associations that may be beneficial to future researchers who wish to extract information on patient medication use from EHRs.

ACKNOWLEDGMENTS
This publication was supported by the National Center for Advancing Translational Sciences (UL1TR000445) and the National Cancer Institute (R37CA237452) of the National Institutes of Health. We gratefully acknowledge Dr. Tom Lasko for help in gaining data access to the VUMC SD. We also acknowledge Julia Silge for her excellent tutorial on the `stm` package*.

REFERENCES

* Available at https://juliasilge.com/blog/evaluating-stm/

APPENDIX

All topics from the best STM model (60 topics, documents with $\gamma_t>0.5$) and the filtering CTM model can be found at https://bit.ly/3xwarPO. The third worksheet contains a list of words <3 characters that were included in the final token list.
Tracking the COVID-19 outbreak in India through Twitter: Opportunities for social media based global pandemic surveillance

Sahithi Lakamana, MS¹, Yuan-Chi Yang, PhD¹, Mohammed Ali Al-Garadi, PhD¹, Abeed Sarker, PhD¹

¹Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, GA 30322

Abstract

We investigated the utility of Twitter for conducting multi-faceted geolocation-centric pandemic surveillance, using India as an example. We collected over 4 million COVID19-related tweets related to the Indian outbreak between January and July 2021. We geolocated the tweets, applied natural language processing to characterize the tweets (e.g., identifying symptoms and emotions), and compared tweet volumes with the numbers of confirmed COVID-19 cases. Tweet numbers closely mirrored the outbreak, with the 7-day average strongly correlated with confirmed COVID-19 cases nationally (Spearman r=0.944; p=0.001), and also at the state level (Spearman r=0.84, p=0.0003). Fatigue, Dyspnea and Cough were the top symptoms detected, while there was a significant increase in the proportion of tweets expressing negative emotions (e.g., fear and sadness). The surge in COVID-19 tweets was followed by increased number of posts expressing concern about black fungus and oxygen supply. Our study illustrates the potential of social media for multi-faceted pandemic surveillance.

Introduction

The global outbreak of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), commonly referred to as COVID-19, has been one amongst the worst pandemics known the World history¹, resulting in adverse social, political and economic consequences around the globe². Due to the unprecedented nature of this pandemic, traditional public health surveillance methods struggled to tackle this scenario, and different strategies were adopted by different governments around the world for conducting surveillance³–⁵. As of August 2021, COVID-19 outbreaks continues globally, with the highly-infectious B.1.617.2 (delta) variant being the driving force behind the waves of outbreak at the time⁶–⁸. This variant, which was first detected in India, ravaged the country resulting in the largest national lockdown across the world⁹. As of 20th July, 2021, India reported over 31 million confirmed cases and over 400,000 deaths. The rapid prevalence the outbreak in India, and other countries, exposed the weaknesses of traditional surveillance systems, exhibiting that most traditional mechanisms were not designed to meet the challenges of this pandemic¹⁰. For example, national surveillance methods based on testing symptomatic people, which worked effectively for certain countries, are unrealistic for others¹⁰,¹¹. The latest research suggests that the vaccines developed for COVID-19 are effective for the delta variant⁶, but considering the unvaccinated population across the world, it is likely that outbreaks caused by this variant will continue around the world. Even within largely vaccinated communities, many breakthrough infections have been reported due to this variant¹². There is also a possibility of new variants of the virus emerging over time, causing future outbreaks, and perhaps at faster rates. Consequently, there is a need to identify and deploy novel surveillance methods that can complement traditional surveillance approaches¹³.

One potential resource for conducting effective, real-time surveillance of COVID-19 is social media. Currently, social media adoption and usage around the globe is at an all-time high¹⁴, and social media has a global reach, with hundreds of millions of monthly active users. Despite its potential, social media has been largely underutilized for conducting close to real-time surveillance. While social media has potential negative aspects, which have been highlighted in recent literature¹⁵, the opportunities it present have not been explored sufficiently. Due to the huge global user base of social media, and the recent advances in data-driven information management approaches, such as natural language processing (NLP) and machine learning, this resource may enable the real-time monitoring of localized outbreaks. In addition to detecting outbreaks, the knowledge generated from social media may be used for conducting syndromic surveillance and understanding population-level perceptions about the pandemic. The knowledge generated over social media have been utilized in recent research for a variety of tasks such as sentiment analysis¹⁶, pharmacovigilance¹⁷, toxicovigilance¹⁸, studying mental health-related topics¹⁹, and other health related tasks²⁰.

In this study, we explored the potential utilities of social media for close to real-time pandemic surveillance. We use the recent outbreak caused by the delta variant in India to explore the potential utilities of social media. We specifically utilize the data generated on Twitter, and apply NLP methods to explore utilities beyond simple outbreak
detection. Our retrospective analyses of Twitter data illustrate that it may potentially be used for gaining near real-
time insights about various aspects of the pandemic. In this paper, we discuss the following possibilities:

- Utilizing the volume of COVID-19 data generated on Twitter, combined with geolocation-related metadata, for
detecting outbreaks.
- Conducting syndromic surveillance through the active monitoring of user-reported symptoms on Twitter via the
use of NLP.
- Assessing the mental/psychological impacts of outbreaks via automated emotion analysis of COVID-19-related
Twitter chatter.
- Detecting emerging concerns related to the pandemic in targeted populations via automated Twitter chatter
analysis.

We present our methods and findings in the following subsections. We have also publicly reported our findings here
publicly available via an interactive, web-based dashboard.†

Materials and methods

Study setting

Our study is based on data generated publicly on the social network Twitter. Twitter is one of the most popular social
networks in the world, with close to 400 million monthly active users in 2021. Twitter is a particularly attractive
resource for real-time data analysis because most of the data on this platform is public. Posts on Twitter, referred to
as ‘tweets’, are essentially ‘microblogs’ and are typically publicly available. Posts on Twitter may also have associated
meta-data, which can be leveraged to obtain additional information, such as geolocation-specific statistics.

Data collection

We collected data from Twitter via its streaming application programming interface (API) that was specifically created
for conducting COVID-19 research. The COVID-19 stream API was released by the company to enable researchers
study the conversation surrounding COVID-19, and the authors of this manuscript were granted special permission to
access the stream. Unlike the traditional public streaming API, which only provides access to a sample of tweets
posted at any time, the COVID-19 API stream delivers all the conversations about the topic without any rate
limitations. We collected all posts in English from this API using COVID-19-related keywords (eg., ‘COVID’,
‘COVID19’, ‘corona virus’),‡ and utilized the metadata associated with the tweets to geolocate the sources,
specifically tweets that originated from India. When available, we used the geolocation coordinates of tweets to
identify specific regions of India from which each tweet originated. We also collected data about India during this
period by adding a second layer of filter containing term ‘India’ to understand global response. For the experiments
described in this paper, we used data collected in this manner from the beginning of 2021 until July 2021. The
streaming big data was stored in a mongoDB database and NLP methods were applied to derive insights from the
data.

Data analyses

We conducted analysis to explore several aspects of the Indian outbreak, specifically (i) outbreak timeline and location
analysis, (ii) real-time syndromic surveillance, (iii) population-level emotion analysis during the outbreak, and (iv)
emerging topic detection. We described the methods applied for the specific analyses in the following subsections.

Outbreak timeline and location

We used the volume of COVID-19 related tweets over time geolocated from India to track the outbreak timeline. We
compared the timeline with other events in India (eg., opening of public places such as shopping complexes and movie
theatres). Whenever possible, we mapped the tweets to the states from which they were posted. We used 2 methods
to detect the geolocation origins of the tweets. First, for tweets that had geolocation coordinates available, we mapped
them to the specific state in India. For tweets that did not have geolocation coordinates in the meta-data, we used an
existing package called geo-carmen. This package uses the meta-data provided by the Twitter API to extract their
location from their geo-tagging information, user profile etc. The geolocation information is further possibly

segmented into country, state and county level. This helps in identifying the precise locations and to identify patterns specific to certain areas.

We additionally used a previously-developed machine learning classifier to detect tweets that represented self-reports of COVID-19 positive tests (i.e., users who reported that they had tested positive for COVID-19). The classifier was trained to use posts from Twitter which mentioned COVID-19-related keywords. These posts were manually annotated to indicate self-report or otherwise. To prepare the texts of the tweets for this classifier, we had to perform some basic preprocessing on the text. We first tokenized the texts by breaking lines into words, called tokens, followed by converting all the data into lower case. We also removed stopwords, punctuations, numbers, extra spaces, and special characters using “cleantext” package. The detection of self-reports was modeled as a binary classification task, and we applied a state-of-the-art method called bidirectional encoder representations from transformers (BERT) on manually annotated data.

In addition to tracking posts originating from India, we also tracked the global response to the Indian COVID-19 outbreak by identifying tweets emerging from outside India that contained both: COVID-19 related keywords and references to India. We geolocated tweets at the country-level to identify which countries, other than India, had high interest in the outbreak.

**Syndromic surveillance**

We applied a previously-developed COVID-19 Twitter symptom lexicon to detect specific symptoms that were reported by users from India. The lexicon contains non-standard expressions and misspellings that are commonly found in social media data. To detect symptoms from the text, we applied NLP to perform inexact matching. This enabled us to detect symptom expressions that were lexically similar to those encoded in the lexicon, but not necessarily identical. The lexicon was used to map symptom expressions to standardized IDs in the Unified Medical Language System (UMLS). We then computed the frequency of each symptom.

**Emotion analysis**

Our intent was to analyze the emotions expressed via tweets before, during and after the outbreak, that could also potentially detect changes in emotions over time. We performed linguistic emotion analysis of the tweets using the lexicon curated by the National Research Council, Canada, which contains a comprehensive list of approximately 14,182 English words related to anger, fear, anticipation, trust, surprise, sadness, joy, sentiment (negative and positive), and disgust. In addition to this, we also quantified the aggregated anxiety levels over time, as expressed by the tweets. We tried to identify the changes in user’s emotions pre-outbreak and post-outbreak. We considered pre-outbreak period as January and February while March, April, and May as post-outbreak period. The intensity of each emotion is measured between 0 to 1 at the tweet level, where 0 is the least and 1 to be the highest.

**Detecting emerging concerns**

We used frequency distributions to identify emerging concerns and interests associated with the outbreak. For detectable emerging concerns, we tracked their distribution over time by tracking frequencies of word bigrams and trigrams (Figure 1). Bi-grams and tri-grams are collectively called n-grams, which represent contiguous sequences of n words. This led us to discover multiple topics—black fungus, a disease that widespread incidence in India.
following the COVID-19 outbreak, and vaccine-related chatter, specifically represented by the keywords ‘CoviShield’ and ‘CoVaxin’, which represent the two vaccines that were available in India at the time.

**Results**

**Outbreak and location**

Between January and July 2021, we collected over 4 million tweets about the outbreak in India, of which over 500,000 were geolocated to be from India, with 9,700 having specific geolocation coordinates. Globally, 3.56 million tweets were posted on India from other countries. Figure 2 presents the timeline of COVID-19 tweets geolocated which have been posted from India between early January 2021 to early June, 2021. The figure also shows the timeline of daily confirmed COVID-19 cases in India, and the timelines for two important national events—the opening of public places such as shopping centers and state-level elections. From the figure, it can be seen that the daily volume of tweets closely followed the number of confirmed COVID-19 cases. We also found a statistically significant correlation (Spearman r=0.944, P=0.001) between the 7-day moving average of tweet count and 7-day moving average of COVID cases reported.

Figure 2. Comparison of weekly volume of COVID-19 related tweets from India (top) and the number of confirmed COVID-19 cases per day (bottom).

Figure 3 shows the state-level distribution of tweets during this timeframe. Darker shades represent higher numbers of tweets. Significant correlations for the COVID-19 cases in the states were found with the volume of Twitter data available (Spearman r = 0.84, p = 0.0003). The highest number of tweets were from the states Maharashtra (~24%), Karnataka (~11.5%), Uttar Pradesh (~7.5%) and Tamil Nadu (~7%). The correlation between tweet count and COVID-19 cases recorded for the top 4 states (Maharashtra, Karnataka, Uttar Pradesh, and Tamil Nadu) is also strong, however not statistically significant (Spearman r = 0.8, p= 0.200) due to the low number of available data points. Maharashtra, which also has one amongst the largest cities in India (Mumbai) also had the highest number of COVID-19 cases during this time, than the other states. The three other states with the highest number of tweets were among the next top 5 states in terms of highest numbers of COVID-19 cases. Kerala and Andhra Pradesh were the two other states that had high numbers of confirmed COVID-19 cases but relatively lower number of tweets. The number of self-reports detected via supervised classification was relatively low, peaking at 109 reports on April 18th. In total, 374 COVID-19 self-reports recorded in the first half of 2021. This finding is different from the other conducted by similar recent studies similar recent studies focusing on other geolocations, which showed high numbers of self-reports during early COVID-19 outbreaks.

**Syndromic surveillance**

The most commonly discussed/reported symptom was fatigue, followed by cough and dyspnea (shortness of breath). The number of mentions about of fatigue were more than double the next highest reported symptom (cough). Other detected symptoms included headache, anosmia (loss of smell) and loss of appetite. All these symptoms were among the top 8 symptoms of acute COVID-19 that
were detected to be reported by COVID-19 positive Twitter users. Relatively speaking, two symptoms that were underreported were pyrexia (fever) and body ache & pain.

*Emotion analysis*

Compared to the pre-outbreak time period, there was a detectable surge in negative emotions during the outbreak, namely fear, sadness, and anger (Figure 4). As much as six times the pre-outbreak count of tweets prone to fear, anger and sadness were posted during the months of April and May, which coincided with the outbreak period.

*Emerging concerns*

In addition to increased volume of COVID-19 tweets during the outbreak, there is a high levels of negative emotions, and high levels of anxiety expressed in the tweets, our NLP-driven analyses specifically discovered three topics that has emerged during the Indian outbreak—black fungus, oxygen supply, and COVID-19 vaccines. Many cases of mucormycosis, commonly referred to as black fungus, were detected in Asian countries, particularly India, following the outbreak of the delta variant of the COVID-19 virus\(^30\). In general, the main source of mucormycosis infections is from multiple contaminated sources and tend to infect diabetic patients quickly. As the number of COVID-19 cases increased in India, the number of black fungus infected particularly among diabetic patients, also increased\(^31\). Chatter about black fungus originating from India started rising from early May which peaked on the 20\(^{th}\) of the month, approximately two months after the outbreak-related chatter surged on Twitter (Figure 5a). The timeline of the rise in black fungus chatter coincided with the extraordinarily high numbers of post-COVID infections of cerebral mucormycosis infections in the country\(^32\).

Vaccine-related chatter also surged during this wave of outbreak in India, although, unlike black fungus, the increase in such chatter was steady from early 2021 and continued to increase during the outbreak months. Two prominent vaccines have been widely used in India—CoVaxin and CoviShield. Among the vaccine-related chatter, nearly 44% of the users discussed about CoVaxin while 56% about CoviShield. Among the vaccine-related chatter, early on in the year, CoVaxin was the most commonly discussed vaccine, but it was later surpassed by CoviShield (Figure 5b). Interestingly, the increase in CoviShield-related chatter relative to CoVaxin coincided with a reduction in supply of the latter, around late April. Another topic of interested identified through twitter chatter was oxygen supply. As the COVID-19 infections rapidly increased, India a faced crisis in oxygen supply. The requirement of oxygen was observed in the Twitter chatter as the topmost mentions about oxygen included oxygen supply, oxygen bed, oxygen concentrators, and oxygen cylinders, as shown in Figure 6.

*Early detection*

Our findings also suggest that Twitter may be used for predicting the outbreak of COVID-19, by detecting symptom-mentioning posts. We found a strong and significant correlation between the number of covid cases recorded and the counts of symptomatic tweets mentioned a week earlier in the chatter (Spearman \(r=0.897\), \(p=0.000\)). Early detection/prediction may have a significant impact by enabling us to estimate future hospitalization needs.

*Discussion*

Social media, specifically Twitter, chatter encapsulates information in abundance regarding COVID-19. The knowledge contained within this resource can potentially be leveraged to obtain real-time insights about the current pandemic and also future pandemics. Our explorations on large-scale Twitter data generated specifically during the delta variant outbreak in India suggests that such data can be utilized to obtain multifaceted insights, adding to the detection of geolocation-specific outbreaks. While we used the Indian outbreak as our chosen topic, the methods outlined in this paper may be applied to any specific region of the world and over any social network. For our study, Twitter was a suitable social network as India has the third highest number of Twitter users in the world, following the United States and Japan\(^33\). For our study, the COVID-19 streaming API made it possible to collect user-posted
data in real time. Most social networks provide APIs that can be leveraged to obtain real-time insights, thus, for monitoring pandemics in regions with lower numbers of Twitter users, the most relevant social networks can be used.

One of the primary challenges of using social media data for obtaining such multi-faceted insights is its difficulty of mining knowledge from the noisy text-based data that is generated. The text-based knowledge that is generated is typically hidden in large volumes of noise, non-standard terminologies, misspellings, and ambiguous expressions. While it is possible to extract relevant knowledge, by manual inspection of each data point, the large volume of data makes manual curation on a continuous basis impossible, particularly in real-time. Thus, such manual curation and analyses of data are generally retrospective in nature.

Importantly, manual curation is typically limited to small data samples, and relying on such curation methods takes away one of the major advantages of social media—the availability of big data.

Thus, leveraging social media data optimally requires the development of advanced data science, NLP and machine learning methods, which can effectively characterize streaming data. In addition to addressing the above-mentioned challenges associated with social media data, such approaches also need to address emerging problems on social media, as in misinformation often referred to as an infodemic.

One among the key findings of this work is the high correlation between the COVID-19 case numbers and the volume of Twitter chatter. States in India with higher numbers of COVID-19 cases also tended to have high volumes of chatter. We observed that states with large metropolitan cities such as Mumbai (Maharashtra), Bengaluru (Karnataka), and Chennai (Tamil Nadu) tend to produce higher volumes of chatter associated with the outbreak. There are two explainable reasons behind this: (i) large metropolitan cities are generally the epicenters of outbreaks, which was no different in case of the Indian outbreak; and (ii) such metropolitan cities also have large numbers of technologically adept, young people, who make up to the vast majorities of social media users. In addition to mirroring the outbreak, the Twitter chatter also revealed symptoms reported/discussed by the users, which may be used to conduct syndromic surveillance. In areas with low number of testing locations, social media based syndromic surveillance may provide early signals about upcoming outbreaks. Interestingly, although there was a large volume of COVID-19 related chatter emerging from India, unlike prior studies conducted on populations from other countries (eg., United States), we found low number of self-reports on COVID-19 positive status. While exploring the reasons behind this is outside the scope of our work, we suspect it might be because of stigma associated with COVID-19.

It is possible that the tweet volumes were in response to the rising case numbers, and not necessarily predictive.

In addition to outbreak and syndromic surveillance, our study shows that Twitter data can be effectively utilized to assess population-level emotions associated with the pandemic. The rise in negative emotions may not only be a response to COVID-19, but also to the social distancing and ‘lockdown’ measures implemented by regional governments. A number of recent studies have elaborated on the negative mental health consequences of the pandemic, and our study suggests that social media data at the time of an outbreak maybe utilized for better understanding on geolocation-specific mental health impacts (eg., by studying/detecting expressed emotions). Interventions or social programs may be guided by population-level mental health assessments at specific times and places. While we attempted to assess emotions in a relatively simplistic manner, more sophisticated approaches under the umbrella of sentiment analysis which might be employed for more targeted information. For example, sentiment
Analysis methods have been employed in the past to study people’s perceptions about vaccines\textsuperscript{45} and treatments\textsuperscript{46}. A recent systematic review discussed many applications of sentiment analysis approaches during pandemics, such as the current one, and infectious disease outbreaks\textsuperscript{47}.

Strengths and limitations

The methods we described in this paper have several advantages compared to more other traditional surveillance approaches. Firstly, the social media based surveillance methods discussed are purely data-centric, and they do not incorporate hypothesis-driven biases. For emerging health crises with many unknown issues, like the current COVID-19 pandemic, data-driven approaches can provide insights on topics that may have been overlooked under other circumstances. Secondly, streaming social media based monitoring can be done at real-time or close to it. This can significantly reduce the lag times associated with many traditional surveillance approaches that rely on tools such as surveys or require compiling data from multiple sources (eg., reports from testing centers). Rapid outbreaks, such as those observed during the COVID-19 delta variant, require rapid responses, and data-centric approaches over social media data may aid such processes. And finally, the large user base of social networks means that they can potentially provide access to hard-to-reach populations. While our study is exploratory by nature, the methods we applied may be built on to establishing participatory surveillance through social media\textsuperscript{10,48}. In our work, information flow was unidirectional, but there is a potential for establishing bidirectional social media based communication channels to complement the existing public health education, surveillance and intervention methods. Such bidirectional communication models have been explored in recent literature for targeted topics, such as the use of chatbots for mental health support\textsuperscript{49}, but their utility for pandemic surveillance and response has not been focused.

Our study has several limitations as well. The primary limitation stems from the user base of Twitter—users tend to be younger than the general population and is thus not representative of the general population. However, over recent years, social media adoption has increased significantly among older demographics\textsuperscript{50}. Also, our study only focused on tweets that were in English. This limitation was introduced because the NLP tools we employed were not designed for multilingual text processing. Recent research in the field of computational linguistics has focused on developing multilingual corpora to aid multilingual NLP, and similar efforts have been reported for clinical texts\textsuperscript{51}. Purely data-centric methods such as ours are also vulnerable to potential data manipulation by bots or automated accounts, and to misinformation. Both of these problems persist in all social media based studies, and currently researchers and social network administrators are actively engaged in reducing the impacts of these. Twitter, for example, has recently adopted a zero tolerance stance for accounts spreading misinformation, and actively suspends or closes such accounts.

Conclusion

Social media, specifically Twitter, chatter encapsulates an abundance of information regarding COVID-19. The knowledge contained within this resource can potentially be leveraged to obtain real-time insights about the current pandemic and may help in forecasting future pandemics. While our study focuses solely on India, the same methods can be applied to conduct real-time surveillance in other countries, including the United States. It must be noted that...
while we have outlined the utilities of social media for pandemic surveillance, we do not advocate the replacement of traditional methods of surveillance by social media based ones. Traditional surveillance methods, such as those relying on testing center numbers, hospital admissions, and contact tracing methods, to name a few, have been established over the years through evidence-based research. Our findings suggest that social media has high potential for complementing traditional surveillance methods, and as the user base of social media grows, the utility of such platforms may further increase in the future. Future research efforts should investigate further how social media can complement traditional surveillance methods, and also how they may be utilized for participatory surveillance and interventions.

References
7. Torjesen I. Covid-19: Delta variant is now UK’s most dominant strain and spreading through schools. BMJ. 2021;373:n1445. doi:10.1136/BMJ.N1445


CycleGAN with Dynamic Criterion for Malaria Blood Cell Image Synthetization

Zhaohui Liang, PhD Candidate, Jimmy Xiangji Huang, PhD
York University, Toronto, Ontario, Canada

Abstract

We present a cycle-consistent adversarial network (Cycle GAN) with dynamic criterion to synthesize blood cells parasitized by malaria plasmodia. The result shows 100% of the synthetic images are correctly classified by the pretrained classifier compared to 99.61% of the real images, 76.6% generated by the Cycle GAN without the dynamic criterion. The average score of Frechet Inception Distance (FID) of the generated images by the enhanced Cycle GAN is 0.0043 (Std=0.0005), which is significantly lower than the FID score of the variational autoencoder (VAE) model (0.0085 (Std=0.0007)). We conclude that the new Cycle GAN model with dynamic criterion can generate high quality malaria infected blood cell images with good diversity. The new method provides new augmentation technique to enhance the image diversity where the acquisition of well-annotated images is highly restricted, and to improve the robustness of medical image automatic processing by deep neural networks.

Introduction

Malaria is a tropical infectious disease caused by the infection of plasmodia. The microorganism is usually transmitted by mosquito bites and parasitizes in human blood cells (particularly red blood cells). According to the statistics by WHO, there are 409,000 deaths related to malaria in 2019 and the total death toll is accumulated to 7.6 million since 2000. The standard malaria fast screening diagnosis technique is microscopic malaria infected blood cell counting by medical professionals. This method is inefficient because it is not only time and labor consuming but also highly affected by individual expertise. To solve this problem automatic image processing technology has been applied to malaria diagnosis since 2005. In 2016, our team developed a convolution neural network (CNN) with 6 convolutional layers for classification of malaria infected blood cells. The CNN was trained with a dataset with 27,578 blood cell images (ratio: 1:1) and the average accuracy is 97.37%. The following studies also report extremely high classification accuracy. However, these results are all achieved based on a large, well annotated dataset for training the CNN models. In most cases, big annotated medical image datasets are difficult to acquired. If the medical images are annotated by non-medical persons, the quality of the image data is suspicious due to the lack of expertise. Therefore, we should seek for a solution to minimize the human expertise intervention to the deep neural network (DNN) optimization.

Another drawback of DNN is that the specific medical image patterns are different from general-purposed images such as those in the ImageNet dataset. As the result, when using transfer learning with DNN models trained by the ImageNet to fine tune a new model for the medical images, the pretrained feature extractors usually cannot effectively capture the medical significant patterns through the complex architecture but simply develop meaningless combinations for the final decision. In our previous work on CNN for the malaria blood cell image classification, the transfer learning approach has lower accuracy (91.99%) than the randomly initialized CNN (97.37%). A recent study reveals that the seemingly high-performance DNN models for COVID-19 chest X-Ray image detection are vulnerable from network attacks.

The common strategy to improve DNN performance is to enhance the diversity of training image data by augmentation such as random rotation, flipping, and jittering. However, these conventional augmentation methods are unsuitable to the most medical tasks like images of histological cells and tissues, or X-Ray photos. The image-based medical diagnosis usually requires structure completeness and correct image alignment because the diagnosis is usually based on the comparison between normal and abnormal structure. The random augmentation techniques are likely to break the structure completeness or position alignment. As a result, the DNN models are likely to capture wrong combination of patterns or artifacts instead of the correct ones matching the human knowledge. This is a possible explanation for the DNN vulnerability for medical image processing. As response to the challenge, we switch to apply the generative models for image augmentation. A generative model can learn the data distribution in the unsupervised manner, then it can generate new data with reasonable variations given the learned distribution. In this work, we respectively apply two approaches of generative learning: variant autoencoder (VAE) and generative adversarial networks (GAN). The
difference is that a VAE is to maximize the evidence lower bound (ELBO) of the data distribution, while the GAN is optimized by achieving an equilibrium between the generator model and the discriminator model. The merit of using generative models for image augmentation is that both VAE and GAN generate synthetic images with an acceptable extent of randomness, which can effectively simulate the real context of medical practice.

In the rest parts, we will briefly introduce the rationale of VAE and GAN, then we will present the details of our new cycle-consistent adversarial network (Cycle GAN) with a dynamic criterion to compute an extra loss term of the objective, the setting of the experiments, and the result of the comparison of the VAE, new Cycle GAN, and a conventional GAN. Finally, we will draw a conclusion based on the analysis and propose the solution for malaria blood cell synthetization.

Rationale and Methods

Deep generative learning or generative DNN, is an unsupervised learning approach to learn the training data distribution with a given DNN architecture, then it can generate new data points belonging to the learned distribution with random variance. However, the generative DNN cannot either explicitly or implicitly learn the identical distribution of the training data, but it can approximate the true parameters by different modeling techniques. There are two main methods for generative DNN: variational autoencoder (VAE) and generative adversarial networks (GAN). In this study, we mainly use the GAN approach to synthesize the malaria infected blood cell images. The VAE model is implemented for synthetic image comparison.

VAE is a generative model introduced in 20137. Given the observed dataset $X = \{x^{(1)}, x^{(2)}, \ldots, x^{(i)}\}$, a VAE is composed of two networks. The encoder is a DNN parameterized by $\phi$ to estimate the posterior distribution of the latent variable $z$ given $X$: $q_\phi(z|X)$, where the training data points are taken as observations to estimate the parameters of the conditional distribution of the latent representation $Z$. The decoder is another DNN parameterized by $\theta$ to estimate the conditional distribution of the observed data $p_\theta(X|z)$, where the input is a sample $z$ (usually the outputs from the encoder). The optimization objective of a VAE can be written as:

$$-\mathcal{L}(\theta, \phi; X) = \omega \cdot D_{KL}(q_\phi(z|X) || p_\theta(z)) + \mathbb{E}_{q_\phi(z|X)}[-\log p_\theta(X|z)] \quad (1)$$

where the reverse Kullback-Leibler (KL) divergence is to measure the distance between posterior distribution of $z$ ($q_\phi(z|X)$) parameterized by the encoder and the prior distribution of $z$ ($p_\theta(z)$) parametrized by the decoder. The second term of the right side of Equation (1) is the expected negative log-likelihood to measure the expected error of reconstructing the data points belonging to $X$ from the latent space $Z$. We aim to maximize the log-likelihood of $\log p_\theta(x) \geq \text{ELBO}$, where ELBO is the evidence lower bound. We let $\text{ELBO} = \mathcal{L}(\theta, \phi; X)$. The $\log p_\theta(x)$ will be maximized when the negative ELBO is minimized. Given the GPU support, we choose to compute the analytic KL divergence and do not use the reparameterization trick. In addition, the weight ($\omega$) of the KL divergence term is a crucial hyperparameter for VAE performance. A too small $\omega$ cannot effectively regularizing the $q_\phi(z|X)$ term so the $z$ sampled from $q_\phi(z)$ will be from a very low-density position of $q_\phi(z|X)$. On the contrary, when $\omega$ is too large, the distance between the posterior distribution and prior distribution is too close, resulting in the loss of diversity. In our study, we set the $\omega=0.01$ as the KL divergence weight or VAE optimization. For image generation, we use 2D convolutional layers to down-sampling (stride=2) the feature maps for the VAE encoder and use 2D transpose convolutional layers to up-sampling (stride=2) the latent variables back to images (64-by-64-by-3) for the VAE decoder.

The new cycle-consistent adversarial network with dynamic criterion, or Cycle GAN with dynamic criterion we present in this work, is used to generate high quality synthetic malaria infected blood cell images from real images with human eye detectable randomly diversity from the original ones. Cycle GAN is the state-of-the-art conditional generative adversarial network (cGAN) for unpaired image to image translation8. A typical Cycle GAN uses two generators and two discriminators to learn the mapping of two distributions by optimizing with a complex objective and reaching a state of adversarial equilibrium. In our new model, we add a pretrained binary classifier as the criterion, which is a residual neural network trained by the training dataset with the accuracy of 99.61%. The criterion is to calculate an extra critic loss term for optimizing both generators during the Cycle GAN training. The term “dynamic” means if the GAN model generates images too far away from the real malaria positive blood cells, the criterion will yield large penalty to the critic loss term to pull the generator back to the acceptable scope. On the other hand, if the generated images are in the acceptable scope for real malaria blood cells, the critic loss term will be dynamically minimize to keep the diversity of the synthetic images.
The original Cycle GAN architecture has two GAN models to learn and generate images respectively belonging to the source domain X and target domain Y. X represents the distribution of the normal cell images and Y represents the distribution of the malaria infected cell images. There are two pairs of GAN models in the Cycle GAN: generator G and discriminator D aim to adversarially generate and distinguish the generated / real malaria infected blood cell images, i.e., \( \min G, \max D \) \( L_{GAN}(G, D, X, Y) \), and generator F and discriminator D aim to adversarially generate and distinguish the generated and real normal blood images, i.e. \( \min F, \max D \) \( L_{GAN}(F, D, Y, X) \). The summation of the two terms: \( L_{GAN}(G, D, Y, X) + L_{GAN}(F, D, X, Y) \), is the adversarial loss term of the Cycle GAN. Two more loss terms are added to form the original total generator loss: cycle consistency loss, or \( L_{cycle}(G, F) \), is to measure the error when the images are reversely translated back to their original domains, i.e., \( x \rightarrow G(x) \rightarrow F(G(x)) \approx x \) (forward cycle consistency), and \( y \rightarrow F(y) \rightarrow G(F(y)) \approx y \) (backward cycle consistency). It helps to transfer uncommon style elements such as the dots representing the parasitizing plasmodia in the infected blood cells and random dyed organelles in normal cell, while it remains the comment features such as the shape of the blood cell during image translation. The identity loss, or \( L_{iden}(G, F) \) is added to compute the total objective loss for the whole model. It is to measure whether the generators \( (G \text{ and } F) \) can produce a real image from a real image, i.e., \( x \rightarrow F(x) \approx x \), and \( y \rightarrow G(y) \approx y \). This adjustment can raise the magnitude the gradient to further stabilize the adversarial train, and it also helps to enhance the background diversity of the generated images. The total generator loss is written as:

\[
L_{total} = [L_{GAN}(G, D, Y, X) + L_{GAN}(F, D, X, Y)] + \lambda L_{cycle}(G, F) + \lambda L_{iden}(G, F)
\]

\( \lambda \) is the weight for the cycle consistency loss and identity loss during optimization. According to original work, the \( \lambda \) term is set to 10.0 for Cycle GAN optimization\(^8\). However, we find that \( \lambda=10.0 \) is too low for the model optimization, thus empirically \( \lambda=80.0 \) is used in our Cycle GAN implementation. Our enhanced Cycle GAN model introduces a new criterion loss term. It consists of two terms, i.e., the cycle criterion loss and the identity criterion loss, which can be jointly written as:

\[
L_{critics} = L_{c-cycle} + L_{c-identity}
\]

Like the cycle loss and identity loss, the cycle criterion loss \( L_{c-cycle} \) quantitatively measures whether the back-translated images are still be classified as the original class, and the identity criterion loss \( L_{c-identity} \) quantitatively measures whether the trained generators can produce real images from a real observed sample that still consistent to the same class. After adding the new criterion loss term, the total generator loss in Equation 2 is revised as:

\[
L_{total} = [L_{GAN}(G, D, Y, X) + L_{GAN}(F, D, X, Y)] + \lambda L_{cycle}(G, F) + \lambda L_{iden}(G, F) + \kappa (\varphi L_{critics})
\]

In Equation 4, \( \varphi \) and \( \kappa \) are the parameters to respectively adjust the importance of the different loss terms during the whole Cycle GAN architecture optimization. The classification loss injection is controlled by a function \( \kappa \) to determine the frequency of classification loss injection. We set \( \kappa \) to be once every five steps in this work because too frequent injection of classification loss will shift the adversarial equilibrium and reduce the fidelity of the synthetic images. The term \( \varphi \) determines the importance of the criterion loss when it is injected into the total generator loss. Empirically, the criterion loss contributes a large proportion of the total generator loss at the beginning of the Cycle GAN optimization. When the Cycle GAN training reaches an adversarial equilibrium, the criterion loss can periodically add an extra oscillation momentum to the stable condition to push the generator progress to learn more details. The new \( L_{critics} \) term is considered as a regularization method to prevent the saturated status of the GAN optimization because it provides a method to make the GAN training controllable to a certain degree. In summary, the total loss of the generators in our enhanced Cycle GAN consists of four parts:

- **Adversarial loss**: \( L_{GAN}(G, D, Y, X) + L_{GAN}(F, D, X, Y) \)
- **Cycle consistency loss**: \( L_{cycle}(G, F) \)
- **Identity loss**: \( L_{iden}(G, F) \)
- **Criterion Loss**: \( L_{critics} = L_{c-cycle} + L_{c-identity} \)

The generators follow the U-NET architecture with skip connections to reduce the input feature size from 64 by 64 to 1 by 1 then to restore to 64 by 64. The discriminators follow the PatchGAN architecture with an output of 8-by-8-by-1 feature map to determine with the images are real or fake. We use the binary cross entropy as the objective function for the discriminator loss and the adversarial loss terms for the generators. The cycle consistency loss and the identity loss use the mean of absolute error (MAE) function as the objective. The terms of the criterion loss are measured by
the sparse categorical cross entropy as the same method as how the pretrained criterion was optimized. Some studies recommend using the unbounded smooth loss function such as to optimize the GAN models such as Wasserstein loss or least square loss (MSE)\textsuperscript{8,9}. Empirically, the choice of loss functions is mainly based on the components of the total loss objective. If all errors can be measured within similar scales, using the unbounded loss functions is straightforward and easier for the overall GAN optimization. However, if the GAN architecture consists of many components like this case, using hyperparameters to adjust the importance of different terms or to determine the frequency of loss injection to the total loss can provide a more flexible option for GAN optimization as described in Equation 4. Our new Cycle GAN architecture is illustrated in Figure 1.

![Figure 1. Architecture of Cycle GAN with Dynamic Criterion](image)

**Experiments and Results**

We use an open-source dataset contains 24 thousand parasitemic (malaria positive) and normal (malaria negative) segmented blood cell images (ratio 1:1) hosted by National Library of Medicine (NLM) as we did our previous work\textsuperscript{3}. The dataset is accessible at ftp://lhcftp.nlm.nih.gov/Open-Access-Datasets/Malaria/NIH-NLM-ThinBloodSmearsPF/ for the development of an Android based automatic malaria screener\textsuperscript{11}. One benefit of using the Cycle GAN architecture is that the model can be optimized by a relatively small dataset (e.g., hundreds of images). To save the runtime, we randomly choose 19,578 images (9,789 from each class) for the Cycle GAN optimization and the rest 8,000 image for the following tests. Given the original image size, they are resized to 64-by-64-by-3 to fit the model input. And the models are respectively optimized by 140 epochs on the Google Colab Pro Cloud supported by a Tesla P100 GPU. The average optimization runtime of a single epoch is 16 seconds for the VAE model, and 55 seconds for the enhanced Cycle GAN model. A sample of the real blood cell images is illustrated in Figure 2.

![Figure 2. Original Blood Cell Images](image)
From Figure 2, we find that it is difficult to discriminate the uninfected blood cells (malaria negative) from the parasitic cells (malaria positive) without medical expertise because the images from both classes have similar background color and randomly dyed dots inside the cells. Since our hypothesis is that the malaria positive cell images and the malaria negative cell images belong to two separable distributions. Therefore, we can use the VAE to learn the distribution parameters of the malaria positive images, and we can also use the new Cycle GAN with dynamic criterion to learn the mapping parameters between the two domains.

We implement the VAE models as mentioned above and optimized them with different weights of KL divergence. We use the Adam (adaptive moment estimation) optimizer\(^3\) with the initial learning rate of \(2 \times 10^{-4}\). The VAE model are optimized with 100 epochs with the mini-batch size of 128 given the GPU memory limitation on Google Colab.

The results show that the VAE decoder can generate plausible blood cell images from the KL weight=0.10. The outcomes become better when the KL weight=0.05. The generated images are shown in Figure 3. Noted that the all the synthetic images by the VAE are all classified as malaria positive images by the pretrained criterion.

![Figure 3. Synthetic Blood Cell Images by VAE](image)

In the next experiment, we implement the new Cycle GAN with dynamic criterion with \(\lambda = 40.0\) and \(\varphi =0.20\). The models are optimized with the Adam optimizer\(^4\) with the initial learning rate of \(4 \times 10^{-4}\) for 150 epochs. The mini-batch size is set to 128 given the limitation of GPU memory on Google Colab. The synthetic blood cell images respectively generated from the malaria positive images and from the negative images are shown in Figure 4.

From the results, we conclude that the trained generator of the enhanced Cycle GAN can synthesize malaria positive from both positive and negative real images. All the generated images are accurately classified by the pretrained criterion (accuracy=100%). However, from Figure 4, we find that thought all generated images are classified as malaria positive by the criterion, the synthetic images generated from positive images are obviously more plausible than those generated from negative images. In additional, artifacts are easily observed from all generated images, but they are effectively restricted to a certain degree by the Cycle GAN optimization. In contrast, if we remove the criterion and optimize the original Cycle GAN architecture, the generated images will have less control resulting in about 23.3% of the generated images will be classified as malaria negative by the pretrained criterion. (See Figure 5)
When we observe the changes of different loss term values during the optimization of the enhanced GAN with dynamic criterion (Figure 6), we find that the whole architecture will reach and maintain an adversarial equilibrium after about 5 epochs. As a result, the learning process of the generators will be slowed down due to the stable low gradient. A periodical injection of the criterion loss term can give extra oscillation momentum to slightly break the
adversarial equilibrium, so that the learning process can be expedited by this dynamic loss injection. The proper choice of the weight and frequency of the criterion loss value is important. From the bottom graphs of Figure 6, we find that the values of the two criterion loss terms are very large at the beginning (particularly during the first 40 epochs), therefore, they will pull the whole architecture away from the approaching adversarial equilibrium. Conversely, the value of the two criterion loss terms become much smaller in the end of the training (after the 80 epochs), so they will improve the learning process when the gradients become saturated.

Figure 6. Change of loss values during enhanced Cycle GAN Optimization

Finally, we use the Frechet Inception Distance (FID) to quantitatively measure the generated images by different generative models. FID is a metric to evaluate the distance between the feature vectors by an inception network trained by the ImageNet dataset\textsuperscript{13}. The FID score provides quantitative evaluation for the quality of generated images by generative models. In general, lower FID scores indicates the generated images are well correlated to the real ones and with higher quality images. However, since the images in the ImageNet dataset and the histological images are obviously heterogeneous, the FID scores for the generated blood cell images are likely to be very low because the pretrained inception net is likely to classify all generated cell images to the same class. However, the quantitative difference between the generated images by different models still can provide an objective metrics for comparison. The FID scores of the synthetic images by different generative models are shown in Table 1. It indicates that the new Cycle GAN with dynamic can generate the best quality images compared to those by the VAE model.

Table 1. Comparison of image quality and performance of enhanced Cycle GAN and VAE

<table>
<thead>
<tr>
<th>Model architecture</th>
<th>Classification Accuracy</th>
<th>Training Runtime (sec / epoch)</th>
<th>FID Mean (Std)</th>
<th>Image Quality (Subjective)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cycle GAN with dynamic criterion</td>
<td>100%</td>
<td>55</td>
<td>0.0043 (0.0005)</td>
<td>High with good diversity</td>
</tr>
<tr>
<td>Cycle GAN without criterion</td>
<td>76.6%</td>
<td>52</td>
<td>0.0051 (0.0003)</td>
<td>High with good diversity</td>
</tr>
<tr>
<td>Convolutional VAE</td>
<td>100%</td>
<td>16</td>
<td>0.0085 (0.0007)</td>
<td>Fair with good diversity</td>
</tr>
</tbody>
</table>
Conclusion

This study demonstrates a new Cycle GAN with dynamic criterion can generate high quality synthetic blood cell images from real segmented histological blood cell images. Compared with the convolutional variant autoencoder (VAE), another trendy deep generative model, the enhanced Cycle GAN produces synthetic blood cell images with higher quality and good diversity. The criterion provides a dynamic control to the GAN architecture to generate images belonging to the desirable class with complex discriminative pattern associated with medical expertise. We believe this new method is a state-of-the-art solution to improve the balance of the training dataset and further the final performance of the other DNN based machine learning tasks. Therefore, it is helpfully to solve the common machine learning issue on the inaccessibility of well-annotated medical images relying on medical expertise, and it will finally become a low-cost and feasible method to improve the AI performance in the medical imaging domain.

Acknowledgment

This work is supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the York Research Chair (YRC) program.

References

MetBERT: a generalizable and pre-trained deep learning model for the prediction of metastatic cancer from clinical notes

Ke Liu, Ph.D.\textsuperscript{1, 2, #}, Omkar Kulkarni, M.S.\textsuperscript{3, #}, Martin Witteveen-Lane M.S.\textsuperscript{3}, Bin Chen, Ph.D.\textsuperscript{1, 2, *}, Dave Chesla, MHA.\textsuperscript{3, 4, *}

\textsuperscript{1}Department of Pediatrics and Human Development, College of Human Medicine, Michigan State University, Grand Rapids, MI, USA; \textsuperscript{2}Department of Pharmacology and Toxicology, College of Human Medicine, Michigan State University, Grand Rapids, MI, USA; \textsuperscript{3}Office of Research, Spectrum Health, Grand Rapids, MI, USA; \textsuperscript{4}Department of Obstetrics, Gynecology and Reproductive Biology, College of Human Medicine, Michigan State University, Grand Rapids, MI, USA

#: Equal contribution; *correspondences: Bin Chen (chenbi12@msu.edu), Dave Chesla (david.chesla@spectrumhealth.org)

Abstract

Distant metastasis is the major cause of cancer-related deaths; however, early diagnosis of cancer metastasis remains a significant challenge. The recent advances in pre-trained natural language processing models coupled with the accumulation of publicly available Electronic Health Records (EHR) data provide an unprecedented opportunity to computationally tackle the challenge. Here, we fine-tuned multiple state-of-the-art BERT-based models using discharge summaries from the open MIMIC-III dataset and derived MetBERT, a novel model tailored to predict cancer metastasis from clinical notes. MetBERT achieved high performance (AUC=0.94) on our in-house validation dataset, suggesting its high generalizability. In addition, MetBERT enabled determining the date of cancer metastasis using the rich information in clinical notes and therefore could be potentially deployed as a tool for early diagnosis. Finally, we interpreted MetBERT at different scales and revealed a possible association between radiation therapy and metastasis risk in multiple cancer types.

Introduction

Cancer metastasis is the primary cause of cancer mortality; however, early diagnosis of cancer metastasis remains a big challenge, partially because micrometastatic niches are difficult to detect by conventional diagnostic approaches (such as CT-scan)\textsuperscript{1}. Since the onset of metastatic cancer is associated with multiple clinical symptoms, computational patient phenotyping using unstructured data such as charge summaries and progress notes in Electronic Health Record (EHR) systems likely assists early diagnosis of cancer metastasis.

The recent innovations in deep neural network-based methods have disrupted many fields including natural language processing\textsuperscript{2, 3, 4, 5}. However, training a language model using deep neural networks from scratch requires a huge amount of labeled data and computing resources, which are usually lacking in small institutes. Utilizing pre-trained model weights followed by fine-tuning for downstream domain-specific tasks has therefore become a popular approach as it often results in better performance with lower training overhead on smaller datasets. With the advances in transfer
learning, more advanced deep neural network based approaches (such as transformer-based representations which rely on attention mechanism \(^6\)), and pretrained contextualized language models (such as Bidirectional Encoder Representation from Transformers (BERT) \(^7\)), have achieved state-of-the-art performance in multiple types of tasks (e.g., text classification, question answering and named entity recognition). With the success of BERT based models in general domain, clinical researchers have developed models such as BioBERT (pre-trained using data from PubMed abstracts and PMC full-text articles), clinicalBERT (pre-trained with MIMIC-III data), BlueBERT (pre-trained with the combination of MIMIC-III and PubMed data), and PubmedBERT (pre-trained using domain-specific pretraining from scratch on PubMed data) \(^8,9,10,11\). These models have shown promising performance on clinical domain tasks. In addition, researchers also reported that language models pre-trained on biomedical text performed better on downstream biomedical domain tasks than those trained on general domain text corpora such as Wikipedia data \(^9,11\).

Here, we aim to address the clinical challenge in cancer metastasis diagnosis using recent transfer-learning techniques and frameworks (Figure 1). We started by utilizing the publicly available MIMIC-III dataset to fine-tune five different BERT-based models and then evaluated the performance of the best fine-tuned model (which we call MetBERT) in an independent dataset prepared from the Epic system at Spectrum Health. In addition, we showed an example to demonstrate that MetBERT has an enormous potential to be applied for the early diagnosis of metastatic cancer. Finally, we interpreted MetBERT with gradient-based methods.

**Figure 1.** Overview of the study design. The green and purple parts denote the fine-tuning and validation phase, respectively.
Methods

Our training data was compiled from MIMIC-III, a publicly available (and de-identified) EHR dataset consisting of clinical information of over 40,000 patients. All the 1,610 expert-annotated discharge summaries used for model fine-tuning came from previous work. Among them, 178 were labeled with “Advanced Cancer” and considered as positive samples (class 1).

Our in-house testing data was pulled out form the Epic system of Spectrum Health. In total, 5024 discharge summaries (corresponding to 1,478 Cancer Registry patients with metastatic cancer) were used for validating MetBERT.

Before fine-tuning, we took multiple steps to pre-process discharge summaries. For example, de-identified fields (encapsulated by [***data***]), punctuations, information leading up to chief complaint, and numbers were removed; in addition, acronyms for medications prescribed were replaced with full names (e.g. 'b.i.d.' with ‘twice a day’, ‘q.i.d.’ with ‘four times a day’, 'p.r.n.' with ‘as needed’).

Pre-trained weights of all the fine-tuned BERT models were provided by the HuggingFace library. For each model, we added a dropout layer for regularization and a fully connected classification layer (which outputs scores) for fine-tuning. The softmax and argmax functions were used for mapping scores to probabilities and class assignment, respectively. The cross-entropy loss function was used to calculate the total loss. We fine-tuned each model for 4 epochs, the batch size and learning rate (Adam optimizer) were set to 32 and 3e-5, respectively.

MetBERT is implemented in Python3 and freely available at https://github.com/Bin-Chen-Lab/MetBERT.

Results

Fine-tuning of BERT models

We fine-tuned five types of BERT-based models (BERT, BlueBERT, BioBERT, ClinicalBERT, and PubmedBERT) and assessed their performances by measuring precision, recall, and F-1 scores of positive samples in the training dataset. As shown in Table 1, the PubmedBERT model performed the best while baseline BERT came a close second. The results are not too surprising given that PubmedBERT used specific pre-training methods as well as more representative in-domain vocabularies, boosting its performance in document classification. We call the fine-tuned PubMedBERT as MetBERT (Metastatic cancer related BERT) and all downstream analyses are based on this model.

Table 1. Performance of different BERT models on a training set.

<table>
<thead>
<tr>
<th>Name</th>
<th>Precision</th>
<th>Recall</th>
<th>F-1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT (Baseline)</td>
<td>0.700</td>
<td>0.777</td>
<td>0.736</td>
</tr>
<tr>
<td>BlueBERT</td>
<td>0.722</td>
<td>0.722</td>
<td>0.722</td>
</tr>
</tbody>
</table>
Validating MetBERT on an independent dataset

We validated the performance of MetBERT on our in-house dataset containing 5,024 clinical notes (corresponding to 1,478 Cancer Registry patients with metastatic cancer). We found the distribution of MetBERT probability values was rather bimodal (Figure 2a), with one peak centered around 0 and the other one around 0.75; however, when applying MetBERT to clinical notes of random patients (sampled from the Epic system of Spectrum Health) the derived probability values were highly concentrated around zero (Figure 2b).

Since there are patients with more than one clinical note, we further computed a probability for each patient by taking the maximum MetBERT probability of the notes corresponding to the patient. Wilcoxon rank-test suggested that the patient-level probability distributions were significantly different between random and metastatic cancer patients (Figure 2c). When using the patient-level probability to predict metastatic cancer patients, the AUC (area under the curve) achieved 0.94 (Figure 2d).

Determine the date of cancer metastasis with MetBERT

Knowing the date of cancer metastasis is important for physicians to make treatment plans. In our case, although Cancer Registry provides the date of diagnosis with metastasis, we show that a more accurate estimation could be recovered using MetBERT. For example, we identified a patient who had seven clinical notes of high MetBERT probability (Figure 3); surprisingly, the earliest note was generated 641 days ago before day 0 (date of cancer metastasis according to Cancer Registry), suggesting the patient might have been with metastasis for at least 1.75 years before his (or her) tumor samples was analyzed by Cancer Registry, or the date was mislabeled during the curation. Therefore, the timestamp of the earliest high-probability note should be a better surrogate of the date of cancer metastasis.
Figure 2. Validation of MetBERT. (a) Histogram of MetBERT probabilities of discharge summaries of metastatic cancer patients. (b) Histogram of MetBERT probabilities of discharge summaries of random patients. (c) Boxplot of patient-level probability values. The p-value is computed with the Wilcoxon rank-test. (d) ROC curve to show the performance of predicting metastatic cancer patients with patient-level probability.
Figure 3. Determine the date of cancer metastasis with MetBERT. Each dot represents a discharge summary, the x-axis represents the time in days (zero mean the date of diagnosis of cancer metastasis recorded in Cancer Registry), and the y-axis represents MetBERT probability. The solid line represents Y=0.8.

Interpretation of MetBERT

It has been shown that gradient-based explanations gave relatively better model interpretation when using fine-tuned model architectures. To get a sense of how MetBERT internally makes the prediction, we utilized CAPTUM library to interpret it. Figure 4a showed the interpretation result of a random positive sample from the MIMIC-III test set. Interestingly, we found “gemcitabine” and “oxaliplatin” were high attribution tokens of the model. In clinics, gemcitabine and oxaliplatin are widely used in treating metastatic cancers; therefore, the presence of their names in a discharge summary may imply that the corresponding patient has metastatic cancer.

In addition to interpreting MetBERT with a single sample, we generated model interpretations with multiple samples. In detail, we picked out ten Spectrum Health discharge summaries with the highest MetBERT probability and pooled the top 30 high-attribution tokens of each summary together; then, the “token pool” was visualized with a word cloud (Figure 4b). It is reasonable to observe that words such as “cancer”, “oncology”, “metastatic”, “metastasis” were with high frequency. Interestingly, the token “radiation” was also present in the word cloud. Since we did not purposely pick out the patients of a specific cancer to train our model, such observation may suggest a pan-cancer association between radiation therapy and higher risk of distance metastasis.

Figure 4. Interpreting MetBERT with gradient-based methods. (a) Visualization of interpretation results for a MIMIC-III discharge summary with probability larger than 0.8. Highlighted in green are high attribution tokens from the model’s perspective. (b) The word cloud visualizing the interpretation results based on multiple samples.
Conclusion

In this paper, we utilized transfer-learning technology to train a model (MetBERT) which predicts cancer metastasis patients based on EHR clinical notes and then validated its performance on an independent dataset. Although MetBERT was fine-tuned using MIMIC III dataset, the high performance (AUC = 0.94) on Spectrum Health dataset suggested that it is generalizable to other datasets, and such feature enables a potential usage of MetBERT for early diagnosis of metastatic cancer. Training a model using open datasets and deploying to an internal system is very appealing in many small institutes including ours where the initial labeled data is often scarce. This work has also provided helpful guidance for training BERT-based models that are used to predict other phenotypes (e.g., advanced heart disease) or even more specific labels (e.g., tumor stage).

One limitation of this study is that highly imbalanced label distribution is present within the training data. We tried using weighted cost function and weighted sampler for balancing the labels but did not see significant improvement in results. In the future, as more and more EHR datasets are being released, this issue should be mitigated. In the model-tuning process, high precision and low recall rate were observed for some of the models, which may be due to one of the drawbacks of BERT-based models, that is, a maximum of 512 tokens are allowed for training/fine-tuning due to the high computational cost of analyzing longer text instances. Oftentimes discharge summaries consist of long paragraphs of clinical text where important information might be enclosed in start, mid or even last part of discharge summary. As we are only selecting the first 512 tokens for fine-tuning, the model fails to capture nuances of phonemic traits related to advanced cancer that could be present in the tail end of discharge summary, resulting in a lower recall rate. There are some instances where we found the first few sentences of discharge summary including information relevant to social history and admission details but not relevant to cancer phenotyping rendering the model to misclassify the discharge summary as negative while in fact the actual label was positive. Employing advanced truncation methods such as head truncation (selecting only first 512 tokens), tail truncation (selecting last 512 tokens) and mixed selection (head + tail) might give us better results.

Acknowledgement

The research is supported by the MSU-Spectrum Health Alliance fund. The content is solely the responsibility of the authors and does not necessarily represent the official views of sponsors.

Competing interests

The authors declare no competing interests.

Contributions

K.L. and O.K. conceived the project. K.L., O.K., and M.W. performed the data analyses. K.L. and O.K. wrote the manuscript, with all authors contributing to writing and providing the feedback. B.C. and D.C. supervised the study. All the authors read and approved the final manuscript.
References


Extracting Radiological Findings With Normalized Anatomical Information Using a Span-Based BERT Relation Extraction Model

Kevin Lybarger, PhD, Aashka Damani, Martin Gunn, M.B., Ch.B., Özlem Uzuner, PhD, Meliha Yetisgen, PhD

University of Washington, Seattle, WA, USA; George Mason University, Fairfax, VA, USA

Abstract
Medical imaging is critical to the diagnosis and treatment of numerous medical problems, including many forms of cancer. Medical imaging reports distill the findings and observations of radiologists, creating an unstructured textual representation of unstructured medical images. Large-scale use of this text-encoded information requires converting the unstructured text to a structured, semantic representation. We explore the extraction and normalization of anatomical information in radiology reports that is associated with radiological findings. We investigate this extraction and normalization task using a span-based relation extraction model that jointly extracts entities and relations using BERT. This work examines the factors that influence extraction and normalization performance, including the body part/organ system, frequency of occurrence, span length, and span diversity. It discusses approaches for improving performance and creating high-quality semantic representations of radiological phenomena.

Introduction
Radiology reports contain detailed descriptions of diverse clinical abnormalities based on radiologists’ interpretation of medical imaging. Although structured reports with semantic representations of medical concepts have been developed, nearly all radiology reports convey findings through unstructured text. Semantic representations of radiological findings could be automatically generated using natural language processing (NLP) information extraction techniques. These automatically derived semantic representations would enable a wide range of applications, including ground-truth labeling for artificial intelligence applications of medical images, translation of reports into lay-language for patients, integration with clinical decision support, cross-specialty diagnosis correlation, automated impression generation, semantic searching of reports, and timely follow-up of recommendations. We are currently conducting a large-scale clinical and economic analysis of incidental findings (incidentalomas) in radiology reports, focusing on six organ systems with the highest probability of incidental malignancy (thyroid, lung, adrenal glands, kidneys, liver, and pancreas). Incidentaloma identification requires the extraction of radiological findings and conversion of these findings to a structured semantic representation.

To develop data-driven extraction models, we designed an event-based annotation schema and annotated computed tomography (CT) reports. Each finding event is characterized by a trigger and set of attributes (assertion, anatomy, characteristics, size, size-trend, size count). In this paper, we use this gold standard corpus to explore the extraction of radiological findings with normalized anatomy information. We extract radiological findings and associated anatomies as a relation extraction task, where the extracted anatomies are normalized to a set of 56 pre-defined anatomy labels. We investigate this relation extraction task using Eberts and Ulges’s Span-based Entity and Relation Transformer (SpERT). SpERT is a state-of-the-art BERT model that jointly extracts entities and relations using span and relation output layers. To better understand the anatomy normalization task and the role of context, we use the gold standard anatomy spans to explore anatomy normalization, without extraction. In this normalization experimentation, anatomy phrases are normalized at 0.89 F1 micro. In the extraction experimentation, finding spans are extracted at 0.83-0.92 F1, anatomy spans are extracted at 0.72-0.79 F1, and finding-anatomy relations are extracted at 0.63-0.72 F1. We explore the relationship between extraction performance, span length and diversity, and anatomy frequency. This work leverages state-of-the-art transformer-based extraction approaches and provides insight into the extraction of key finding and anatomy information from radiology reports.

Related Work
There is a large body of biomedical entity normalization work exploring the mapping of text spans to fixed vocabularies. A frequently explored ontology is the Unified Medical Language System (UMLS), which includes the
Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) and RxNorm. The 2019 National NLP Clinical Challenges (n2c2)/Open Health NLP (OHNLP) shared task explored the normalization of pre-defined text spans in clinical notes to SNOMED CT and RxNorm concepts. Top performing teams used dictionary and string matching, cosine distance, retrieval and ranking, and deep learning, with the highest performing system utilizing deep learning.\(^\text{11}\)

With large concept vocabularies, a frequently explored approach utilizes a two-step process, where a retrieval model identifies top candidate concepts and then a reranking model identifies the single best concept.\(^\text{12–15}\) Chen et al. normalized biomedical entities to SNOMED CT concepts, using knowledge sources to identify candidate concepts and an ensemble of machine learning approaches is used to identify target concepts.\(^\text{12}\) Datta et al. and Ji et al. explore biomedical entity normalization tasks using BM25 to identify top concept candidates and BERT to select the top concept.\(^\text{13, 14}\) In the n2c2 challenge, Xu et al. uses a Lucene-based search that utilizes the UMLS and a BERT-based reranker.\(^\text{15}\) In our exploration of anatomy normalization, the anatomy label vocabulary is relatively small (56 labels) and does not necessitate a retrieval step for identifying top label candidates; however, we do utilize BERT-based models for extracting anatomy phrases and mapping the phrases to a vocabulary of anatomy concepts. Tutubalina et al. investigate the normalization of medical concepts in social media posts to SNOMED CT concepts using a bidirectional recurrent neural network (RNN) and attention network to classify spans, incorporating semantic information from the UMLS.\(^\text{16}\) Wang et al. explore a hierarchical anatomy normalization task with nine body parts (e.g. head and chest) and 41 sub-body parts (e.g. skull and brain).\(^\text{17}\) Wang et al. use Wikipedia as an anatomical knowledge source and explore different scoring functions for comparing anatomical entities to anatomical wiki pages.

Recent work also explores both the extraction and normalization of biomedical entities, including anatomical spans. Tahmasebi et al. implement an unsupervised approach where anatomical phrases are identified using SNOMED CT and grammar-based patterns.\(^\text{18}\) Anatomical phrases are normalized by representing each phrase as the weighted sum of word embeddings and comparing the cosine similarity between anatomical phrases and target concept labels. This unsupervised approach outperforms a stacked bidirectional RNN and conditional random fields (CRF) model. Tahmasebi et al. identify 56 anatomical class labels corresponding to SNOMED CT IDs, which we use in this work. In a sequence tagging task, Zhu et al. predict eight anatomy classes (brain, breast, kidney, liver, lung, prostate, thyroid, and other) using a stacked bidirectional long short-term memory network (bi-LSTM) and CRF that incorporates sentence-level context vectors that are learned to predict the presence of each anatomical class in the sentence.\(^\text{19}\) Zhu et al. experiments with incorporating sentence-level and report-level context and finds that incorporating report-level context improves classification performance. Similar to Zhu et al., we also explore the role of context in normalizing anatomical spans. Our work is differentiated from this prior work in that we extract anatomical information related to medical findings, and the anatomical phrases are normalized to a larger anatomy vocabulary.

Methods

Data

This work utilizes an annotated data set created by Lau, et al.\(^\text{20}\) which includes 500 CT reports from an existing clinical data set. The 500 annotated reports were randomly selected from a pool of 706,908 CT reports authored from 2008-2018 at the University of Washington Medical Center and Harborview Medical Center. The annotated reports use an event-based annotation scheme to characterize two types of findings: (1) lesion findings (e.g. mass or tumor) and (2) other medical problem findings (e.g. fracture or lymphadenopathy). Each event includes a trigger that identifies the lesion or medical problem finding. The trigger connects to arguments that characterize the finding across multiple dimensions, including assertion (e.g. present vs. absent), anatomy,
count, size, and other attributes. In the initial rounds of annotation, Lau, et al. doubly annotated 30 reports to assess inter-rater agreement. The inter-rater agreement for the event annotations is 0.83 F1.

Although the corpus is annotated with several attributes related to findings, including lesions, this work focuses on the extraction of findings and the associated anatomical information. We collectively refer to the lesion findings and other medical problem findings as *Finding*. Lau, et al.’s annotated corpus includes *Anatomy* annotations without anatomy normalization labels. We augment the *Anatomy* annotations to include the *Anatomy Subtype* labels defined in Table 1. These *Anatomy Subtype* labels are based on Tahmasebi et al.’s work identifying anatomical terms using unsupervised learning.\(^{18}\) The terms have associated SNOMED CT concept identifiers and represent all human organ systems, anatomic labels, and body regions. The *Anatomy Subtype* labels normalize the *Anatomy* spans, allowing the extracted finding and anatomy information to be more readily used in secondary use analyses.

<table>
<thead>
<tr>
<th>Anatomy Subtype labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abdomen</td>
</tr>
<tr>
<td>Gallbladder</td>
</tr>
<tr>
<td>Nasal sinus</td>
</tr>
<tr>
<td>Seminal vesicle</td>
</tr>
<tr>
<td>Adrenal gland</td>
</tr>
<tr>
<td>Head</td>
</tr>
<tr>
<td>Neck</td>
</tr>
<tr>
<td>Spleen</td>
</tr>
<tr>
<td>Back</td>
</tr>
<tr>
<td>Heart</td>
</tr>
<tr>
<td>Nervous*</td>
</tr>
<tr>
<td>Stomach</td>
</tr>
<tr>
<td>Bile Duct</td>
</tr>
<tr>
<td>Integumentary*</td>
</tr>
<tr>
<td>Nose</td>
</tr>
<tr>
<td>Testis</td>
</tr>
<tr>
<td>Bladder</td>
</tr>
<tr>
<td>Intestine</td>
</tr>
<tr>
<td>Ovary</td>
</tr>
<tr>
<td>Thorax</td>
</tr>
<tr>
<td>Brain</td>
</tr>
<tr>
<td>Kidney</td>
</tr>
<tr>
<td>Pancreas</td>
</tr>
<tr>
<td>Thyroid</td>
</tr>
<tr>
<td>Breast</td>
</tr>
<tr>
<td>Laryngeal</td>
</tr>
<tr>
<td>Pelvis</td>
</tr>
<tr>
<td>Trach.</td>
</tr>
<tr>
<td>Cardio*</td>
</tr>
<tr>
<td>Liver</td>
</tr>
<tr>
<td>Penis</td>
</tr>
<tr>
<td>Upper limb</td>
</tr>
<tr>
<td>Diaphragm</td>
</tr>
<tr>
<td>Lower limb</td>
</tr>
<tr>
<td>Pericardial sac</td>
</tr>
<tr>
<td>Urethra</td>
</tr>
<tr>
<td>Digestive*</td>
</tr>
<tr>
<td>Lung</td>
</tr>
<tr>
<td>Peritoneal sac</td>
</tr>
<tr>
<td>Uterus</td>
</tr>
<tr>
<td>Ear</td>
</tr>
<tr>
<td>Lymphatic*</td>
</tr>
<tr>
<td>Pharynx</td>
</tr>
<tr>
<td>Vagina</td>
</tr>
<tr>
<td>Esophagus</td>
</tr>
<tr>
<td>Mediastinum</td>
</tr>
<tr>
<td>Pleural sac</td>
</tr>
<tr>
<td>Vas deferens</td>
</tr>
<tr>
<td>Eye</td>
</tr>
<tr>
<td>Mouth</td>
</tr>
<tr>
<td>Prostate</td>
</tr>
<tr>
<td>Vulva</td>
</tr>
<tr>
<td>Fallopian tube</td>
</tr>
<tr>
<td>MSK*</td>
</tr>
<tr>
<td>Retropertoneal</td>
</tr>
<tr>
<td>Whole body</td>
</tr>
</tbody>
</table>

Table 1: *Anatomy Subtype* labels. Abbreviated terms include Cardiovascular (Cardio), Musculoskeletal (MSK), and Tracheobronchial (Trach). * indicates systems, like the Nervous System.

We approach this radiological information extraction task as a relation extraction task, where spans are identified, mapped to a fixed set of classes, and linked through relations. Figure 1 presents example annotations. The entity types include *Finding* and *Anatomy*, although the phrases are not strictly noun phrases. Unlike a typical entity annotation, the *Anatomy* entities include *Anatomy Subtype* labels corresponding to the 56 anatomies defined in Table 1. We represent the *Finding-Anatomy* pairs as asymmetric relations, where the relation head is a *Finding* entity and the tail is an *Anatomy* entity. There is only a single relation type, *has*, so the *Finding-Anatomy* pairing can be interpreted as a binary classification task (connected vs. not connected).

The annotated corpus includes 500 CT reports, with 10.4K *Finding* entities, 5.0K *Anatomy* entities, and 6.3K *Finding-Anatomy* relations.\(^{20}\) There are more *Finding-Anatomy* relations than *Anatomy* entities, because a given *Anatomy* entity can be associated with multiple findings. The corpus includes 19K sentences and 203K tokens and is randomly split into training (70%), validation (10%), and test (20%) sets. Figure 2 presents the 20 most frequently annotated *Anatomy Subtypes* in the training set. Musculoskeletal system (MSK), Cardiovascular system (Cardio), and Lung are the most frequent *Anatomy Subtypes*, and there are many subtypes that occur infrequently or are absent from the data set. This skewed distribution is the result of randomly sampling the 500 CT reports from the larger data set.

**Information Extraction**

We extract the radiological findings and related anatomy using Eberts and Ulges’s SpERT model.\(^{9}\) SpERT jointly extracts entities and relations using a pre-trained BERT\(^{21}\) model with output layers that classify spans and predict the relations between spans. SpERT achieves state-of-the-art performance in three entity and relation extraction tasks, including open domain information extraction (CoNLL04), science information extraction (SciERC), and adverse drug
event extraction (ADE). The SpERT framework is presented in Figure 3.

**Input encoding:** Each sentence is tokenized and converted to BERT word pieces. BERT generates a contextualized representation for the sentence, yielding a sequence of word-piece embeddings \( (e_{CLS}, e_1, e_2, \ldots, e_t, \ldots, e_n) \), where \( e_{CLS} \) is the sentence-level representation associated with the \([CLS]\) token, \( e_t \) is the \( t^{th} \) word piece embedding, and \( n \) is the sequence length.

**Span Classification:** The span classifier predicts labels for each span, \( s = (t, t+1, \ldots, t+k) \), where the width of the span is \( k+1 \) word pieces. A learned matrix of span width embeddings, \( w \), is used to incorporate a span width prior in the classification of spans and relations. A fixed length representation of the \( i^{th} \) span, \( e(s_i) \), is created by max pooling the associated BERT embeddings and looking up the relevant span width embedding, as

\[
e(s_i) = \text{MaxPool}(e_t, e_{t+1}, \ldots, e_{t+k}) \odot w_{k+1},
\]

where \( \odot \) denotes concatenation. The span classifier input for the \( i^{th} \) span, \( x^s_i \), is the concatenation of the span embedding, \( e(s_i) \), and sentence-level context embedding, \( e_{CLS} \), as

\[
x^s_i = e(s_i) \odot e_{CLS}.
\]

The span classifier consists of a single linear layer, as

\[
y^s_i = \text{softmax}(W^s \cdot x^s_i + b^s).
\]

For our task, the span classifier label set, \( \Phi^s \), includes a null label, Finding, and the 56 Anatomy Subtypes in Table 1: \( \Phi^s = \{\text{null}, \text{Finding}, \text{Abdomen}, \text{Adrenal gland}, \ldots, \text{Whole body}\} \) (\( |\Phi^s| = 58 \)). The null label indicates no span prediction. We experimented with several multi-layer, hierarchical span classifiers, where the first classification layer predicts the entity labels, \( \{\text{Finding}, \text{Anatomy}\} \), and the second layer resolves the 56 Anatomy Subtype labels for the Anatomy spans. However, none of the hierarchical span classifier configurations outperformed the base SpERT model, so these hierarchical configurations are not presented. By directly predicting the Anatomy Subtypes, the span classifier identifies and normalizes the Anatomy spans. Only spans with a width less than a predefined maximum are included in modeling to limit time and space complexity.

**Relation Classification:** The relation classifier predicts the relationship between a candidate head span, \( s_i \), and a candidate tail span, \( s_j \), with input

\[
x^r_{i,j} = e(s_i) \odot e(s_i,s_j) \odot e(s_j),
\]

\[342\]
where $e(s_i)$ and $e(s_j)$ are the head and tail span embeddings and $c(s_i, s_j)$ is the max pooling of the BERT embedding sequence between the head and tail spans. The relation classifier consists of a single linear layer, as

$$y^r_{i,j} = \text{softmax}(W^r \cdot x^r_{i,j} + b^r).$$

(5)

For our task, the relation classifier label set, $\Phi^r$, includes a null label and the relation types: $\Phi^r = \{\text{null, has}\}$ ($|\Phi^r| = 2$). Only spans predicted to have a non-null label are considered in the relation classification, to limit the time and space complexity of the pairwise span combinations.

**Training:** The span and relation classifier parameters are learned while fine-tuning BERT. For each training batch, the cross entropy loss for each classifier is averaged, and the averaged loss values are summed using uniform weighting. The training spans include all the gold spans, $S^g$, as positive examples and a fixed number of spans with label null as negative examples. The training relations include all the gold relations as positive in samples, and negative relation examples are created from all the entity pairs in $S^g$ that are not connected through a relation.

**Baseline:** As a baseline for evaluating the performance of SpERT, we implement a multi-step BERT approach (BERT-multi) where entities are first extracted and then relations between entities are resolved. BERT-multi is implemented by adding entity extraction and relation prediction layers to a single pretrained BERT model. For entity extraction, we implement a common BERT sequence tagging approach, where Begin-Inside-Outside (BIO) labels are predicted by a linear layer applied to the last BERT hidden state. For evaluation, the word piece predictions are aggregated to token-level predictions by taking label of the first word piece of the token. For relation prediction, we implement a common BERT sentence classification approach, where relation predictions are generated by a linear layer applied to the $[CLS]$ encoding. For each pair of predicted entities, a modified input sentence is created where the identified entities are replaced with special tags. For example, the first sentence in Figure 1 would become, “Lungs: @Finding$\$ of the @Lung$\$”. When enumerating candidate head-tail pairs, only Finding entities are included as potential heads and only Anatomy Subtype spans are included as potential tails. No such constraint is imposed in SpERT. Each training batch involves (i) generating sequence tag predictions and (ii) predicting relations for the identified spans, and the loss is backpropagated after both (i) and (ii).

**Experimentation:** The primary focus of this work is the extraction and normalization of anatomy information associated with findings. We use SpERT to extract Finding and Anatomy spans, normalize the Anatomy spans to the Anatomy Subtypes, and resolve Finding-Anatomy relations. The data set only includes anatomy annotations for anatomical information connected to findings, so not all anatomy phrases are annotated in the reports.

We include normalization-only experimentation, where the Anatomy Subtype labels are predicted for gold anatomy phrases. The normalization-only experimentation is incorporated to explore the difficulty of the anatomy normalization task separate from span extraction and investigate the role of context in anatomy normalization. This normalization-only experimentation uses the same input encoding (including width embedding) and span classifier as SpERT (see Equations 2-3). To investigate the role of context in anatomy normalization, we implement phrase-only models where the input is the anatomy phrase (e.g. “right lower lobe”) without any context and sentence context models where each anatomy phrase is contextualized in the sentence in which it is located (e.g. “Lungs: Compressive atelectasis of the the right lower lobe.”) Both normalization models use the gold labels to identify the anatomy phrases.

Model architectures and hyperparameters were selected by training models on the training set and evaluating performance on the validation set. Final performance was evaluated on the withheld test set. Common parameters across all models include: pretrained transformer=Bio+Clinical BERT, optimizer=Adam, maximum gradient norm=1.0, and learning rate=5e-5. Normalization parameters include: dropout=0.05, batch size=50, and epochs=15. SpERT parameters include: dropout=0.2, batch size=20, epochs=20, learning rate warmup=0.1, weight decay=0.01, negative entity count=100, negative relation count=100, max span width=10, and maximum span pairs=1000. BERT-multi parameters include: batch size=50, epochs=20, dropout=0.2, negative relation count=100, and maximum span pairs=1000. To account for the variance associated with model initialization, each model configuration was trained on the training set 10 times with the selected hyperparameters and evaluated on the test set to generate a distribution of performance values. The mean and standard deviation (SD) of the performance values is presented (mean±SD). Significance is assessed using a two-sided t-test with unequal variance.

**Evaluation:** Performance is assessed using precision (P), recall (R), and F-score (F1). Each entity, $z$, can be repre-
sented as a double, \( z = (s, \phi^s) \), where \( s \) is the span \( (t, t + k) \) and \( \phi^s \) is the span label in \( \Phi^s \). Entity extraction performance is assessed using two sets of equivalence criteria: exact match and any overlap. Under the exact match criteria, a gold entity, \( z \), is equivalent to a predicted entity, \( \hat{z} \), if the span and span label match exactly, as \( (s \equiv \hat{s}) \land (\phi^s \equiv \hat{\phi}^s) \). Under the more relaxed any overlap criteria, \( z \) is equivalent to \( \hat{z} \), if there is at least one overlapping token in the gold and predicted spans and the span labels match, as \( (s \text{ overlaps with } \hat{s}) \land (\phi^s \equiv \hat{\phi}^s) \). We include this any overlap assessment, because the Anatomy Subtype labels capture clinically relevant information, even if there are discrepancies in spans. In the example of Figure 1, the span “right lower lobe” is labeled as Anatomy with Anatomy Subtype Lung. If the span classifier predicts the span “lower lobe” to have the Anatomy Subtype label Lung, the gold and predicted spans would not match, and the sidedness information associated with “right” would not be captured. However, a majority of the clinically relevant information would be captured, namely that the Finding is associated with the Lung. Each relation, \( r \), can be represented as a triple, \( r = (z^h, \phi^r, z^t) \), where \( z^h \) is the head, \( \phi^r \) is the relation label in \( \Phi^r \), and \( z^t \) is the tail. A gold relation, \( \hat{r} \), and predicted relation, \( \hat{r} \), are equivalent if \( (z^h \equiv \hat{z}^h) \land (\phi^r \equiv \hat{\phi}^r) \land (z^t \equiv \hat{z}^t) \), where entity equivalence can be assessed using the exact match or any overlap criteria.

Results
Normalization

This section presents the normalization results where Anatomy Subtype labels are predicted for gold anatomy phrases. Table 2 presents the anatomy normalization performance on the withheld test set averaged across the 10 randomly instantiated models for each input configuration: phrase only and sentence context. The F1 scores in Table 2 are micro averaged across the 56 Anatomy Subtype labels. The phrase only model achieves relatively high performance, indicating a high proportion of the anatomical phrases include strong cues for normalization. The inclusion of the sentence context improves normalization performance from 0.86 F1 to 0.89 F1 with significance \( p < 0.05 \), indicating there are some ambiguous anatomy phrases that require intra-sentence context to resolve. For example, the term “cervical” can be related to the neck or the uterus, and sentence context is needed to resolve ambiguity. Early experimentation with context beyond the sentence of the anatomy phrase did not improve performance.

Table 3 presents the most frequently confused Anatomy Subtypes, averaged across the sentence context model predictions. We omit the full confusion matrix because of the high number of labels and sparsity of the matrix. In general, organs and body regions are the most confused anatomy subtypes as either could be applied. Cardio and MSK are among the most frequently confused labels, with 53% of all errors involving Cardio or MSK labels as the gold or predicted labels. Cardio and MSK labels are organ systems that extend throughout the body and therefore overlap with body region labels. Moreover, these labels are the most frequent in the data set. Other frequently confused labels include co-located body parts and organ systems, like Abdomen-Intestine and Head-Neck.

Entity and Relation Extraction

This section presents the entity and relation extraction performance. Tables 4a and 4b present the extraction performance on the withheld test set for SpERT and BERT-multi, averaged across 10 randomly instantiated models. Table 4a includes the span labeling performance for Finding and Anatomy entities and the micro-averaged Anatomy Subtype labels. An Anatomy label is assigned is any span with an Anatomy Subtype label. SpERT outperforms BERT-multi for
Table 4: Average extraction performance on the withheld test set, as mean and standard deviation for 10 trained models. †indicates best performance with significance ($p < 0.05$).

Table 4b presents the relation extraction performance. SpERT outperforms BERT-multi for Finding-Anatomy and Finding-Anatomy Subtype relations with significance. As expected, the relation extraction performance is lower than the span labeling performance because of cascading errors. For both architectures, the magnitude of the performance drop from span labeling to relation extraction is roughly consistent with the accumulation of Finding and Anatomy span labeling errors, suggesting that the performance of the relation classifiers is relatively high.

Figure 4 presents the recall of SpERT as a function of the gold span length, in number of tokens (not word pieces). The recall is aggregated for the 10 model runs and reported for Finding, Anatomy, and Anatomy Subtype labels. The maximum span width for SpERT is set to 10 tokens, so the exact match recall is zero for all spans longer than 10 tokens. Under the exact match criteria, the Finding recall drops from approxi-
mately 0.9 for shorter spans to approximately 0.2-0.3 for long spans (9-10 tokens). Under the any overlap criteria, the Finding recall remains relatively high for all span lengths, as the extractor only needs to identify a portion of the gold span for a match. Under the exact match criteria, the Anatomy and Anatomy Subtype remains relatively steady across span lengths from 1-10. Under the any overlap criteria, the Anatomy and Anatomy Subtype recall tends to increase with span length.

Figure 5 presents summary statistics and performance for the 15 most frequent Anatomy Subtypes in the test set. Figure 5 includes the label counts (# gold), number of unique lower cased spans (# unique). It also includes the normalization, span labeling, and relation extraction performance.

The normalization performance is associated with sentence context models summarized in Table 2, and the span labeling and relation extraction performance is associated with the SpERT model summarized in Table 4. There is a large imbalance in the distribution of Anatomy Subtype labels with Cardio, MSK, and Lung accounting for approximately 50% of the labels. The diversity of the anatomy spans varies significantly by Anatomy Subtype. For example, MSK has 168 unique spans in 204 occurrences (ratio of 0.8), while Mediastinum has 7 unique spans in 37 occurrences (ratio of 0.2). The span labeling and relation extraction performance does not drop off for infrequent labels and appears to be more related to the span diversity.

Error Analysis

Generating a correct relation prediction requires identifying the Finding (head), identifying the Anatomy and Anatomy Subtype (tail), and pairing the head and tail (role). The results in Table 4 suggests the biggest source of error is identifying Anatomy entities, followed by identifying Finding entities. Error is also introduced in the Anatomy Subtype normalization and Finding-Anatomy pairing; however, entity extraction is the most challenging aspect of this task. Table 5 presents example SpERT false negative spans for Finding and the most frequent Anatomy Subtypes. These false negatives are assessed using the any overlap criteria, to identify text regions related to findings and anatomy that the model completely missed.

The short Finding examples are relatively straightforward targets and the cause of these missed spans is unclear. The long Finding examples include medical problems coupled with anatomy, resulting in longer spans that are generally

<table>
<thead>
<tr>
<th>Span label</th>
<th>Short examples</th>
<th>Long examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finding</td>
<td>“hemia”</td>
<td>“poor opacification of these vessels distally”</td>
</tr>
<tr>
<td></td>
<td>“lesion”</td>
<td>“expanded thoracic aortic aneurysm”</td>
</tr>
<tr>
<td>Anatomy</td>
<td>“scapula”</td>
<td>“soft tissues of the posterolateral left chest wall”</td>
</tr>
<tr>
<td>MSK</td>
<td>“left third rib”</td>
<td>“subcutaneous fat in the right groin”</td>
</tr>
<tr>
<td>Anatomy</td>
<td>“aorta”</td>
<td>“proximal descending thoracic aorta”</td>
</tr>
<tr>
<td>Cardio</td>
<td>“coronary arteries”</td>
<td>“arteries of the right lower extremity and abdomen”</td>
</tr>
<tr>
<td>Anatomy</td>
<td>“left lung”</td>
<td>“lateral aspect of the right major fissure”</td>
</tr>
<tr>
<td>Lung</td>
<td>“right lower lobe”</td>
<td>“dependent portions of the left upper lobe adjacent”</td>
</tr>
</tbody>
</table>

Table 5: Example false-negative spans
more difficult to extract. The inclusion of anatomical information in the Finding spans creates annotation inconsistencies, where anatomical information may be labeled as Finding or Anatomy. We are building on this work as part of an exploration of incidentalomas and updated the annotation guidelines to separate finding and anatomy information, to create shorter, more consistently annotated spans. For example the Finding span, “expanded thoracic aortic aneurysm”, would be annotated as a the relation triple (Finding=“aneurysm”, role=“has”, Anatomy=“thoracic aortic”).

All of the short Anatomy examples are concise descriptions of anatomy that use common anatomical terminology. There are multiple contributing factors to these errors. In the corpus, only anatomy associated with findings is annotated, so many anatomy descriptions are not annotated. As previously discussed, anatomy information is frequently incorporated into Finding annotations, which introduces annotation inconsistencies. The long Anatomy examples are more nuanced descriptions of anatomy that often describe multiple systems or body parts in relation to each other. For example, the Cardio span, “arteries of the right lower extremity and abdomen”, contains references of three Anatomy Subtypes: Cardio, Lower limb, and Abdomen. Annotating such examples with the Anatomy Subtype labels can be challenging, and more nuanced anatomy descriptions are likely to have noisier annotations.

Conclusions

This work explores a novel radiological information extraction task with the goal of automatically generating semantic representations of radiological findings that capture anatomical information. We extract and normalize anatomical information connected to findings in CT reports, using state-of-the-art extraction architectures. This extraction task is both novel and important because it couples extracted anatomical information with radiological findings and normalizes the anatomical information to a commonly used ontology. Linking the anatomy to findings and normalizing the anatomy yields a more complete semantic representation, which can more easily be incorporated into secondary use applications. We demonstrate that the span-based SpERT model, which jointly extracts entities and relations, outperforms a strong BERT baseline that separately extracts entities and relations in a pipelined approach. The explored extraction task involves three subtasks: identifying Finding and Anatomy entities, normalizing Anatomy entities to Anatomy Subtypes, and pairing Finding and Anatomy entities through relations. Entity extraction is the most difficult of these subtasks. We find that extraction performance for Finding entities decreases as span length increases; however, Anatomy extraction performance is relatively constant across span lengths. In an exploration of performance by Anatomy Subtype, we find span extraction performance is influenced more by the diversity of the associated spans than the frequency of the Anatomy Subtype labels.

This work is limited by the annotated data set, which only utilizes data from a single hospital system and incorporates a single type of imaging report (CT). The extraction models trained on this annotated data set may not generalize well to other institutions or radiology modalities. We are currently expanding the annotated data set to other radiology modalities, including magnetic resonance imaging (MRI) and positron emission tomography (PET) reports.

The 56 Anatomy Subtypes used in this work provide moderate granularity in resolving the anatomical location of radiological findings. In our current incidentaloma research, we anticipate representing anatomical locations with finer resolution. We will build on the work presented here and explore learned approaches for characterizing anatomical spans through multiple attributes. For example the phrase “right lower lobe” could be characterized through a semantic representation describing the body part/organ (Lung), sidedness (right), and vertical location (lower). This type of detailed semantic representation could facilitate a wide range of impactful use cases.

Acknowledgements

This work was supported by NIH/NCI (1R01CA248422-01A1) and NIH/NLM (Biomedical and Health Informatics Training Program - T15LM007442). Research and results reported in this publication were partially facilitated by the generous contribution of computational resources from the University of Washington Department of Radiology.

References


Improving Pharmacovigilance Signal Detection from Clinical Notes with Locality Sensitive Neural Concept Embeddings

Justin Mower, PhD1, Elmer Bernstam, MD, MSE, MS, FACP, FACMI 2, Hua Xu, PhD, FACMI 3, Sahiti Myneni, PhD 2, Devika Subramanian, PhD 1, Trevor Cohen, MBChB, PhD, FACMI 3

1Rice University, Houston, Texas; 2University of Texas Health Science Center, Houston, Texas; 3University of Washington, Seattle, Washington

Abstract

Although pharmaceutical products undergo clinical trials to profile efficacy and safety, some adverse drug reactions (ADRs) are only discovered after release to market. Post-market drug safety surveillance – pharmacovigilance - leverages information from various sources to proactively identify such ADRs. Clinical notes are one source of observational data that could assist this process, but their inherent complexity can obfuscate possible ADR signals. In previous research, embeddings trained on observational reports have improved detection of such signals over commonly used statistical measures. Moreover, neural embedding methods which further encode juxtapositional information have shown promise on analogical retrieval tasks, suggesting proximity-based alternatives to document-level modeling for signal detection. This work uses natural language processing and locality sensitive neural embeddings to increase ADR signal recovery from clinical notes, with AUCs of ~0.63-0.71. Constituting a ~50% increase over baselines, our method sets the state-of-the-art for these reference standards when solely leveraging clinical notes.

Introduction

While the aim of pharmaceutical intervention is to ameliorate or remedy medical issues in patients, such interventions come with risk of side-effects, or adverse drug reactions (ADRs). With nearly half of the US population - and nearly ninety percent of those over the age of 65 - utilizing at least one prescription drug every year1, significant economic and health burdens are associated with unintended effects of medication therapy, even when these are relatively rare. Limiting these adverse effects could both save billions in healthcare expenditure2 and reduce unnecessary patient morbidity and mortality3,4.

Pharmacovigilance systems that surveil on-market drugs or therapeutic biologic products for new or emerging safety concerns, seek to reduce ADRs by actively monitoring sources of observational data, such as the FDA Adverse Event Reporting System (FAERS)5-8. Recently, multi-modal approaches that integrate numerous data sources, such as the literature or claims data, have seen increased investigation9-15. Clinical notes in electronic health records are one source of potentially voluminous and rich ADR signal, containing documentation of patient conditions and treatments, including ADRs and medications captured in real time14,16 and without the reporting biases that have been documented with adverse event report data17,18. Despite this richness, use of clinical notes for pharmacovigilance faces many challenges. Their ever-growing volume necessitates scalable computational models while their inherent complexity as natural language makes computational modeling challenging16. Clinical notes contain narratives generated by healthcare professionals primarily in support of patient care by humans and not primarily for secondary informatics research. In contrast to FAERS reports, which are intended to facilitate secondary research in pharmacovigilance and report drugs based on their suspicion of causing an ADR, drugs and problems may occur in the same clinical note for many reasons, including direct therapeutic relationships or therapeutic relationships with comorbidities. Consequently, drug-ADR signals can be obfuscated by noisy signals unrelated to drug safety. Clinical natural language processing (NLP) is a burgeoning field aimed at addressing these challenges and enabling downstream secondary research19,20.

One way to limit the complexity of unstructured clinical text data is to extract clinical concepts of interest using NLP methods (e.g., named entity recognition and normalization to concepts in a vocabulary)21. This does not provide estimates of association per se but does restrict the complexity and vocabulary involved, facilitating downstream analysis. While this alone could enhance signal detection for ADRs compared to unprocessed clinical text, the choice of signal detection methodology can be critically important. In the realm of pharmacovigilance, signal detection is done by calculating discrete statistical measures, such as the Proportional Reporting Ratio (PRR)22 or Reporting Odds
Ratio (ROR)\textsuperscript{23}, which are known as disproportionality metrics. These measures reveal signals by calculating the relative abundance of report-level co-occurrence for concepts or terms of interest, such as a drug and ADR. However, such measures do not model the similarity between drugs or side-effects - each drug and ADR are considered independently and discretely based purely on co-occurrence between queried terms. Consequently, these methods cannot draw associations between drugs with similar actions (such as members of the statin family), or side-effects with related pathophysiological mechanisms (such as myocardial infarction and stroke). One family of methods which do infer and encode similarity based on shared contexts is neural embedding approaches, as popularized by the skipgram-with-negative-sampling (SGNS) architecture provided in the widely-used word2vec software package\textsuperscript{24,25}.

In pharmacovigilance, a similar neural embedding technique, Portanova et al’s aer2vec\textsuperscript{26}, was found to outperform discrete statistical measures on recovery of curated drug-ADR associations when trained on FAERS reports\textsuperscript{31,26}. A distinguishing feature of the aer2vec approach vis a vis typical word2vec style models is the recapitulation of the training objective at test or query time. In typical word2vec approaches using SGNS, input weights and output weights of a neural network are updated during training to optimize predicting context terms given an observed term as in the optimization objective:

$$\arg\max (O, C) \sum_{(o, c) \in D} \log \sigma (\bar{o} \cdot \bar{c}) + \sum_{(o, \neg c) \in D^t} - \log \sigma (\bar{o} \cdot \neg \bar{c})$$

where $O$ and $C$ represent observed (input) term and context (output) term weight matrices, where $\bar{o}$ and $\bar{c}$ represent single vectors of $O$ and $C$ respectively, drawn from the set $D$ of observed term-pairs $o, c$; $\sigma$ represents the sigmoid function; and $\neg \bar{c}$ represents a random negative sample vector of $C$ drawn from other contexts not likely occurring in the set $D$ (that is, the unobserved context set $D^t$).\textsuperscript{1} Subsequent analysis is performed using the trained input weights only, with the output weights $c$ discarded. In contrast, with aer2vec models the output weights are also retained, and at query time the sigmoid of the dot product between the input weight for a given observed term and the output weight for the suspected context term is calculated. This not only recapitulates the training objective but also provides a probabilistic interpretation of the association between the query terms. Such a formulation satisfies via probabilistic approximation queries of the form: what is the probability given a query term, that I would expect to observe a given second query term in its context? Such queries lend naturally to the field of pharmacovigilance: given some set of observational data, what is the probability of observing a given context in the context of a specified ADR? More importantly, such embedding techniques produce similar probabilities for these queries for analogous drugs and analogous ADRs. This is encoded during the training process, with entities sharing similar contexts having similar representations - a distinguishing feature from discrete statistical approaches.

One important modeling decision in embedding models is how to define context. In the case of aer2vec, context is defined at the document level. If an ADR occurs in a FAERS report, it will be used to predict all drugs in that report\textsuperscript{6}. However, FAERS reports do not contain natural language of the form seen in clinical narratives, where the order of the entities as they appear in context could be informative and important. That is, it may not be advantageous in a clinical narrative for a given drug to inform the prediction of every other entity within that narrative. Instead, it may be advantageous to examine defined, locally constrained context windows, as is the case with standard SGNS models that operate on a sliding window basis\textsuperscript{24}. In such a case, context may be defined by a narrow window of terms on either side of an observed term (e.g., a window radius of 5 around the observed term, for a total window size of 10 plus the observed term). We hypothesized that this narrower definition of context would constrain the associations learned by embedding models to emphasize those of greater utility for pharmacovigilance. Moreover, the relative position of juxtaposed terms within a context window could add additional information, as was observed on general domain tasks in Cohen and Widdows’ recent work\textsuperscript{27}. The derived model, Embeddings Augmented by Random Permutations (EARP), has several variants, each exploiting some element or combination of relative direction and distance from an observed term. That is to say, EARP variants encode not just if a term is likely to appear in a given context window with an observed term, but also what direction (to the left or right) and/or distance away from that

\begin{enumerate}
\item Negative samples are not rigorously guaranteed not to occur in the set $D$. For further information concerning this process, the interested reader is encouraged to see Mikolov et al.
\item Portanova et al introduced two variants of the aer2vec architecture: aer2vec+ which predicts $P(\text{drug|ADR})$ and aer2vec-, which predicts $P(\text{ADR|drug})$. We restricted our investigations to the aer2vec+ model as this performed better in the pharmacovigilance task described in this study.
\end{enumerate}
term might a query term be most likely to appear within that context window. In Cohen and Widdows’ experiments, this improved performance on analogical reasoning tasks; clinical narratives may contain ordered information (e.g., it may be more likely for a drug to precede the utterance of an associated ADR) which could likewise see improved signal detection of drug-ADR associations.

In this paper, we propose an innovative computational method to recover ADR signal from NLP processed clinical notes. We hypothesized that extracting relevant concepts using an established clinical NLP tool would allow for recovery of drug-ADR signals with discrete statistical methods, more so with aer2vec models on account of their encoding of distributional similarity, and even more so when including proximity and direction-based features by generating locality sensitive neural concept embeddings with SGNS and EARP variants. Moreover, we deploy these variants in a novel way, reprising the training objective at query time as with previous aer2vec evaluations. In doing so we provide answers to fundamental questions about how best to model context when representing clinical notes for pharmacovigilance signal detection.

Methods
The data used in this analysis was a subset derived from The University of Texas Health Science Center at Houston (UTHSC-H) clinical data warehouse, with clinical notes spanning 2004 to 2015 collected at UT Physicians clinics and comprising a total of 4.67 million notes. This study has been approved by the Committee for the Protection of Human Subjects (the UTHSC-H IRB) under protocol HSC-SBMI-13-0549. The text from these clinical notes was processed with the Clinical Language Annotation, Modeling and Processing (CLAMP) NLP tool, a tool which has won numerous clinical NLP challenge tasks. We used the “run_comprehensive_pipeline” script included with the software package. Among other output, this produces structured data with semantic annotations for named entities and concepts. For our analysis, we retained only those records which contained at least one concept tagged as “problem” and at least concept tagged as “drug,” resulting in a final collection of 2,545,152 clinical notes. Additionally, for each of these tagged entities extracted by the CLAMP tool, we retained their Unified Medical Language System (UMLS) concept unique identifiers (CUIs) for problems or their RxNorm RXCUIs for drugs. CLAMP additionally extracts these concepts in order of appearance in the clinical narrative, allowing for an ordered list of problem (e.g., ADR) CUIs and drug RXCUIs. For statistical (ROR/PRR) and note-level embedding (aer2vec) procedures, we took the additional step of ensuring that each record had a unique list of CUIs of both types by removing duplicates to mirror the adverse event reports these methods are intended to model (which generally do not contain repeated mentions of the same drug or ADR within a single report).

For comparison to typically deployed statistical methods in pharmacovigilance we utilized the PRR and ROR methods which have seen use by the European Medicines Agency (EMA), Lareb, the FDA, and others. Additionally, we generated note-level embeddings using aer2vec. Note-level embeddings were trained by treating each note as a bag-of-words and training a model to predict every drug in a note given each observed problem. This model, but with reports rather than notes as the unit of analysis, was deployed in previous work leveraging FAERS reports and performed better on the reference standards (subsequently described) than PRR and ROR methods using that data source. Both the PRR and ROR statistics as well as the aer2vec embeddings were generated with the Semantic Vectors (SV) Java package. Aer2vec embeddings were trained similarly to previous work, with 5 negative samples, 5 training epochs and real valued vectors of 200 dimensions.

To process the ordered list of extracted concepts with locality sensitive embeddings, SGNS and EARP embeddings were utilized via the SV package. While SGNS uses a sliding window to restrict training context to a given radius, it does not encode the relative direction or distance from an observed concept within a given window. In contrast, EARP variants seek to encode that directional and proximity specific information. Directional (EARP_dir) and proximity based (EARP_prox) variations were used. In brief, EARP_dir and EARP_prox both consider a window of several terms around a given observed concept in context as in SGNS, but additionally seek to encode information about the direction (EARP_dir) or the direction and proximity (EARP_prox) of the context terms relative to an observed term. Encoding of this positional or locality specific information was accomplished via the use of permutations (operators that shuffle vector coordinates) in both EARP variants used in this analysis. Separate sets of vectors (or neural weights) are kept for observed terms and context terms, which the SV software denotes as embedding and elemental vectors, respectively. Permutation vectors are also kept which distinguish left of an observed term or right of an observed term for EARP_dir and direction and distance to left or right from an observed term for EARP_prox. During training, EARP_air optimizes the sigmoid of the scalar dot product between the embedding vector for the observed term permuted by the
left or right permutation vector (whichever is applicable for the given context term) and the context vectors occurring within its window radius, as in the optimization objective:

$$\arg\max(O, C) \sum_{(o,c) \in D} \log \sigma (\tilde{o} \cdot \Pi_p \tilde{c}) + \sum_{(o,\neg c) \in D'} -\log \sigma (\tilde{o} \cdot \Pi_p \neg\tilde{c})$$

where $O$ represents the weight matrix for the observed terms (the input weights) with $\tilde{o}$ a vector of $O$, $C$ represents the weight matrix for context terms (the output weights) with $\tilde{c}$ a vector of $C$, and $p$ represents the position of $c$ relative to $o$, in the set $D$ of observed positional term-pairs $o,c$; $\sigma$ represents the sigmoid function; $\Pi$ represents the permutation associated with position $p$; and $\neg\tilde{c}$ represents a random negative vector of $C$ drawn from other contexts not likely occurring in the set $D$ (that is, $D'$). EARP$_{prox}$ is the same, only the permutation encoding the direction and distance from the observed term is utilized instead of one just encoding relative direction. These permutations are deliberately constructed such that similarity as measured by the scalar product will degrade gracefully as the sliding window is traversed. For example, with $x$ as a random vector and $\Pi_p$ the permutation for position $p$, $\Pi_1(x) \cdot \Pi_2(x) > \Pi_1(x) \cdot \Pi_2(x)$. Figure 1 depicts the difference for an example text window and specific observed term within that window. Additional details on EARP and its variants can be found in Cohen and Widdows for the interested reader. In every case, models were trained with 5 negative samples, a window radius of 5, 5 training epochs and real valued vectors of 200 dimensions.

Given a window around a cue concept, interferon beta-1a:

... acute liver failure, mefenamate, interferon beta-1a, insulin, acute kidney failure...

Figure 1. Example schematic depicting the process for optimizing the representation of a given observed concept based on the embedding methods deployed: aer2vec, SGNS, EARP$_{dir}$, and EARP$_{prox}$. In all cases, the prediction of nearby terms is done using the sigmoid of the dot product between the observed concept (permuted if applicable) and the applicable context terms. In this study, concepts would be replaced with UMLS CUIs for ADRs and RxCUIs for drugs as extracted by the CLAMP tool from clinical notes but are presented here in human readable form.

The ability of each method to recover ADR signal was measured by the ability to rank positively labeled drug-ADR pairs before negatively labeled pairs using Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) and the Average Precision (AP) score using two commonly used ADR reference standards. The first reference standard is the Exploring and Understanding Adverse Drug Reactions (EUADR) reference standard, comprising 44 positive and 50 negative drug-ADR pairs for a wide range of drug classes across 10 ADRs. The second is the reference standard developed by Ryan et al for the Observational Medical Outcomes Partnership (OMOP), now known as the
Observational Health Data Sciences and Informatics (OHDSI) program. The OMOP reference standard comprises 165 positive and 234 negative drug-ADR pairs for a wide range of drug classes for 4 serious ADRs of interest. In either case, the examples are manually curated drug-ADR pairs reviewed for known interactions (positive examples) or lack of association (negative examples). After mapping to CUIs and RxCUIs and removal of example pairs for which an entity did not occur in the clinical notes used, 93 and 39 example pairs were retained for the EUADR and OMOP sets respectively. The dropped EUADR example pair was nimesulide with acute liver injury, a positive example in the reference set. Nimesulide has never filed for FDA approval and reassuringly does not appear in the clinical notes used in this study. The two OMOP examples dropped were both positive example pairs: alatrofloxacin with acute liver injury (a prodrug of another example, trovafloxacin, that mapped to the same CUI as that example) and capreomycin with kidney injury (capreomycin did occur in the data but was not retained in the mapping process to RxCUIs). PRR/ROR were ranked according to their values, and aer2vec and SGNS embeddings were ranked according to the sigmoid of the dot product of the trained ADR and drug vectors (input and output weights, respectively). Directional and proximity-based EARP embeddings were ranked according to the maximum and average sigmoid of the dot product of the trained ADR vector looking to the left and to the right, and drug vectors. For EARP_{dir} this is achieved by indexing the ADR vector by the generated permutation vectors for the right and left directions of the observed term. For EARP_{prox} this accomplished by indexing the ADR vector by the closest permutation vector to the right (P_{1} in Figure 1) and left (P_{-1} in Figure 1) of the observed term. In either case, the resulting set of sigmoided values are averaged together (denoted μ) or the largest is selected (denoted max).

### Results

Performance of each method is shown in Table 1. The best performance on the OMOP and EUADR sets are obtained with the EARP_{dir} and SGNS models, respectively, with EARP_{dir} performing best overall as estimated by a weighted average across the sets. On both sets, aer2vec performs substantially better than the ROR/PRR measures. However, the use of narrow context definitions and encoding of directional or proximity information results in further – and larger – improvements, with best performing models attaining to a relative increase in AP over statistical baselines of ~50% and ~55% for the EUADR and OMOP sets respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>OMOP AUC</th>
<th>OMOP AP</th>
<th>EUADR AUC</th>
<th>EUADR AP</th>
<th>Average AUC</th>
<th>Average AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRR/ROR</td>
<td>0.484</td>
<td>0.391</td>
<td>0.530</td>
<td>0.470</td>
<td>0.493</td>
<td>0.406</td>
</tr>
<tr>
<td>aer2vec</td>
<td>0.577</td>
<td>0.477</td>
<td>0.547</td>
<td>0.537</td>
<td>0.571</td>
<td>0.488</td>
</tr>
<tr>
<td>SGNS</td>
<td>0.629</td>
<td>0.576</td>
<td>0.709</td>
<td>0.651</td>
<td>0.644</td>
<td>0.601</td>
</tr>
<tr>
<td>EARP_{dir}</td>
<td>0.645</td>
<td>0.599</td>
<td>0.647</td>
<td>0.651</td>
<td>0.645</td>
<td>0.609</td>
</tr>
<tr>
<td>EARP_{prox}</td>
<td>0.650</td>
<td>0.604</td>
<td>0.658</td>
<td>0.663</td>
<td>0.652</td>
<td>0.615</td>
</tr>
</tbody>
</table>

Table 1. Performance comparison across tested methods. PRR/ROR were scored according to their ranked values (simply denoted * here). The small variations between PRR and ROR formulas did not change the ordering of the example pairs, and so they resulted in the same AUCs and APs for both sets. Ranking by sigmoid of the dot product of the example pairs [denoted σ(*)] was used for aer2vec and SGNS embeddings. Ranking by the maximum [max(σ(*))] and average [μ(σ(*))] of the left and right permuted ADR and drug pair examples is reported for EARP variants. Average denotes the performance when taking the average weighted by the number of example pairs in each reference standard. Bold values indicate best performance for a given row.

Notably, EARP based models consistently achieve their best performance when taking the average of the sigmoid of the dot product [μ(σ(*))]. Document level methods PRR/ROR and aer2vec both have lower relative performance than sliding window / locality sensitive methods, regardless of data set or scoring measure. Comparing document level
methods, aer2vec performed consistently better than PRR/ROR, consistent with previous findings\(^{11,26}\). EARP\(_{dir}\) performed better than EARP\(_{prox}\) and SGNS on OMOP, whereas EARP\(_{prox}\) performed better than EARP\(_{dir}\) on EUADR but worse than SGNS. In terms of overall results, calculated by the weighted average of performance on both reference standards, EARP\(_{dir}\) has the best overall performance, followed by EARP\(_{prox}\), SGNS, aer2vec and then PRR/ROR.

Results are compared to previous work in Table 2. These results are not strictly comparable as Li et al\(^{13}\) and Malec et al\(^{15,41}\) both use different subsets of the OMOP reference standard and different clinical note sources. In this study, we utilize 397 of 399 example pairs in the OMOP reference standard. Malec et al 2018 resolved 396 of the pairs and Malec et al 2021 resolved 163 of the pairs as a result of enforcing more strict inclusion criterion. Li et al utilized 214 of 399 example pairs, again due to strict inclusion criterion. For our study, we required only that an entity occur at least once in the clinical notes to be included. For clinical notes, Li et al used 0.3 million notes and structured data from New York Presbyterian Hospital at Columbia University Medical Center, and Malec et al used 2.2 million clinical notes from UTHSC-H spanning 2004 – 2012. It should also be noted that Li et al used additional sources of information, including claims and FAERS data, to reach a final adjusted AUC of 0.89, supporting the promise of multi-modal data integration to improve performance on this task. Similarly, Malec et al improved performance over their 2018 work utilizing information from the literature to identify confounding variables for modeling of clinical notes in their 2021 paper. Malec et al was included in Table 2 on the account that literature data was used to adjust signal from the clinical notes but does not contribute signal itself. Neither group has thus far deployed their methods on the EUADR reference standard using clinical notes.

![Table 2](image)

Table 2. AUC performance comparison to other clinical note methods utilizing the OMOP reference standard, with EUADR results presented for additional context. Bold indicates best performance for a given row. PRR/ROR results calculated in this study are presented as a baseline. In the unadjusted results, the best performing method utilizing only information from clinical notes is presented. Adjusted results for comparators are after adjustment via reference set supervised confounding analysis (Li et al) or literature-derived estimates of confounding (Malec et al). Results for Li et al and Malec et al were curated from the publications. This study does not perform any confounding adjustment and solely uses clinical note-derived data. These comparisons are not strictly comparable on account of differing subsets of the OMOP reference standard in addition to different clinical data sources and are presented only for context.

Discussion

To our knowledge, this is the highest performance achieved on the OMOP reference standard using information solely from clinical notes without any additional sources of information or context. That proximity methods provide a significant 50% lift or more over baseline statistical methods or even document level is a key finding. With the increase of SGNS results over aer2vec, and the overall performance increase of EARP methods, juxtapositional information is a main contributor to recovering signal on these reference standards with NLP extracted clinical concepts. EARP\(_{dir}\) did perform the best overall, indicating that proximity can provide additional signal, though EARP\(_{prox}\) did not yield additional improvements. Notably, this supports the hypothesis that proximity information (e.g. how closely a side-effect is mentioned in a clinical narrative to a putative drug) can improve signal yield from clinical notes in drug-ADR prediction tasks, and may be equally beneficial for other tasks. Interestingly, in further experiments (not shown in Table 1), using only the cosine between the input weight embedding vectors for both drugs and ADRs, as is typically the case with word embedding analyses (where output weights are discarded, prohibiting recapitulation of the training objective), had relatively poor performance compared to the permuted embedding vectors for ADRs and context vectors for the drugs. This is to be expected. The training objective of the SGNS and EARP methods optimize for the sigmoid of the dot product between an embedding vector (positionally permuted in the case of EARP) and its context vectors; using the same objective in the testing procedure should produce the optimal result when estimating the...
probability of association (i.e., \( P(\text{drug}|\text{ADR}) \)), and this was observed here. Still, as many word vector embedding approaches discard the trained output weights for context vectors, this finding is notable, and further underscores the utility of this fundamentally different way of retrieving encoded information from neural embeddings when modeling observational data (originally demonstrated by aer2vec). Additional research into the potential utility of context embeddings generated during training for other applications may be warranted.

In line with the multi-modal performance improvements from our previous work leveraging FAERS and biomedical literature and the success of Li et al and Malec et al, future work should explore the use of these locality sensitive embeddings with other sources of ADR signal. Additionally, we did not perform grid searching to optimize window radius, negative sampling, or any other hyperparameters for models trained here, nor did we perform any modifications to the CLAMP clinical note preprocessor. It seems possible that optimizing these parameters and/or the text preprocessing could further improve performance. The reference standards themselves are human curated artifacts and subject to potential bias and are limited to a total of 10 ADRs of interest. Further study is warranted as to the generalizability of this method beyond these reference standards. Our work was also confined to a subset of clinical notes collected in the Houston metropolitan area during a fixed time window. It is possible that other clinical note repositories, such as that used by Li et al, would not see the same performance breakdown as observed here. Future work is planned to investigate if these findings hold true across NLP preprocessing tools and across clinical data repositories. Furthermore, retrofitting and subword embeddings have been used to enhance FAERS-based aer2vec performance on this task\(^\text{10}\), and could be used in future research for clinical-based aer2vec and EARP models. Li et al used time dependency for their inclusion criterion in preprocessing their clinical notes and Malec et al sought to causally model drug-ADR assertions and reduce confounders, both of which could be complementary to the work presented here. Finally, the methods developed here could have generalization beyond the task of pharmacovigilance, extending the use of clinical note modeling for drug repurposing or other tasks.

**Conclusion**

This study presents a locality sensitive embedding technique applied to NLP processed clinical notes to achieve state-of-the-art performance for models using clinical notes alone on two ADR reference standards. NLP preprocessing combined with document level statistical approaches in PRR/ROR did not produce appreciable signal when scored against the two reference standards. Document level embeddings of this same data did recover more signal than statistical approaches, consistent with previous research\(^\text{26}\). While performance variations between specific embedding techniques are present, all locality sensitive embedding techniques substantially outperformed statistical and document level embedding techniques, underscoring the added utility of retaining proximity features.

**Acknowledgements**

This work was supported in part by US National Library of Medicine grant (R01 LM011563); the National Center for Advancing Translational Sciences (NCATS) under awards UL1TR000371 and U01TR002393; the Cancer Prevention and Research Institute of Texas (CPRIT), under award RP170668 and the Reynolds and Reynolds Professorship in Clinical Informatics. We’d also like to acknowledge UT staff: Susan Guerrero, Alejandro Araya, and Joseph Applegate for their support with UT clinical data access and IT support, and Yaoyun Zhang and Firat Tiryaki for their assistance with the CLAMP tool.

**Conflicts of Interest**

Dr. Xu and The University of Texas Health Science Center at Houston have research-related financial interests in Melax Technologies, Inc.

**References**


28. Clinical Data Warehouse - UTH BIG - The University of Texas Health Science Center at Houston (UTHealth) School of Biomedical Informatics [Internet]. [cited 2021 Aug 24]. Available from: https://redcap.uth.tmc.edu/cdwstats/stats-mpi.htm


35. Poluzzi E, Raschi E, Piccinni C, Ponti FD. Data mining techniques in pharmacovigilance: analysis of the publicly accessible FDA adverse event reporting system. In: (AERS),” in Data Mining Applications in Engineering. 2012.


Predicting Hormonal Therapy Medication Discontinuation for Breast Cancer Patients using Structured Data in Electronic Medical Records

Congning Ni, ME¹, Jeremy L. Warner, MD, MS², Bradley A. Malin, PhD¹,², ³, Zhijun Yin, PhD¹,²

¹Department of Computer Science, Vanderbilt University, Nashville, TN USA; ²Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN USA; ³Department of Medicine, Division of Hematology/Oncology, Vanderbilt University Medical Center, Nashville, TN USA; ⁴Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, U.S.

Abstract

Hormonal therapy (HT) reduces the risk of cancer recurrence and the mortality rate for patients with hormone-receptor-positive breast cancer. However, it is estimated that half of the patients fail to complete the standard 5-year adjuvant treatment protocol. We investigate the extent to which certain types of structured data in electronic medical records (EMRs), namely conditions, drugs, laboratory tests and procedures, as well as when such data is entered EMRs, can forecast HT discontinuation. Our experiments with EMR data from 2,251 patients showed that machine learning models based on these data types achieve fair performance (AUC of 0.65). More importantly, the performance was not statistically significantly different when fitting a model using all or only one feature type, suggesting that the model is robust to missing information in the EMR.

Introduction

Breast cancer is the most common invasive malignancy in the United States¹ and the second leading cause of death for women diagnosed with cancer². In 2018, there were over 20 million newly-diagnosed breast cancer patients worldwide (11.6% of all new cancer patients), while over 600,000 breast cancer patients died (6.6% of all cancer-related deaths)³. Breast cancer is difficult to cure and requires prolonged treatment, even at an early stage, to ensure the effectiveness of treatments and mitigate recurrence⁴,⁵.

The treatment of breast cancer varies by stage and the comorbidities exhibited by a patient. A combination of surgery, radiotherapy and chemotherapy is the most common initial treatment for early breast cancer (EBC), inclusive of stages I through III⁶,⁷. Moreover, for patients with hormone-receptor-positive EBC, it has been shown that hormonal therapy for at least five years can substantially reduce recurrence rates for at least 15 years after diagnosis⁸. As such, once the initial phase of treatment is complete, it is recommended that patients with hormone-receptor-positive breast cancer take long-term hormone therapy medications to maintain the therapeutic effect and prevent cancer recurrence⁹,¹⁰.

Despite the benefits, patients with EBC generally exhibit a high rate of hormonal therapy discontinuation¹¹. This is a result of numerous factors that include, but are not limited to, the side effects of the prescribed medications, the development of local recurrence or metastatic disease while on therapy, socioeconomic pressures, and the personality or tendencies of the patient¹²-¹⁴. The discontinuation of hormone therapy will reduce the therapeutic effect¹⁵,¹⁶, leading to cancer recurrence and increased mortality that could have been prevented¹⁷. To date, the most common approach to stave off discontinuation is an intervention. Over the years, there have been various investigations into how to design and test practical strategies to assist patients with treatment adherence¹⁸-²⁰. However, the past decade has also witnessed continuous growth in the number of patients with breast cancer, as well as the application of hormonal therapy. Given that there is a general shortage of healthcare providers, which is more pronounced amongst specialists, healthcare systems would benefit from support in allocating resources, such that patients with a higher risk of medication discontinuation receive appropriate attention.

In this paper, we examined the effectiveness of structured data in electronic medical records (EMRs)²¹ for predicting medication discontinuation of a five-year course of hormonal therapy. We focused on approximately 2,000 breast cancer patients prescribed hormonal therapy medications at Vanderbilt University Medical Center (VUMC) between 1998 and 2012. We examined how four types of common structured EMR data as features: 1) medical conditions, 2)
prescribed and administered drugs, 3) laboratory tests, and 4) procedures can enable discontinuation prediction. Particularly, we focused on two research questions:

**RQ1:** To what extent do EMR data documented before the initiation of hormonal therapy (before-HT) and in the first six months of hormonal therapy (early-HT) support discontinuation prediction?

**RQ2:** How robust are discontinuation prediction models to missing certain types of features in the EMR?

The answer to RQ1 helps characterize the predictive ability of features crafted in different timespans. By contrast, RQ2 examines whether a lack of certain types of features in the EHR influences our ability to predict if a patient will discontinue a medication regimen. This is important because there is substantial variability in the availability of the various feature types in the EMRs of breast cancer patients. Simply neglecting records with missing information could lead to concerns of bias and fairness in the resulting models.

**Methods**

**Data** The data for this study was collected from a de-identified version of the VUMC’s EMR system. All dates in this resource were shifted by a random -1 to -365 days, consistently applied on a per-patient basis. We focused on patients diagnosed with stage I, II or III breast cancer and prescribed any of the following hormonal therapy medications: anastrozole, exemestane, letrozole, raloxifene and tamoxifen. Considering that the latest record from our dataset was November 1, 2017 (after which VUMC switched to the Epic system), we removed patients who started hormonal therapy after November 1, 2012. As such, we confined our study cohort to EBC patients who prescribed their first hormonal therapy medication between January 1, 1998, the earliest record in our dataset, and November 1, 2012.

A medication discontinuation event is said to have occurred if the time between the start of hormonal therapy and the final medication recording date is shorter than four years and three months. We applied a nine-month grace period before the end of the five-year treatment plan for several reasons. First, it is general practice for breast cancer patients on hormonal therapy to visit an oncologist every six months. Moreover, the intervals between appointments are unlikely to be precisely six months. Lastly, a typical prescription is for a 90-day supply. This study was approved as non-human subjects research by the Vanderbilt University Institutional Review Board (#180761).

![Research Pipeline Diagram](image_url)

**Figure 1.** The research pipeline designed to predict hormonal therapy discontinuation from structured EMR data.

**Research Pipeline** Figure 1 illustrates the research pipeline. To examine the two research questions, we designed our experiments as follows. For each patient, we partitioned the EMR data into two non-overlapping periods: 1) the before hormonal therapy (HT) phase, or before-HT, which corresponds to the time between the diagnosis date and the initiation of hormonal therapy medications and 2) the early-HT phase, which corresponds to the time after the first six months of HT (RQ1). For each phase, we trained 1) four models – one for each type of structured EMR data (i.e., medical conditions, prescribed and administered drugs, laboratory test results, or procedures) and 2) a model with all
four types of data together. In doing so, we examined if the prediction task was likely to succeed when a patient had only one type of feature available, which we compared to the situation when all four types were available (RQ2).

In model training, we focused on the existence of a feature, as opposed to its exact value. For example, when training a model using Drugs, if a patient took Tramadol, then the corresponding feature value for this patient was set to 1 and 0 otherwise. We also included cancer stage (0 for stage I or II; and 1 for stage III) and (normalized) age at the diagnosis in each model. It is notable that continuous values (e.g., laboratory test results) were binarized as to whether the laboratory test was performed.

**Model Fitting and Evaluation** We applied nested cross-validation to train and test the models. First, an inner 10-fold cross-validation (i.e., splits of training and validation datasets) was applied to tune the hyperparameters of the model. Next, the model was refit with the entire training and validation dataset to obtain the updated coefficients. The resulting model was then applied to the test dataset. To assess the generalizability of a model, this process was repeated 10 times through the outer 10-fold stratified shuffle splits (i.e., splits of training plus validation and test datasets). We utilized a stratified shuffle split strategy to ensure the ratio of the number of patients discontinuing hormonal therapy medications to patients who completed a 5-year treatment protocol was equal to the original dataset. We applied the same nested cross-validation splits for each sub-cohort, such that we were able to directly compare model performance using the same feature type in the before and early-HT phases.

We investigated five machine learning methods: $k$-nearest neighbors, logistic regression with LASSO regularization, logistic regression with Ridge regulation, naïve Bayes, random forests, and support vector machines with a linear kernel. Due to space limits, we only presented the logistic regression model with Ridge regularization since it exhibited the best performance, on average. We reported the average and standard deviation (SD) for 1) accuracy, 2) precision, 3) recall, 4) F1, the weighted average of precision and recall, and 5) area under the receiver operating characteristic curve (AUC). A Student’s t-test was applied to determine if the performance of a model fit with each type of feature was statistically significantly different between the before-HT and early-HT phases. We also applied a t-test to compare the performance of models with one feature type only against the combination of all four types.

**Most Predictive Features** For each model with a single data type (i.e., conditions, drugs, laboratory tests, or procedures), we ranked the features according to the average of their coefficients in the outer 10-fold split. For each data type, we compared and analyzed the 10 most informative features of the models between the before-HT and early-HT phases.

**Sensitivity Analysis on Decision Threshold** We investigated how the model performance, especially precision and recall, changes as a function of the decision threshold. This is because, in practice, end users of the model (e.g., healthcare providers) may want to maximize the chance that they recognize breast cancer patients at risk for medication discontinuation. In such a situation, they may prefer to apply a lower decision threshold, which will ensure a higher recall at the cost of a lower precision.

![Diagram](image)

**Figure 2.** Summary statistics stratified by data types. *Unique* represents the number of categories in the machine learning model of the corresponding feature type.
Results

Summary of Study Data Figure 2 provides basic statistics about the patient population. There were 4,606 patients who initiated hormonal therapy medications, and 2,649 of which started between 1998 and 2012. We restricted the population to patients who were diagnosed with cancer at stage I, II and III and had at least one of the four feature types of interest in both the before-HT and early-HT phases. This yielded 2,251 patients diagnosed with EBC and initiated hormonal therapy treatment. This group had an average (SD) age of 53.3 (17.3) at cancer diagnosis. Based on the medication discontinuation criteria, 47.7% of the patients discontinued hormonal therapy medications before the completion of a five-year treatment protocol. There were 1,257 patients with four types of features available in both the before-HT and the early-HT phase.

Model Performance Table 1 summarizes the performance of the Ridge regression model with respect to each of the EMR data types during the before-HT and early-HT phases.

There are several notable observations to highlight. First, each EMR data type in the before-HT and early-HT phase outperformed a random guess in terms of accuracy, precision, and AUC, which indicates that the information from both phases supported discontinuation prediction (RQ1). However, none of these features resulted in a statistically significantly better recall compared to a random guess. Additionally, in before-HT, all features except the laboratory test results performed similar or better than random guess regarding F1.

Secondly, the performance of the before-HT and early-HT phases varied according to the feature types. For example, the model trained with before-HT conditions outperformed the model trained with early-HT conditions on accuracy, precision, F1 and AUC. By contrast, the model trained with early-HT drugs outperformed the model trained with before-HT drugs in all the metrics. For laboratory test results, however, the model trained with early-HT features outperformed the model trained with before-HT features on accuracy, precision, F1 and AUC. The results suggest there is no statistically significant difference in supporting prediction performance among various sole data types.

Finally, it is notable that the combination of features from all four data types failed to achieve a statistically significant advantage over models based on each feature type independently. This is intriguing because it suggests that there is redundancy in the information across the EMR data types. As such, it should be possible to predict the medication status of patients who did not have all of the data types available.

Table 1. Performance of logistic regression with Ridge regularization for different EMR feature types. *p < 0.05, **p < 0.01 and ***p < 0.001 under a Student’s t-test to measure whether the performance was significantly different between the before-HT and early-HT phase. The random guess performance was based on the 47.7% medication discontinuation rate observed in the study cohort. The best score for each metric is shown in bold.

<table>
<thead>
<tr>
<th>Features</th>
<th>HT Phase</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conditions</td>
<td>Before</td>
<td>0.63 (0.02)**</td>
<td>0.62 (0.04)**</td>
<td>0.46 (0.02)</td>
<td>0.53 (0.02)**</td>
<td>0.63 (0.02)*</td>
</tr>
<tr>
<td></td>
<td>Early</td>
<td>0.60 (0.03)</td>
<td>0.56 (0.04)</td>
<td>0.44 (0.05)</td>
<td>0.49 (0.03)</td>
<td>0.60 (0.03)</td>
</tr>
<tr>
<td>Drugs</td>
<td>Before</td>
<td>0.59 (0.02)</td>
<td>0.57 (0.03)</td>
<td>0.50 (0.04)</td>
<td>0.53 (0.03)</td>
<td>0.61 (0.03)</td>
</tr>
<tr>
<td></td>
<td>Early</td>
<td>0.61 (0.02)*</td>
<td>0.61 (0.03)**</td>
<td>0.52 (0.05)</td>
<td>0.56 (0.03)*</td>
<td>0.65 (0.02)**</td>
</tr>
<tr>
<td>Laboratory</td>
<td>Before</td>
<td>0.61 (0.03)</td>
<td>0.54 (0.05)</td>
<td>0.40 (0.05)</td>
<td>0.45 (0.04)</td>
<td>0.61 (0.03)</td>
</tr>
<tr>
<td>tests</td>
<td>Early</td>
<td>0.64 (0.02)*</td>
<td>0.58 (0.03)*</td>
<td>0.42 (0.04)</td>
<td>0.49 (0.03)*</td>
<td>0.64 (0.03)*</td>
</tr>
<tr>
<td>Procedures</td>
<td>Before</td>
<td>0.60 (0.02)</td>
<td>0.57 (0.03)</td>
<td>0.47 (0.04)</td>
<td>0.51 (0.03)</td>
<td>0.62 (0.02)</td>
</tr>
<tr>
<td></td>
<td>Early</td>
<td>0.61 (0.03)</td>
<td>0.58 (0.04)</td>
<td>0.50 (0.05)</td>
<td>0.54 (0.04)</td>
<td>0.65 (0.02)**</td>
</tr>
<tr>
<td>Combined</td>
<td>Before</td>
<td>0.62 (0.03)</td>
<td>0.58 (0.02)</td>
<td>0.51 (0.02)*</td>
<td>0.53 (0.02)</td>
<td>0.64 (0.01)</td>
</tr>
<tr>
<td></td>
<td>Early</td>
<td>0.63 (0.03)</td>
<td>0.56 (0.05)</td>
<td>0.46 (0.06)</td>
<td>0.51 (0.05)</td>
<td>0.65 (0.04)</td>
</tr>
<tr>
<td>Random Guess</td>
<td></td>
<td>0.50</td>
<td>0.48</td>
<td>0.50</td>
<td>0.49</td>
<td>0.50</td>
</tr>
</tbody>
</table>
Most Predictive Features  Figure 3 illustrates the most predictive features from the conditions-only model. In the before-HT phase, these include 1) Mitral valve disorder, 2) Secondary malignant neoplasm of bone and 3) Edema, a condition characterized by an excess of watery fluid collecting in the cavities or tissues of the body. In the early-HT phase, these include 1) Hypercalcemia, a condition in which the calcium level in blood is above normal; 2) Secondary malignant neoplasm of brain and spinal cord; and 3) Congestive heart failure.

![Figure 3](image)

Figure 3. The most predictive features for the before-HT (left) and early-HT (right) phases for the model, based solely on conditions. Each bar represents the mean (±1 SD) feature importance score. (PMN: primary malignant neoplasm; SMN: secondary malignant neoplasm).

Figure 4 illustrates the most predictive features when using only drugs in the model. The most predictive features in both the before-HT and early-HT phases include 1) Lasix (furosemide), a medication to treat congestive heart failure; 2) Prednisone, a class of drugs known as corticosteroids to decrease the immune system's response to various diseases; 3) Sodium (Docusate in before-HT) and 4) high blood pressure treatments, such as Toprol in the before-HT phase, and Nifedipine in the early-HT phase. Moreover, the third most predictive drug in early-HT is Xeloda (capecitabine), a chemotherapy medication used to treat breast cancer that is not typically observed in before-HT.

Figure 5 illustrates the most predictive features when using only laboratory tests in the model. Here, we summarize several notable findings. First, in both phases, Microcytosis, a condition in which red blood cells are unusually small as measured by their mean corpuscular volume, was one of the most predictive features. Second, in the before-HT phase, the most predictive features are related to infections and tests for their detection, such as gram stain and Antigen C. difficile. The other predictive features in the before-HT phase include rheumatoid factor, a test applied to diagnose the presence of rheumatoid arthritis, and prolactin, a test investigating the unexplained flow of breast milk, which is a prognostic biomarker for breast cancer recurrence. Third, in the early-HT phase, the most predictive features are eosinophils, investigation of possible acute interstitial nephritis; HER2 fish, determining if a breast cancer is HER2-positive; and estradiol, a blood test to measure the amount of estradiol in a person’s blood which is sometimes used to check the effectiveness of aromatase inhibition. Note that one of the most predictive features in the before-HT phase is Lead whole blood, which suggests a relationship between lead exposure and female breast cancer.

Figure 6 illustrates the most predictive features when using only procedures in the model. While many of the top features are designed to detect occult or suspected metastatic disease, there are several additional notable findings. First, [SNOMED] 447421006 - prophylactic mastectomy, a surgery to remove one or both breasts to reduce the risk of developing breast cancer, is among the most predictive features in the before-HT phase. Second, [SNOMED] 22418005 - bilateral simple mastectomy, as well as [CPT4] 11970 - replacement of tissue expander, and certain psychotherapy services are also among the most predictive features in the early-HT phase. Additionally, we observed certain psychotherapy services in early-HT, like [CPT4] 99354 - Prolonged evaluation and management or
psychotherapy service(s) (beyond the typical service time of the primary procedure) in the office or other outpatient setting requiring direct patient contact beyond the usual service, indicating that psychological distress may lead to treatment discontinuation.\textsuperscript{28}

Figure 4. The most predictive features for the before-HT (left) and early-HT (right) phases for the model, based solely on drugs. Each bar represents the mean score ($\pm 1$ SD) of feature importance.

Figure 5. The most predictive features for the before-HT (left) and early-HT (right) phases for the model, based solely on laboratory tests. Each bar represents the mean score ($\pm 1$ SD) of feature importance.

Finally, it should be noted that Cancer Stage is among the most predictive features in several models. For example, it ranks first (coefficient of 0.46) in the model with before-HT procedures, fourth (coefficient of 0.58) in the model with before-HT conditions, fifth (coefficient of 0.60) in the model with before-HT Drugs, and fourth (coefficient of 0.48) in the model with early-HT Drugs.
Figure 6. The most predictive features for the before-HT (left) and early-HT (right) phases for the model, based solely on procedures. Each procedure is represented as vocabulary concept code. The concepts are available online.  

Sensitivity Analysis on Decision Thresholds It should be noted that the model performance reported in Table 1 is based on a default decision threshold of 0.5. Figure 7 shows how the precision and recall change with different decision boundaries in a logistic regression model with only drugs in the early-HT phase. It can be seen that when the decision threshold is shifted to 0.3, the recall would increase to 0.95 with a precision of 0.50. By contrast, with a decision threshold of 0.72, the model would have a recall of 0.10 and a precision of 0.8. The AUC is 0.65 with an optimal threshold of 0.52 based on Youden’s J statistic. The preferable threshold of the prediction model could be adjusted based on different requirements of user needs.

Figure 7. Change in model performance as a function of decision threshold for the logistic regression model based on drugs in the early-HT phase.

Discussion
This study showed that structured EMR features documented in the before-HT phase and the early-HT phase have the potential to assist in the prediction of breast cancer patients at risk for discontinuation in several ways. The two phases
of features resulted in predicting models with an average AUC of 0.64 and 0.65, respectively. Despite the fair model performance, we demonstrated that by varying the threshold, we could end up with a model with a high recall (0.95) at a cost of a low precision (0.50). This could be helpful in practice that we want to have all the patients with medication discontinuation risk detected. By contrast, we could also choose a threshold that ends up with a higher decision (0.80) threshold to avoid false alarms but at a cost of a low recall (0.10).

Prior to this study, it was shown that the EMR data we used may help in the prediction of treatment discontinuation (e.g., the abandonment of treatment after the start of antidepressants\textsuperscript{32}, the worsening of asthma\textsuperscript{33}, etc.). However, the moderate prediction performance achieved by such models suggests that structured EMR data is unlikely to be sufficient to represent all of the associated factors. For example, it has been further shown that there is a non-trivial association between discontinuation behavior and financial issues\textsuperscript{34}, psychological status\textsuperscript{24,35}, and social support\textsuperscript{36}. It suggests that unstructured text, such as clinical notes, clinical communications, and online communications between patients and healthcare providers in a patient portal may provide such information\textsuperscript{37,38} and can be useful in cancer research\textsuperscript{39,40}.

Our investigation into the robustness of prediction models (RQ2) indicated that the models fitted with a single type of feature shared quite similar performance with the model fitted using the combination of the four types of features, suggesting strongly correlated (or perhaps complementary) information across the four types of features. This is not surprising because it is possible that a patient underwent certain lab tests, was diagnosed with several conditions, and then was subject to procedures and prescribed medications based on the composite of the findings. Our findings confirmed our conjecture that a key benefit of only relying on a single type of feature in the EMR is needed for, instead of all the four types of features, is that many patients can be predicted on their medication discontinuation status prediction, which allows for more flexible prediction models in practice. For example, in the dataset used in this study, we observed that 44.2\% of the 2,251 patients were lacking data about one or more feature types. If we only relied on the model built with four types of features, all these patients would have been excluded from prediction, leading to concerns of bias and fairness.

Our review of the most informative features yielded several findings of note. First, advanced-stage cancer (i.e., stage III) ranked in the most predictive features for drug and procedure models. Though cancer stage was not in the most predictive features for the model based on conditions, the highest ranking conditions included secondary neoplasms, presumably representing a metastatic spread of the original early breast cancer. Thus, cancer relapse was a strong predictor that led to hormonal therapy medication discontinuation. For example, several of the most predictive conditions in both the before-HT and early-HT phases were related to cancer relapse, such as secondary malignant neoplasms. Several features related to the diagnosis of conditions like congestive heart failure (CHF) in the early-HT phase, suggested that a subset of the patients discontinued treatment early due to complications in their cancer journey. Specifically, CHF is a known complication of the commonly used chemotherapeutic doxorubicin, as well as the HER2 inhibitor trastuzumab, which would be used in the cases of dual-positive (ER+/HER2+) breast cancer. To the best of our knowledge, this study is the first to suggest a direct correlation between this adverse event and subsequent hormonal treatment discontinuation, a finding which bears further study. Furthermore, cancer antigen 27.29 (a tumor biomarker) and prolactin are often tested when a clinician has a strong suspicion of cancer relapse and demonstrate value even without accounting for the actual test results. Third, in the early-HT phase, women undergoing bilateral simple mastectomy and replacement of tissue expander may require additional breast reconstructions. These surgeries can result in a high rate of dissatisfaction\textsuperscript{41} and, thus, affect their hormonal treatment continuation as well.

We acknowledge that there are limitations to this study, which we believe serve as a basis for future work. First, the best AUC our models achieved was 0.65. In future work, we can improve model performance by strategies like including more EMR data types as features, increasing data volume and using advanced models. Second, we only considered the presence of each feature instead of its actual value in the EMR system. This certainly obscures information that is likely to be relevant for predictive purposes. For example, it is known that laboratory test values are informative when classifying hospital-acquired complications\textsuperscript{42}. Future work can integrate such information into the models. Third, there were redundant HCPCS codes, a collection of standardized codes that represent medical procedures, supplies, products and services, in the EMR data. For example, we treated J3487 (Injection, zoledronic acid (Zometa), 1 mg) and Zoledronic acid (5 ML zoledronic acid 0.8 MG/ML Injection) as two separated features, whereas they are presented differently because of varying vial sizes, which help billing purposes but have no clinical distinction. We can leverage domain knowledge to intentionally aggregated features with the same meaning before the analysis to achieve better performance in future work. Fourth, we only considered four EMR data types. This set could be augmented by incorporating additional structured data types (e.g., vital signs, menopause status or breast cancer subtype) and data from unstructured EMRs, such as clinical notes and communications. Fifth, the medication
records that we relied on to identify the medication discontinuation status do not necessarily indicate that a patient actually took the medication. Another notable limitation of this study is that the split into before and early-HT phases is subject to several possible errors: 1) A patient may be prescribed a hormonal therapy but may not begin it immediately (at the explicit direction of the clinician or of their own volition); 2) The before-HT phase can include chemotherapy and radiotherapy, especially for the more advanced stages, and some of these interventions (e.g., capecitabine) may affect the early-HT phase. Finally, our ascertainment of treatment discontinuation does not include a reason for discontinuation. For example, disease progression may lead to reasonable discontinuation of treatment. As our review of the informative features, a non-trivial number of patients was discontinued due to metastatic recurrence. To ensure appropriate interventions, such scenarios will need to be distinguished in future work.

Conclusion

Long-term hormonal therapy can reduce breast cancer recurrence and mortality. However, hormonal medication discontinuation is not uncommon among breast cancer patients. As an increasing number of EBC patients take oral medication as their only systemic therapy, early detection of medication discontinuation can help build more effective, targeted interventions. Our investigation demonstrated that structured EMR data has the potential to contribute to the prediction of medication discontinuation behavior. Though the predictive performance was moderate, this study serves as a foundation upon which more powerful predictive models can be learned.

Acknowledgments

This work was supported by the National Cancer Institute of the National Institutes of Health under award numbers R37CA237452 and UL1TR002243.

References


Assessing the Association Between Network-based Provider Communities and Patient Mortality in the Medicare Population with Multiple Chronic Conditions

Yoonyoung Park, Sc.D1, Panagiotis D Karampourniotis, Ph.D1, Issa Sylla, BA1, Gigi Yuen-Reed, PhD2, Amar K Das, MD, PhD1
1IBM Research, Cambridge MA; 2IBM Watson Health, Tampa FL

Abstract

Understanding the complexity of care delivery and care coordination for patients with multiple chronic conditions is challenging. Network analysis can model the relationship between providers and patients to find factors associated with patient mortality. We constructed a network by connecting the providers through shared patients, which was then partitioned into tightly connected communities using a community detection algorithm. After adjusting for patient characteristics, the odds ratio of death for one standard deviation increase in degree centrality ratio between primary care providers (PCPs) and non-PCPs was 0.95 (0.92-0.98). Our result suggest that the centrality of PCPs may be a modifiable factor for improving care delivery. We demonstrated that network analysis can be used to find higher order features associated with health outcomes in addition to patient-level features.

Introduction

Multiple chronic conditions (MCC), defined as having two or more chronic conditions, is one of the most pressing challenges regarding care quality and care delivery among Medicare beneficiaries in the United States (US). A greater number of chronic conditions is associated with increased risk of mortality, disability, poor functional status, poor quality of life, adverse drug events, as well as increased service utilization and greater per capita expense.1–5 In 2010, patients with MCC accounted for 93% of Medicare spending.6 The reported prevalence of MCC among adults over 65 in the US ranges from 45.3%3 to 73.5%7 and the number of patients with MCC in the United States is projected to reach 81 million by 2020 with the current aging trend.8 Despite the significant public health and resource implications, however, MCC research lags behind the majority of healthcare research focused on single disease epidemiology or treatment.9 This is also the case for most existing clinical guidelines that concentrate on one specific condition at a time.10 To address the knowledge gap, Federal agencies have promoted research efforts including a strategic framework initiated by the US Department of Health and Human Services (HHS)11 MCC Research Network funded by the Agency for Healthcare Research and Quality (AHRQ).6

Managing multiple conditions often requires care coordination between multiple providers, and this complexity necessitates consideration of provider relationships and care organization structure in research approaches. Network analysis is garnering interests in healthcare domain for its capability to model and quantify such complex relationships between different entities in healthcare.12–21 Recently, the availability of large healthcare data such as administrative claims data coupled with network analysis has enabled researchers to investigate key aspects of healthcare including provider collaboration, referral patterns, prescription patterns, and workplace interactions.22–25 Prior work has shown associations between network-based measures of collaboration or coordination and quality of care, patient outcomes, and cost in different care delivery settings.15,17–19,26

To this date, studies utilizing network analysis have mainly focused on general patient population, not on specific disease groups. Other studies considered limited geographic regions due to data availability. Understanding the provider network and identifying provider collaboration clusters at a national scale can provide important insights regarding the care delivery and management of MCC. We applied network analysis to examine the provider network delivering health services to Medicare beneficiaries with MCC and investigated the characteristics of the provider communities in association with patient characteristics and mortality. We use the term ‘community’ in this study to refer to the naturally forming clusters of tightly connected providers identified with an algorithm, as opposed to formally defined provider communities through institutional or insurance affiliations.
Methods

Data and Patient Population

We used inpatient and outpatient claims (2012-2016) from Medicare Standard Analytic Files (SAF), nation-wide administrative claims data of Medicare beneficiaries. CMS makes this deidentified data set available to researchers with a signed Data Use Agreement and adherence to the data policies. While non-institutional providers or practices are an integral part of health service, we did not have access to the full carrier file. As a result, our provider network comprises providers engaged in institution-based services to Medicare patients.

The patient population of interest was Medicare patients with two or more chronic conditions. Eligible subjects were 65 or older and continuously enrolled in both Medicare Part A and B with at least one claim during 2012-2013. If a patient died during 2012-2013 after being continuously enrolled, we included the patient in the network construction. We restricted the cohort to the majority (> 75%) of patients who had become eligible for Medicare by Old Age and Survival criteria (i.e. age over 65) and excluded those who became entitled for Medicare through ESRD or disability. This allowed focused analysis on general MCC population, not on subpopulations who likely have different utilization patterns and prognoses over time. Subjects with Medicare Advantage benefit or Medicaid dual eligibility at any point during this time period were also excluded since such subjects may have incomplete capture of their data. Lastly, subjects with MCC during 2012-2013 were identified based on definitions available at CMS Chronic Conditions Data Warehouse for the 27 selected chronic conditions.

Provider Network Construction

We constructed a provider network based on the outpatient claims from the eligible MCC patients. We did not consider inpatient provider encounters in network construction under the assumption that the frequency and nature of provider encounters during inpatient admissions would be different from those during outpatient visits and that outpatient encounters better reflect patient and provider relationships as they tend to be more directed by patient choice or provider referrals compared to inpatient encounters. From the encounter-level claims, we obtained provider identifiers (National Provider Identifier; NPI) associated with each claim. Claims missing provider identifier or associated with non-physician professionals except physician assistants and nurse practitioners were excluded. As a result, patients without a claim associated with a physician provider identifier were also removed.

The network was constructed based on patient sharing, similar to the method previously reported. Briefly, we first created a patient-provider bipartite network, linking each patient to every provider associated with the patient’s claims during 2012-2013. We then projected this bipartite network to create a provider-provider network where each provider became a node. In this network, the number of shared patients between a pair of providers became the weight of the edge between the two providers. An edge can represent either an actually co-treated patient or a linkage through a patient who switched between different providers or move from one place to another. We removed the providers who shared zero patients with other providers (isolated nodes) from the network, and subsequently excluded their patients from further analysis.

Provider Community Detection

Considering the geographical scale and diversity in healthcare, we hypothesized that the nationwide network can be partitioned into different clusters or ‘communities’ of providers who work more closely together than with others. Using Clauset-Newman-Moore algorithm based on modularity maximization, we detected non-overlapping provider communities (i.e. each provider belonging to a single community) within the network. In this study, modularity of a provider network partitioned into different communities represents how likely providers within the same communities are more densely connected to each other by shared patients than they are to providers across different communities. Mathematically, modularity can be described as the difference between the fraction of within-community edges in the observed network and in a network where edges are placed between nodes (providers) at random, while maintaining the degree of each node. We chose this algorithm based on performance, scalability and computational efficiency, since the size of the provider network prohibited us from using more computationally expensive algorithms.
Predominant Provider of Care Identification

The way we constructed the provider network allows patients to belong to multiple communities if they are treated by multiple providers. To define characteristics for each provider community and their associated patients, we assigned each patient to a single predominant provider of care and that provider’s community respectively. In our study, predominant provider of care is with whom the patient had the most visits during the 2012-2013 period, following the definition introduced in a previous study where the authors used the same Medicare fee-for-service claims to assign patients to their predominant provider of care. When the greatest number of visits was equal amongst providers, the provider with the greatest time span between the first and last visits was chosen. If there was only one visit for tied providers, then the most recently visited provider was chosen. We added a final step to break a tie using the total charge per provider, choosing the one with maximum charge.

Analysis

We examined the provider network for variations in community level properties, focusing on the ones that can explain the cohesiveness and tightness of a community to understand the impact of provider relationships on patient outcomes (Table 1). Broadly speaking, internal density and conductance of a community quantify how strong the connectivity within and across communities is, respectively. Lower density implies proportionately less provider connections within a community, while lower conductance implies proportionately higher provider connections within the community compared to across communities. Centrality measures how ‘central’ a provider node is within a network, and we focused on degree centrality and betweenness centrality measures.

We examined provider and patient characteristics such as demographics, provider specialty including proportion of primary care providers (PCPs, defined as providers in General Practice, Family Practice, Geriatric Medicine, Internal Medicine, Pediatric Medicine, Preventive Medicine, Obstetrics & Gynecology, or Nurse Practitioners or Physician Assistants), and prevalence and distribution of chronic condition diagnoses among patients. Provider specialty information was available through NPI Downloadable File, a publicly available dataset through CMS. For each community we obtained patient mortality using the death record in Medicare data. In addition, we hypothesized that a provider community formed by patient sharing will have an identifiable geographical boundary because patients typically receive their care nearby. To inspect this hypothesis, we graphed the distribution of provider nodes in each community on a map of the US. Only the subset of communities with at least 1000 provider nodes were included in the community-level analyses, since the remaining communities had 4 or less provider nodes.

We used generalized linear models (GLMs) with logit link to explain the variation in patient mortality, adjusting for the correlation of data among patients in the same community using generalized estimating equations (GEE). We chose GEE approach over mixed model approach because the objective of this study was to understand population effects of network features, not individual patient variability. Also, GEE is a nonparametric method and more robust to covariance structure misspecification. The patient-level features include age over 84 (vs. 84 or less), gender, race, state, urban (vs. rural), and number of chronic conditions. The network features include internal density, conductance, the ratio of degree centrality between PCPs and non-PCPs, and the ratio of betweenness centrality between PCPs and non-PCPs. The degree (betweenness) centrality of PCPs in a community is the average degree (betweenness) centrality of each PCP node in the community. We only include relative centrality measures of PCPs to non-PCPs, since centrality measures are dependent on network size. A p-value less than 0.05 was considered statistically significant. We used Python 3.6 for analysis, including networkx, igraph, scikit learn, and statsmodels packages.

Results

In 2012 and 2013, there were 53,428,800 and 55,087,354 unique Medicare beneficiaries, respectively. After restricting to patients who met the Medicare entitlement criteria and were continuously enrolled without Medicare Advantage or dual coverage for both years, 16,951,784 patients remained in the cohort among whom approximately half (8,477,385) had two or more chronic conditions. After requiring each patient to have at least one claim associated with a physician, physician assistant, or nurse practitioner and excluding the providers who share zero patients with other providers, 6,916,918 patients remained in the final cohort who received health services from 330,243 distinct providers.
Network Feature | Definition |
--- | --- |
Node | In this study, a node in the network represents a single provider, either a physician, a physician assistant, or a nurse practitioner. National Provider Index serves as the unique node identifier. |
Edge | An edge between two provider nodes indicates a single shared (i.e. co-treated) patient for the two providers. When there are multiple shared patients, the total number of shared patients becomes the edge weight. |
Provider Network | The provider network in this study is a weighted network of nation-wide Medicare providers connected through shared patient edges. |
Community | Communities in this study represent non-overlapping clusters of providers within the provider network, detected through modularity maximization algorithm. Providers in a given community are expected to be more densely connected (i.e. have greater number of shared patients) compared to the providers in other communities. |
Degree (Degree Centrality) | Degree of a node in the provider network indicates the number of other providers that the given provider has shared patients with. If a provider node has a large degree (centrality) value, it means that the particular provider creates many linkages through shared patients with other providers by playing a central role in the overall network. |
Internal Density | In this study, internal density is defined as the ratio of number of unweighted within-community edges (i.e. edges that do not leave the community) to total number of possible edges between the provider nodes in the community. If a community has high internal density, it can possibly imply increased communication and care coordination within the same community providers. |
Conductance | In this study, conductance is defined as the ratio between the sum of inter-community edge weights and the sum of all edge weights for the provider nodes in a community. If a community has high conductance value, it can mean that the providers in that community share patients and information between themselves much more than they do with providers outside of that tightly linked community, possibly leading to communication efficiency. |
Betweenness Centrality | Betweenness centrality of a node is defined as the number of shortest paths between all pairs of nodes in the network that pass through that node. It quantifies the influence of a provider on the flow of information or flow of care in the network. If a provider node has high betweenness centrality value, that provider acts as a ‘mutual neighbor’ between many other providers. Such a provider can influence the network information flow because it acts as a shortcut node between many other providers. |

Table 1: Definitions of network features in this study

Overall, patients were 56.6% female, 91.9% White, and 25.3% were 85 or older (Table 2). The top five states in terms of patient number were Florida, Illinois, Texas, California, and Michigan. On average, patients had 4.0 distinct chronic conditions (standard deviation [SD] 1.9) in 2012-2013. The average number of chronic conditions monotonically increased with age category. For those between 65 and 69 the average was 3.4, compared to 4.5 for those over age 89. The most prevalent chronic conditions were hypertension (69.2%), hyperlipidemia (49.4%), and ischemic heart disease (41.2%).

In the provider network with 330,243 nodes, the average degree was 105.6 (SD 131.3), meaning that providers were connected with 106 other providers on average. A total of 130 communities were detected in this network, with the 28 largest communities having 1000 or more providers (Table 3). When plotted on a map, a majority of the communities showed state-level geographical concentration (the 3 largest communities shown in Figure 1). After assigning patients to their predominant providers of care and the providers’ respective communities, the largest community had 26,459 providers providing care to 382,528 patients. Female patients ranged between 52.7% and 60.7%, and white patients ranged between 32.7% and 97.7% across the communities. The average number of chronic conditions per patient ranged from 3.7 to 4.2. The PCPs comprised 15.4% to 42.2% of providers in each community, showing variability.
### Demographic

<table>
<thead>
<tr>
<th>Age Category</th>
<th>N</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>[65, 70)</td>
<td>939,866</td>
<td>13.60%</td>
</tr>
<tr>
<td>[70, 75)</td>
<td>1,280,343</td>
<td>18.50%</td>
</tr>
<tr>
<td>[75, 80)</td>
<td>1,534,085</td>
<td>22.20%</td>
</tr>
<tr>
<td>[80, 85)</td>
<td>1,414,863</td>
<td>20.50%</td>
</tr>
<tr>
<td>[85 or above)</td>
<td>1,747,761</td>
<td>25.30%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>N</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>3,917,593</td>
<td>56.60%</td>
</tr>
<tr>
<td>White</td>
<td>6,353,255</td>
<td>91.90%</td>
</tr>
</tbody>
</table>

### Top 5 States (patient number)

- Florida: 429,928 (6.20%)
- Illinois: 426,610 (6.20%)
- Texas: 404,103 (5.80%)
- California: 397,519 (5.70%)
- Michigan: 344,398 (5.00%)

### Chronic Conditions

#### Ten Most Prevalent Conditions

- Hypertension: 4,787,945 (69.20%)
- Hyperlipidemia: 3,419,358 (49.40%)
- Ischemic Heart Disease: 2,848,119 (41.20%)
- Anemia: 2,176,692 (31.50%)
- Diabetes: 2,127,330 (30.80%)
- Chronic Kidney Disease: 1,732,283 (25.00%)
- Heart Failure: 1,517,884 (21.90%)
- Acquired Hypothyroidism: 1,190,985 (17.20%)
- Rheumatoid Arthritis/Osteoarthritis: 1,002,879 (14.50%)
- Alzheimer’s Disease and Related Disorders/Senile Dementia: 968,008 (14.00%)

<table>
<thead>
<tr>
<th>N Chronic Condition</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Per patient</td>
<td>4</td>
<td>1.9</td>
</tr>
</tbody>
</table>

#### Per Each Age Category

<table>
<thead>
<tr>
<th>Age Category</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>[65, 70)</td>
<td>3.4</td>
<td>1.5</td>
</tr>
<tr>
<td>[70, 75)</td>
<td>3.6</td>
<td>1.6</td>
</tr>
<tr>
<td>[75, 80)</td>
<td>4</td>
<td>1.9</td>
</tr>
<tr>
<td>[80, 85)</td>
<td>4.3</td>
<td>2</td>
</tr>
<tr>
<td>[85 or above)</td>
<td>4.5</td>
<td>2</td>
</tr>
</tbody>
</table>

### Table 2: Characteristics of patients with multiple chronic conditions in Medicare in 2012-2013

| N: Number; SD: Standard deviation |

Based on univariate analyses, all of the patient features were significantly associated with mortality but not the network features. In the multivariate GLM controlling for the correlation (Table 4), male gender, white race, older age (> 84), and higher chronic comorbidity level were significantly associated with higher mortality. Adjusting for all patient features, internal density, betweenness centrality ratio, and conductance were not associated with patient mortality. However, degree centrality ratio between PCPs and non-PCPs was negatively associated with mortality (odds ratio 0.95, 95% confidence interval 0.92-0.98), meaning that controlling for other factors a patient in a community with high PCP to non-PCP degree centrality ratio has lower probability of dying compared to a patient in a community with lower ratio.
Discussion

Using network analysis, we constructed national Medicare provider network for patients with MCC and partitioned the provider network into more tightly connected communities. Multivariate regression results suggest that in addition to patient demographic and clinical factors, PCP degree centrality of the patient’s community relative to that of specialists is weakly associated with patient mortality. The detected provider communities exhibited geographical boundaries, implying that provider interactions through shared patients depend on physical distance.

The provider network characteristics we observed meet our expectations based on literature. Pham et al. examined patient-sharing network using one year (2005) of Medicare claims data and found that the average degree of PCPs in the network was 229 (Interquartile range 125-340). The authors considered physicians responding to a survey and excluded physicians contributing less than 20 hours per week on direct patient care, which may explain the greater average degree compared to our study (105.6) as those physicians are more likely to be actively engaged in care for larger number of patients. While the magnitudes were small, the positive associations we observed for older age and average number of chronic conditions with respect to mortality can be expected based on clinical knowledge. The positive association between proportion of white patients and mortality seems to contradict the reported health disparity among minorities and requires further exploration as it can be a proxy of another explanatory factor. The large number of providers in communities connected by and treating the same group of shared patients may have implication in the suboptimality of care that the MCC patients receive. It has been reported that MCC patients experience difficulty in care coordination and reconciliation across multiple providers.

Associations between provider network features and healthcare utilization or patient outcomes have been reported in a number of previous works. For example, a recent study by Geissler et al. showed that measures of both local and global provider connectivity are associated with healthcare cost and utilization. Our contribution, in addition to creating a provider network specific for MCC patients, is showing the association between relative PCP degree centrality and patient mortality. Higher ratio of PCP degree centrality to non-PCP degree centrality implies a PCP-centric provider community in which PCPs potentially play a significant role in communication and care coordination between providers. The observed 5% decrease in the odds of death associated with one standard deviation increase in degree centrality ratio is significant considering the number and overall mortality of the Medicare population. Additional studies are needed to examine how changes in PCP centrality ratio is associated with patient outcomes and the mechanisms behind it.

The provider communities examined in the study were detected through modularity maximization, which is a data-driven mathematical algorithm. These communities are unlike those formed through organizational membership or health plans. Comparing organizational membership to algorithmic community membership would be an interesting topic for future work. While ascertaining community membership based on algorithmic approach is difficult without ground truth data, empirical evidence such as geographical distribution observed in this study indicates that the algorithmic communities closely represent real-life communities associated with physical distance. Variation in care associated with geography has been observed among Medicare patients with complex conditions, and the

![Figure 1: Geographic distribution of provider nodes in the three largest communities. Community A shows distribution of providers in both Northeast and Florida region. This may reflect the ‘snowbird’ population, a group of people who migrate to warmer regions during the winter months.](image)
Table 3: Characteristics (%) of communities with >1000 nodes detected from the provider network among Medicare beneficiaries with multiple chronic conditions in 2012-2013

<table>
<thead>
<tr>
<th>Comm</th>
<th>N Prov</th>
<th>N Pt</th>
<th>Pt &gt;84</th>
<th>Female Pt</th>
<th>White Pt</th>
<th>CC</th>
<th>Female Prov</th>
<th>PCP</th>
<th>Mortality</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>26,459</td>
<td>382,528</td>
<td>30.8</td>
<td>56.2</td>
<td>88.6</td>
<td>3.9</td>
<td>33.7</td>
<td>16.7</td>
<td>23.7</td>
</tr>
<tr>
<td>B</td>
<td>20,636</td>
<td>390,501</td>
<td>29.5</td>
<td>58</td>
<td>93.6</td>
<td>4.1</td>
<td>37.8</td>
<td>26.4</td>
<td>23.3</td>
</tr>
<tr>
<td>C</td>
<td>19,222</td>
<td>388,982</td>
<td>26.1</td>
<td>56.1</td>
<td>96.7</td>
<td>3.9</td>
<td>43.7</td>
<td>31.1</td>
<td>19.9</td>
</tr>
<tr>
<td>D</td>
<td>18,852</td>
<td>480,083</td>
<td>25.6</td>
<td>57.4</td>
<td>91</td>
<td>4.1</td>
<td>36.5</td>
<td>22.6</td>
<td>21</td>
</tr>
<tr>
<td>E</td>
<td>18,785</td>
<td>416,203</td>
<td>23.0</td>
<td>57.1</td>
<td>91.2</td>
<td>4.1</td>
<td>29.7</td>
<td>19.2</td>
<td>23.3</td>
</tr>
<tr>
<td>F</td>
<td>18,496</td>
<td>488,074</td>
<td>23.1</td>
<td>55.2</td>
<td>92.3</td>
<td>4</td>
<td>29.6</td>
<td>23.3</td>
<td>22.4</td>
</tr>
<tr>
<td>G</td>
<td>16,180</td>
<td>359,534</td>
<td>21.2</td>
<td>56.9</td>
<td>87</td>
<td>3.9</td>
<td>36.8</td>
<td>31.6</td>
<td>21.5</td>
</tr>
<tr>
<td>H</td>
<td>14,479</td>
<td>268,009</td>
<td>27.4</td>
<td>54.9</td>
<td>91.2</td>
<td>3.9</td>
<td>34</td>
<td>22.3</td>
<td>20.8</td>
</tr>
<tr>
<td>I</td>
<td>14,230</td>
<td>307,953</td>
<td>23.8</td>
<td>57.2</td>
<td>94.1</td>
<td>4.1</td>
<td>35.7</td>
<td>26.2</td>
<td>20.7</td>
</tr>
<tr>
<td>J</td>
<td>14,169</td>
<td>249,451</td>
<td>25.6</td>
<td>55.1</td>
<td>95.1</td>
<td>3.9</td>
<td>39.2</td>
<td>33.9</td>
<td>21.4</td>
</tr>
<tr>
<td>K</td>
<td>14,022</td>
<td>174,557</td>
<td>29.9</td>
<td>54.5</td>
<td>87.9</td>
<td>3.9</td>
<td>27.5</td>
<td>15.4</td>
<td>22.8</td>
</tr>
<tr>
<td>L</td>
<td>13,830</td>
<td>317,049</td>
<td>27.8</td>
<td>57.1</td>
<td>97.7</td>
<td>4.1</td>
<td>40.5</td>
<td>35</td>
<td>21.1</td>
</tr>
<tr>
<td>M</td>
<td>12,271</td>
<td>285,149</td>
<td>25.7</td>
<td>56.2</td>
<td>95.7</td>
<td>4</td>
<td>38.9</td>
<td>33.1</td>
<td>21.7</td>
</tr>
<tr>
<td>N</td>
<td>12,161</td>
<td>275,689</td>
<td>22.1</td>
<td>56</td>
<td>91</td>
<td>3.9</td>
<td>32.3</td>
<td>23.4</td>
<td>22.8</td>
</tr>
<tr>
<td>O</td>
<td>12,119</td>
<td>331,291</td>
<td>25.3</td>
<td>57.3</td>
<td>90.8</td>
<td>4.2</td>
<td>34.5</td>
<td>24.3</td>
<td>20.9</td>
</tr>
<tr>
<td>P</td>
<td>12,088</td>
<td>238,073</td>
<td>25.9</td>
<td>56.9</td>
<td>85.3</td>
<td>4</td>
<td>39.6</td>
<td>25.2</td>
<td>22.5</td>
</tr>
<tr>
<td>Q</td>
<td>11,074</td>
<td>267,108</td>
<td>23.6</td>
<td>58</td>
<td>95</td>
<td>4.1</td>
<td>35.4</td>
<td>26.2</td>
<td>22.1</td>
</tr>
<tr>
<td>R</td>
<td>10,513</td>
<td>330,848</td>
<td>24.5</td>
<td>58.3</td>
<td>95.5</td>
<td>4.1</td>
<td>37.1</td>
<td>30</td>
<td>21.6</td>
</tr>
<tr>
<td>S</td>
<td>9,078</td>
<td>212,958</td>
<td>23.6</td>
<td>57.3</td>
<td>86.3</td>
<td>3.9</td>
<td>40.6</td>
<td>30.4</td>
<td>21.2</td>
</tr>
<tr>
<td>T</td>
<td>7,984</td>
<td>122,809</td>
<td>29.4</td>
<td>57.8</td>
<td>95.7</td>
<td>4</td>
<td>45.6</td>
<td>42.2</td>
<td>22.5</td>
</tr>
<tr>
<td>U</td>
<td>7,467</td>
<td>152,641</td>
<td>21.4</td>
<td>58.2</td>
<td>85.1</td>
<td>4.1</td>
<td>35</td>
<td>28</td>
<td>21.8</td>
</tr>
<tr>
<td>V</td>
<td>6,596</td>
<td>119,409</td>
<td>22.9</td>
<td>55.2</td>
<td>94.8</td>
<td>3.8</td>
<td>31.4</td>
<td>35.2</td>
<td>21.4</td>
</tr>
<tr>
<td>W</td>
<td>5,044</td>
<td>69,267</td>
<td>21.9</td>
<td>52.7</td>
<td>90.9</td>
<td>3.8</td>
<td>35.6</td>
<td>27.9</td>
<td>21.8</td>
</tr>
<tr>
<td>X</td>
<td>4,797</td>
<td>76,193</td>
<td>29.1</td>
<td>54.9</td>
<td>95.1</td>
<td>3.8</td>
<td>39.5</td>
<td>29.1</td>
<td>22.5</td>
</tr>
<tr>
<td>Y</td>
<td>3,701</td>
<td>107,003</td>
<td>22.3</td>
<td>56.6</td>
<td>87.3</td>
<td>4.1</td>
<td>35.1</td>
<td>32.2</td>
<td>22.3</td>
</tr>
<tr>
<td>Z</td>
<td>2,444</td>
<td>81,109</td>
<td>21.5</td>
<td>57</td>
<td>94.1</td>
<td>4</td>
<td>30</td>
<td>23.3</td>
<td>21.4</td>
</tr>
<tr>
<td>AA</td>
<td>2,314</td>
<td>8,166</td>
<td>23.8</td>
<td>60.7</td>
<td>75.1</td>
<td>3.7</td>
<td>27.5</td>
<td>35.5</td>
<td>19</td>
</tr>
<tr>
<td>BB</td>
<td>1,008</td>
<td>16,147</td>
<td>27.6</td>
<td>53.9</td>
<td>32.7</td>
<td>3.9</td>
<td>29.7</td>
<td>19.5</td>
<td>19.1</td>
</tr>
<tr>
<td>Min</td>
<td>1,008</td>
<td>8,166</td>
<td>21.2</td>
<td>52.7</td>
<td>32.7</td>
<td>3.7</td>
<td>27.5</td>
<td>15.4</td>
<td>19</td>
</tr>
<tr>
<td>Max</td>
<td>26,459</td>
<td>488,074</td>
<td>30.8</td>
<td>60.7</td>
<td>97.7</td>
<td>4.2</td>
<td>45.6</td>
<td>42.2</td>
<td>23.7</td>
</tr>
<tr>
<td>Avg</td>
<td>11,786.4</td>
<td>247,028.0</td>
<td>25.2</td>
<td>56.5</td>
<td>89.2</td>
<td>4</td>
<td>35.4</td>
<td>27.5</td>
<td>21.7</td>
</tr>
<tr>
<td>SD</td>
<td>6,275.0</td>
<td>136,633.6</td>
<td>2.9</td>
<td>1.6</td>
<td>12.1</td>
<td>0.1</td>
<td>4.6</td>
<td>6.2</td>
<td>1.1</td>
</tr>
</tbody>
</table>

community-level differences observed in our study may partially reflect such variability due to geography.

Our study has multiple strengths. To our best knowledge, this is the first study that examines the network characteristics of provider communities and its effect on the mortality of patients with MCC. As this group of patients drives the cost of Medicare, our study adds valuable insights to the small body of evidence for understanding the care provided to MCC patients. We used multiyear Medicare data that represents all beneficiaries across the nation in a continuous manner, unlike commercial insurance data in which churning is frequent. Provided that the observed associations in our study are indeed causal, we can draw two policy-level public health implications. First, empirical provider network analysis offers important insights on how healthcare delivery through the network can influence patient outcomes and on factors associated with the outcomes that are not observable in conventional patient-level analysis. Second, promoting the central role of PCPs in a given network may be a modifiable factor that can have beneficial impact on the outcomes of patients with MCC, highlighting the importance of primary care in accessible and affordable healthcare.
Table 4: Result from multivariate generalized linear model for patient mortality in 2014-2016, adjusted for correlation within communities

<table>
<thead>
<tr>
<th></th>
<th>OR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>0.73</td>
<td>0.72 0.74</td>
</tr>
<tr>
<td>White</td>
<td>1.13</td>
<td>1.10 1.16</td>
</tr>
<tr>
<td>Age over 84</td>
<td>4.14</td>
<td>4.05 4.23</td>
</tr>
<tr>
<td>Number of chronic conditions</td>
<td>1.53</td>
<td>1.52 1.55</td>
</tr>
<tr>
<td>Urban area</td>
<td>0.98</td>
<td>0.96 1.00</td>
</tr>
<tr>
<td>Internal density</td>
<td>0.98</td>
<td>0.96 1.00</td>
</tr>
<tr>
<td>Degree centrality ratio*</td>
<td>0.95</td>
<td>0.92 0.98</td>
</tr>
<tr>
<td>Betweenness centrality ratio*</td>
<td>1.03</td>
<td>1.00 1.06</td>
</tr>
<tr>
<td>Conductance</td>
<td>1.01</td>
<td>0.99 1.02</td>
</tr>
<tr>
<td>State** (53)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Ratio between PCPs and non-PCPs
* The table does not show OR values for the State variables with 53 categories which was included in the model

The limitations of our study include the fact that we did not have access to Medicare carrier claims data, resulting in a network that represents providers affiliated with institutional care only and not non-institutional providers or practices. Because of this, we are not able to assess if and how the network assignment would change once office-based providers are included. As such, our study results are less applicable for areas where healthcare is mostly provided by small, non-institutional providers, which is likely to be rural parts of the US rather than urban areas. Because medical history is defined by institutional claims only, we may have some misclassification for the patients whose diagnosis are recorded only in the carrier claims. We only examined the network for Medicare beneficiaries and excluded those with Medicaid or Medicare Advantage benefits. There can be significant difference between different payers and hence interpretation of the study results should be restricted to the definition of our study cohort, Medicare patients. Part of the provider connections identified through claims are ‘false positives’ with regard to the actual care coordination, meaning that although some providers appear to be connected by the same patient they are not participating in shared care. Lastly, we defined network based on 2012-2013 data, but the patient-provider relationship could have changed over time. This means that some degree of misclassification of patient-provider relationship might be present, further limiting our ability to interpret the associational result in a causal manner. Future exploration will include addressing the sensitivity of the provider connection definitions and conducting region-specific analyses or temporal analysis to apply and compare different algorithms and their performances.

References


ECG and SpO₂ Signal-Based Real-Time Sleep Apnea Detection Using Feed-Forward Artificial Neural Network

Tanmoy Paul¹2, Omiya Hassan¹, Khuder Alaboud²3, Humayera Islam²3, Md Kamruz Zaman Rana²3, Syed K. Islam, PhD¹, Abu S. M. Mosa, PhD¹,2,3,4*¹
¹Department of Electrical Engineering and Computer Science; ²NextGen Biomedical Informatics Center; ³Institute of Data Science and Informatics; ⁴Department of Health Management and Informatics, University of Missouri, Columbia, MO, USA
*Corresponding Author: mosaa@health.missouri.edu

Abstract
Sleep apnea (SA) is a common sleep disorder characterized by respiratory disturbance during sleep. Polysomnography (PSG) is the gold standard for apnea diagnosis, but it is time-consuming, expensive, and requires manual scoring. As an alternative to PSG, we investigated a real-time SA detection system using oxygen saturation level (SpO₂) and electrocardiogram (ECG) signals individually as well as a combination of both. A series of R-R intervals were derived from the raw ECG data and a feed-forward deep artificial neural network is employed for the detection of SA. Three different models were built using 1-minute-long sequences of SpO₂ and R-R interval signals. The 10-fold cross-validation result showed that the SpO₂-based model performed better than the ECG-based model with an accuracy of 90.78 ± 10.12% and 80.04 ± 7.7%, respectively. Once combined, these two signals complemented each other and resulted in a better model with an accuracy of 91.83 ± 1.51%.

Introduction
Sleep apnea (SA) is one of the most common forms of sleep disorders which is characterized by respiratory disturbance during sleep and affects about 2%-5% of the total adult population and more than 30% of the elderly population in the US [1]. There are mainly three types of sleep apnea: obstructive sleep apnea (OSA), central sleep apnea (CSA), and complex sleep apnea syndrome. OSA is characterized by the repeated pharyngeal collapse that causes shortness (hypopnea) or cessation (apnea) of breathing during sleep [2]. Apnea is often defined as the cessation of breathing for at least 10 seconds and hypopnea is defined by a significant reduction of airflow for a minimum period of 10 seconds accompanied by either 4% desaturation of blood oxygen level or neurological arousal [3]–[5]. CSA is characterized by recurrent apneic events accompanied by a lack of respiratory effort as the brain does not transmit any stimulus to the breathing muscles [6]. Such disturbances often cause arousal from sleep which results in excessive daytime sleepiness and fatigue. Complex Sleep Apnea Syndrome occurs when a patient has both OSA and CSA. Severe forms of SA can lead to cardiovascular dysfunction, ischemic heart disease, and stroke [3]. It is linked with significant cardiovascular morbidity and is one of the major causes of hypertension [7]. Therefore, accurate and timely diagnosis and treatment of sleep apnea are essential for risk minimization.

Currently, polysomnography (PSG) is considered to be the gold standard for apnea diagnosis, which requires a subject to spend a night or two in a sleep laboratory under the supervision of sleep specialists. Usually, multiple sensors and wires are attached to the subject’s body to record various physiological signals. These signals may include brain waves (electroencephalogram, or EEG), eye movements (electrooculogram, or EOG), blood oxygen level (SpO₂), heart rate, and rhythm (electrocardiogram, or ECG). The final diagnosis requires analysis of the recorded data by the specialists. The severity of the SA is addressed by the apnea-hypopnea index (AHI), which is defined as the number of apnea-hypopnea events per hour over the entire sleep period. The process is time-consuming, expensive, and the attachment of multiple sensors and wires causes discomfort to the subjects. As a result, the researchers have made extensive efforts to establish alternatives to PSG with simpler schemes and faster decision-making capabilities.

In the past several years, multiple methods to detect SA have been proposed. These methods differ from one another in terms of the classification methods and the physiological signals used to detect SA. Multiple rule-based techniques have been developed for SA detection [8], [9]. Recently, the machine learning-based approach has become a popular
technique for SA detection and monitoring and researchers have a number of machine learning models at their disposal. Models such as support vector machine (SVM), K-nearest neighbor (KNN), and decision tree have been used successfully in several studies [10]–[14]. There are reports of SA detection systems that used a combination of multiple classifiers as well [3], [15]. Recently, deep learning techniques, such as convolutional neural network (CNN) and recursive neural network (RNN), have emerged as very efficient techniques of SA detection [16]–[19]. Among all the physiological signals used in the above methods, the most commonly used ones are SpO2 and ECG. Many models require appropriate feature extraction from these two signals for efficient performance. However, CNN and RNN based models usually do not require any additional feature extraction from the signals, which makes them extremely effective in real-time SA detection systems with minimum signal analysis. For these networks to perform well without manual feature extraction, complex architecture consisting of feature extraction layer, convolutional layer, pooling layer, recurrent layer, etc. are required. The lightweight deep artificial neural network (ANN), without any of those special layers, has been deployed in multiple systems [20], [21]. But the performance of such a network is yet to be investigated if no feature from the signal is provided.

In this study, we investigated the SpO2 and ECG signals for a real-time sleep apnea detection system. A feed-forward artificial neural network is used to evaluate the performance of the signals for detection of SA. Instead of using heavyweight convolutional neural networks, the proposed scheme demonstrates the use of a simple feed-forward neural network for detecting sleep apnea in real-time without any manual feature extraction. Thus, it enables much simpler and faster implementation of apnea detection system. In addition, the proposed scheme works with two different physiological signals and shows that the dual-channel technique of apnea detection performs better than the single-channel technique and thus substantiates the efforts of further exploration of dual-channel approaches.

**Methods**

**Dataset**

The proposed scheme employs PhysioNet Apnea-ECG dataset [22], [23]. There are a total of 70 records in this dataset. The entire dataset is divided into a training set of 35 records and a testing set of 35 records. The lengths of the records vary from 7 hours to 10 hours. Each of the recordings includes a digitized ECG signal, a set of apnea annotations derived by human experts, and a set of machine-generated QRS annotations. In addition to the ECG signals, 8 of these recordings have chest and abdominal respiratory effort signal, oronasal airflow signal, and oxygen saturation level signal, SpO2. Since the study aimed to build the ANN models by using ECG and SpO2 signals both individually and as a combination, only those 8 recordings in our analysis were used. The signals were sampled at the rate of 100 samples per second. In each recording, the annotation was placed at the start of every minute, followed by a one-minute interval. The annotation ‘A’ signifies that an apneic event was in progress at the beginning of the associated minute. The annotation ‘N’ means that no apneic event was in progress at the beginning of the associated minute. **Figure 1** illustrates the applied annotation criteria in the Apnea-ECG dataset.

![Figure 1](image1.png)

**Figure 1.** Annotation criterion of Apnea-ECG dataset. Elapsed time is indicated by the distance from the left edge. ‘~’s denote the apneic periods, ‘|’s denote the time of apnea annotation (0, 60, 120, …seconds), and apneic and non-apneic intervals are marked by ‘A’ and ‘N’ annotations, respectively.

**Data Segmentation**

Since apnea annotations were provided for each one-minute interval, both the SpO2 and the ECG signals were divided into segments with a duration of 1 minute. **Figure 1** shows that apnea annotation depends only on the presence or absence of apneic events at the beginning of the associated minute. Such annotation scheme may be misleading, as it can be seen in the 6th interval in **Figure 1** where the interval is annotated as ‘N’, although the apneic event was persistent for most of the 1-minute duration. Similarly, the 7th interval is annotated as ‘A’ despite the apneic event lasting for a short time during the associated 1-minute period. To overcome this problem, considered only the 1st 30 seconds of each of the 1-minute segments were considered and the rest were discarded. It made sure that for any ‘N’-annotated 1-minute segment if there were sample points with apneic events, most of these were discarded. Similarly, if there were non-apneic sample points in an ‘A’-annotated segment, most of those were removed. Since an apneic event is marked by abnormal breathing and persistent SpO2 signal for at least 10 seconds, the decision to consider only 30 seconds of each interval is justified as it is more than the required minimum duration.
SpO₂ Signal Processing

The first step of SpO₂ signal processing was to identify and remove all the artifacts. Any SpO₂ value less than 50% is not physiologically possible and was marked as an artifact. Moreover, all changes of SpO₂ values greater than 4% were marked as artifacts [7]. Once those artifacts were removed, the signal was resampled at 1 Hz by using a simple moving average filter. After resampling, only 30 sample points from the 1st 30 seconds of each of the 1-minute segments were retained. Before resampling, if the total number of artifacts in the 30-seconds interval was greater than 10% of the total number of sample points in that interval, the entire associated 1-minute segment was discarded. It was done to make sure that none of the intervals with significant information loss, caused by the removal of sample points marked as artifacts, were used to train the ANN model.

ECG Signal Processing

An ECG graph consists of P wave, QRS complex, and T wave. R-peak is the maximum amplitude in the R wave. The machine-generated QRS annotations associated with each recording were used to detect the R-peaks from the ECG signal within the first 30-seconds interval of each 1-minute segment. The QRS detector used in this database is based on the study by Zong et al. [24]. The R-R interval is simply the time interval between two successive R-peaks. A sliding window technique was implemented to remove the ectopic sample points from the R-R interval series. The window length was 5 and any R-R interval value larger than 20% of the average value of all the R-R interval values within the window was marked as ectopic beats and was removed [10]. Following the removal of the artifacts, the entire associated 1-minute interval was discarded if there were less than 30 sample points from the 30-seconds interval. If there were more than 30 R-R interval points, only the first 30 of these points were considered to be consistent with the number of points in each input vector derived from the SpO₂ signal for training the ANN model. The flow diagram in Figure 2 shows all the steps involved in the data processing sequentially.

![Flow Diagram](image)

**Figure 2:** Illustration of all the data processing steps.

Deep Neural Network

A deep neural network consists of multiple layers of nonlinear processing units where the output of one layer serves as the input of the next one. It is called a feed-forward network since the flow of information is always from the input-end to the output-end [25]. In this study, three deep ANN models were built, one for each SpO₂ and ECG signals and one for a combination the two signals. The network architecture is illustrated in Figure 3. Each model had 5 hidden layers in addition to an input and an output layer. The hidden layers had 100, 50, 25, 10, and 5 neurons respectively. The output layer had only one neuron with an output of 0 or 1, denoting the absence or presence of apneic event respectively. The networks, which were built using SpO₂ and ECG signals individually, had 30 neurons in the input layer to take in 30×1 input vector. Since the third model was created using the combination of two signals, its input layer had 60 neurons.

For optimization, the Adam optimization technique [26] was used while in the training phase, mean-squared loss function was used to achieve the model parameters with optimal values. In the proposed feed-forward ANN model each hidden layer had ReLU activation function (eq. 1) which is a piece-wise linear function that sends the input as output if the value is positive or zero and a forced zero for the negative input values. The output layer had Sigmoid activation function (eq. 2) which successfully categorizes between 0 (Normal Condition) and 1 (Apnea Occurred).
The training was performed for 1000 epochs with 10-fold cross-validation on the training set having a mini-batch size of 10. The performance of a model achieved by the cross-validation technique was evaluated based on its accuracy.

\[
\text{ReLU}(X) = \begin{cases} 
0, & x < 0 \\
 x, & x \geq 0
\end{cases} \tag{1}
\]

\[
\text{Sigmoid}(X) = \frac{1}{1 + e^{-X}} \tag{2}
\]

After having an estimation of the performance of each model employing the cross-validation technique, it was further analyzed to evaluate the performance of the model in a test set. This test set was separated from the entire set of data and was not used in training the model. On this occasion, the performance of a model was evaluated based on accuracy, precision, recall, and F1-score.

![Figure 3: Architecture of the proposed feed-forward deep neural network.](image)

**Results**

Following the data segmentation and the artifact removal steps, different number of sequences were extracted from each type of signal. From the SpO2 signal, a total of 3,815 sequences were obtained, out of which 2,293 corresponded to normal events and 1,522 corresponded to apnea events. Out of the 2,606 sequences extracted from the ECG signal, 1,534 were from normal and 1,072 were from apnea events. From the combined signal, a total of 2,530 sequences were obtained with 1,512 normal and 1,018 apnea events. Each of the combined sequences had 60 sample points and the rest of the sequences had 30 sample points. For each model, the entire associated set of sequences was split into training and testing sets with a ratio of 3:1. Table 1 shows the number of sequences before and after the split with the numbers of associated normal and apnea events.

<table>
<thead>
<tr>
<th>Table 1. Number of Sequences Before and After the Train-Test Split</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td><strong>Entire set</strong></td>
</tr>
<tr>
<td>Normal</td>
</tr>
<tr>
<td>Apnea</td>
</tr>
<tr>
<td>Total</td>
</tr>
<tr>
<td><strong>Train</strong></td>
</tr>
<tr>
<td>Normal</td>
</tr>
<tr>
<td>Apnea</td>
</tr>
<tr>
<td>Total</td>
</tr>
<tr>
<td><strong>Test</strong></td>
</tr>
<tr>
<td>Normal</td>
</tr>
<tr>
<td>Apnea</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>
To build the ANN models from each signal, 10-fold cross-validation was performed using the associated training set. The performance of the model was evaluated based on the accuracy achieved after validation. Figure 4-a) shows the comparison of the cross-validation performance among the models. It can be seen that the best-performing model was the one that was built using the combined signal. The cross-validation accuracy of this model was 91.83 ± 1.51%. SpO2 based model was the 2nd best model with an accuracy of 90.78 ± 10.12%. The ECG signal-based model demonstrated an accuracy of 80.04 ± 7.7%, which was the least among the three models.

Figure 4: a) accuracy (%) of the models after 10-fold cross-validation, b) accuracy, precision, recall, and F1-score of all the models on the test data.

Figure 4-b) shows the performance of the models in the test set, which was separated during the train-test split. For each type of signal, a new model was created using the entire training set, and the performance of the model was evaluated by using accuracy, precision, recall, and F1-score. Although the combined signal provided the best performing model from the cross-validation results, the SpO2-based model outperformed others in predicting the test set. It achieved the highest values of all the performance metrics. It had an accuracy of 94% whereas the ECG and combined signal provided accuracy of 92% and 78%, respectively and its precision, recall, and F1-score were 94%, 89%, and 92%, respectively. As estimated from the validation result, the ECG-based model had the least values of the performance metrics. The combined signal-based model was the 2nd best model with accuracy, precision, recall, and F1-score of 92%, 92%, 88%, and 90%.

Discussion

In this study, three different types of input signals were used to build the feed-forward ANN model. The objectives of this study were to evaluate how the SpO2 and the ECG signals perform to detect sleep apnea and to observe whether the combination of those two signals can provide better performance. From the cross-validation result, it was evident that the combined signal outperformed the models based on individual signals. But for evaluation of the test data, the SpO2-signal based model performed better than the combined signal-based model. A comparison of the accuracy of these two models from the cross-validation demonstrates that their accuracy values were not significantly different. But the standard deviation of all the folds was significantly high for the SpO2 signal. It was 10.12% for SpO2, whereas it was only 1.51% for the combined signal. Such a low standard deviation value indicates that the combined signal-based model is very consistent in detecting with high accuracy from a set of unknown data. On the other hand, although the SpO2-based model performed better on the test set, it may not be as consistent as the combined signal for detecting apnea from an unknown dataset. Therefore, it can be concluded that the combined signal-based model is the most reliable model with a very high accuracy of 91.83%.

Although the ECG-based model did not yield an accuracy as high as the SpO2-based model, once the ECG sequence was combined with the SpO2 sequence, it complemented the SpO2 sequence and as a result, the performance of the combined sequence was better than the individual SpO2 sequence. This is an indication that instead of a single-channel technique, a dual-channel technique using both the SpO2 and the ECG signals should be a better option for real-time apnea detection. Nowadays, there are multiple studies involving wearable sensor-based real-time apnea detection systems [27][28]. With the advancement in the field of sensor technology, researchers are highly interested in such
wearable sensor-based apnea detection systems. The study presented in this work will provide the researchers with a direction towards a dual-channel technique of detection.

In this work, every 1-minute interval of the signal was analyzed, but the classification was performed based on the first 30 seconds of each of the 1-minute intervals. Most of the studies based on the Physionet Apnea-ECG database reported in the literature utilized the entire 1-minute segment. Therefore, these models needed a sequence of 1-minute duration to detect apnea events. On the other hand, the proposed model can generate output by working on a 30-seconds long sequence. Such a shorter processing window makes this model a better contender for a real-time apnea detection system.

Most of the neural network-based apnea detection systems rely on feature extraction from the signal after necessary noise and artifact removal. In this study, a feed-forward neural network model is proposed, which does not need any feature extraction from the SpO2 signal once the noise and artifacts are removed. On the other hand, the proposed ECG-based model takes R-R interval as input. Most of the ANN-based systems using R-R intervals as input require additional features from the R-R interval sequences, such as mean, variance, maximum value, minimum value, etc. Other studies that do not require feature extraction from R-R interval sequence are usually based on more complex networks, such as CNN, RNN, etc. To the best of our knowledge, this study is the first demonstration of a simple feed-forward ANN-based apnea detection system using R-R interval without additional features.

There have been multiple demonstrations of these models yielding higher accuracy than the proposed model [16]. In the era of artificial intelligence embedded hardware systems, simplification of the models is as highly desired as the accuracy for the feasibility of embedding the models on chips. Techniques, such as pruning and quantization can further be applied on the proposed model to make it more suitable for hardware implementation within power constraints. CNN and RNN are too complex to be embedded in hardware even after pruning and quantization.

The major limitation of the study is that the proposed architecture of neural network failed to achieve high values of performance metrics in detecting sleep apnea from the ECG signal. It can be attributed to the erroneous R-R intervals. The R-R interval series was extracted from the raw ECG signal by using the machine-generated QRS annotations. The QRS detector that the database used to generate the annotations was unaudited and contained errors as mentioned in the PhysioNet website (https://physionet.org/content/apnea-ecg/1.0.0/). A better QRS detector could have detected the R-beats at accurate time points with less error. Moreover, only 8 recordings were used in this study. In our future study, we aim to work with a larger dataset and design our own QRS detector, which can lead to better accuracy of the model.

Conclusion

In this study, SpO2 and ECG signals were used both individually and in combination to build a real-time SA detection system with feed-forward ANN. R-R intervals were derived from the raw ECG signal and three models were built using SpO2 and the R-R intervals. The proposed models did not require any additional features from any of those signals for SA detection. According to the results from 10-fold cross-validation, it was evident that the combined signal-based model performed better than the individual signal-based models. Although the accuracy achieved by the ECG-based model was significantly lower than the SpO2-based model, which was attributed to the erroneous QRS detector, in the model employing the combination of SpO2 and ECG the two complemented each other resulting in the highest accuracy among the three models. From this study, it can be concluded that the dual-channel technique is preferable for a real-time apnea detection system and a better performing model can be built by using a more efficient QRS detector.

References

1. Uddin MB, Chow CM, Su SW. Classification methods to detect sleep apnea in adults based on respiratory and oximetry signals: A systematic review. Physiol Meas. 2018
Half the picture: Word frequencies reveal racial differences in clinical documentation, but not their causes

Jacqueline A. Penn¹, Denis Newman-Griffis, PhD²
¹Thomas Jefferson University, Philadelphia, PA; ²University of Pittsburgh, Pittsburgh, PA

Abstract

Clinical notes are the best record of a provider’s perceptions of their patients, but their use in studying racial bias in clinical documentation has typically been limited to manual evaluation of small datasets. We investigated the use of computational methods to scale these insights to large, heterogeneous clinical text data. We found significant differences in negative emotional tone and language implying social dominance in clinical notes between Black and White patients, but identified multiple contributing factors in addition to potential provider bias, including mis-categorization of some healthcare vocabulary as emotion-related. We further found that notes for Black patients were significantly less likely to mention opioids than for White patients, potentially reflecting both inequitable access to medication and provider bias. Our analysis showed that computational tools have significant potential for studying racial bias in large clinical corpora, and identified key challenges to providing a nuanced analysis of bias in clinical documentation.

Introduction

Addressing racial disparities in healthcare requires that mechanisms of disparity be identified and continually monitored. Informatics technologies have the potential to be tools for combatting racial disparities in healthcare, but can also serve to amplify racial biases in the health data they analyze. Understanding what health data are saying—and implying—about patients and the care they receive is key both to improving the quality and equitability of that care and to building equitable informatics technologies. While structural factors such as income inequality are major contributors to racial health disparities, implicit racial biases exhibited by healthcare providers also play a significant role due to their effect on clinical decision-making and patient-provider interactions.¹ Clinical documentation is the best record of providers’ perceptions of their patients, and is central to clinical decision making throughout the course of care. Thus, understanding what implicit bias looks like in clinical documentation is key to both improving provider education, such as modifying medical school curriculums or Continuing Medical Education (CME) for licensed professionals, and designing equitable natural language processing (NLP) methods to analyze the invaluable information in clinical notes.

The impacts of implicit bias on medical care and health outcomes are significant. Previous studies have found that clinicians with higher levels of pro-White implicit bias, as measured by Implicit Association Tests (IATs), are less likely to, relative to White patients, treat Black patients with thrombolytics² or narcotics,³ ⁴ less likely to refer Black patients with chest pain to a specialist⁵, and more likely to diagnose Black patients with less-severe disease.⁵ The pain of Black patients is often underestimated and undertreated, likely due in part to common myths that persist in healthcare. A 2016 study found that many medical students and residents believed myths about Black patients, such as that they have thicker skin or less sensitive nerve endings than White patients. Furthermore, the authors found that belief in such race-based medical myths was correlated with less accurate treatment decisions for Black patients.⁶

Previous studies have shown that provider language can reflect these racial biases. Beach et al.⁷ found higher levels of disbelief-related language in clinical notes for Black and female patients, supporting the idea that Black and female patients’ complaints are not taken as seriously as White or male patients. Hagiwara et al.⁸ analyzed transcriptions of physician-patient interactions, and found greater use of anxiety-related language and first-person plural pronouns (an indicator of social dominance) in racially discordant interactions by physicians with greater levels of IAT-measured implicit bias. Park et al.⁹ also looked at social dominance by hand-analyzing 600 clinical notes and found social dominance to be one of the ways clinicians express negative emotions about the patient, in addition to mechanisms like depicting the patient as untrustworthy or difficult. However, such studies have largely been limited to hand analysis of a small number of notes, limiting their generalizability to broader health data.

In this study, we used well-established NLP tools from computational social science and medical NLP to study evidence of racial bias in a large dataset of critical care clinical notes. Our primary hypothesis was that there are significant differences between critical care notes written for Black and White patients even after pairing based on patient age, gender, and primary diagnosis. We hypothesized that specific differences would be observed in both note
style and note content, reflecting interpersonal and systemic elements of racism, respectively. In particular, we hypothesized that (1) notes for Black patients have a more negative emotional tone than notes for White patients, as well as higher levels of anxiety- and anger-related language; (2) notes for Black patients have more language related to social dominance, as measured by first personal plural pronouns and power-related language; and (3) there are significant differences in the mentions of opioid pain medication between Black and White patients, reflecting previously-observed systemic inequities in perception of pain and access to pain medication for Black patients.

Materials and Methods

Data

All data for this study comes from the Medical Information Mart for Intensive Care, version 3 (MIMIC-III), a public-use database of critical care admissions from Beth Israel Deaconess Medical Center in Boston, MA, between years 2001 to 2012. As the largest and most detailed public-use clinical database available, the iterative releases of the MIMIC dataset have been invaluable resources for clinical informatics research. In addition to its depth of structured data from critical care, including vital signs, lab reports, medications and procedures, etc., MIMIC includes a wealth of unstructured data in over two million free text clinical notes. These unstructured data are much more challenging to deidentify in securing clinical data for research purposes, making MIMIC central to research in medical NLP.

The most recent release of MIMIC to include free text notes, MIMIC-III, has been used to develop benchmark datasets for medical concept normalization, medical question answering, and medical natural language inference, as well as in developing clinical language models that are heavily used in current research. Thus, if patterns of injustice— including implicit provider bias affecting patient interactions and care as well as structural injustice limiting access to high-quality care—are reflected in MIMIC data, then these patterns have the potential to be promulgated or exacerbated by NLP systems built on MIMIC’s foundation.

The representativeness of MIMIC clinical notes with respect to racial identity, and the racialized differences they reflect in the delivery of medical care, have not been investigated. This study presents an initial characterization of racialized differences in MIMIC documentation, and provides insights from our analyses into confounding factors and methodological challenges that may affect investigations into what clinical documentation reveals about the causes behind racial health disparities. By focusing on MIMIC, our analysis is reproducible by other researchers and shines an equity-focused light on a foundational resource for medical NLP research.

We used five tables from MIMIC-III. The patients table includes basic patient information such as gender (only available as binary male/female labels; gender assigned at birth was not explicitly recorded), date of birth, and date of death. The admissions table includes details related to each Intensive Care Unit (ICU) admission, such as admit time and discharge time, as well as patient demographics like ethnicity and insurance provider. We identified Black and White patients by extracting race from the ethnicity variable, which included both race and ethnicity (e.g., “WHITE – RUSSIAN” was mapped to “WHITE”). Primary ICD-9 codes for each admission were incorporated from the diagnoses_icd table, and the names of those diagnoses were taken from the d_icd_diagnoses table. Primary diagnoses were differentiated from secondary diagnoses by restricting to observations in the diagnoses_icd table where seq_num equaled 1. All clinical note information, including note text, is from the noteevents table. The full database consisted of 46,520 patients, 58,976 admissions, 2,083,180 notes, and 7,567 caregivers (e.g., physicians, nurses).

Cohort Construction

We constructed racially-paired cohorts for our analysis. Because different stereotypes and biases exist for different races, we restricted White or Black race only, allowing us to focus exclusively on pro-White, anti-Black racial bias. Additionally, MIMIC has few admissions for patients who are not Black or White (79.9% of admissions are Black or White), limiting our ability to conduct analyses with sufficient statistical power on other racial groups. We also restricted our sample to the two most common primary diagnoses by ICD-9 code, Unspecified Septicemia (ICD-9 code: 0389) and Coronary Atherosclerosis of Native Coronary Artery (ICD-9 code: 41401). We further stratified our sample into male and female cohorts to control for gender-related documentation differences. As our sample was not large enough to have grouped age cohorts, we restricted all cohorts to age 50 years and older to reduce age-related effects. This removed neonates from our sample and allowed us to focus on a relatively older population while still maintaining nearly 50 admissions in our smallest cohorts. We also dropped all notes marked as being made in error. The final sample consisted of 99,936 notes corresponding to 3,903 admissions of 3,748 patients. The notes were written by 1,230 unique caretakers, and our analysis included all available note categories (Table 1). This included all nursing and physician notes as well as notes from pharmacy, social work, rehab services, etc.
Of this sample, 93.4% of patients were White and 66.4% were male. Sample stratification by race, gender, and diagnosis resulted in four pairs of cohorts, each pair differing by race (Table 2). We had 16 patients (2 Black, 14 White) who belonged to 2 cohorts, due to being admitted at separate times for different diagnoses. These patients represent 34 admissions (5 Black, 29 White) and 988 notes (51 Black, 937 White; <1% of all notes in the sample).

**Vocabulary Analysis**

We conducted an exploratory vocabulary analysis to get a sense of what types of words differed in relative frequency by race, and whether any of these words might reflect provider bias. We used SpaCy to tokenize the full note text for all available note categories, normalized each token to the lowercase version of its lemma, and restricted to alpha tokens only. We then calculated the frequency of each of these token lemmas by cohort and calculated the difference in relative frequency (normalized by number of admissions in the cohort) between matched cohorts. In browsing these frequencies, we identified notable differences in words that may reflect note writer bias towards the patient.

To conduct a more topic-based analysis beyond the level of individual words, we employed the Linguistic Inquiry and Word Count (LIWC) software, 2015 edition. LIWC is widely used in computational social science to analyze the frequency of various word categories, including emotion categories, social categories, and syntactic categories. LIWC includes 70 categories in total, and results are reported as the percentage of each note falling within a given category. LIWC allowed us to quantify the degree to which anger-related words are represented in a cohort’s notes, as well as other categories of words that may indicate bias such as emotional tone, negative emotion, and positive emotion. We also explored the themes of anxiety and social dominance, given their relevance in previous literature. We measured social dominance using the LIWC category of first person plural pronouns (such as in Hagiwara et al.). Information on our selected LIWC categories are provided in Table 3.

By applying LIWC software to the note text in our sample, we obtained the percentage of each note belonging to our categories of interest. The category of emotional tone is the exception, as that category is not reported as percentage of note, but rather a number from 0 to 100, with 0 being totally negative in tone, 50 being neutral, and 100 being totally positive. We then used the Mann-Whitney U test on these note-level observations with significance threshold \( p<0.05 \) to test whether each Black/White pair of cohorts differed in their percentage of note text in these categories. The Mann-Whitney U Test was chosen due to the non-normal distribution of the data, given that usually only a small amount of each note’s text belongs to a given LIWC category, if at all.

**Table 1.** Frequency of note categories within our sample.

<table>
<thead>
<tr>
<th>Note Type</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nursing/other</td>
<td>27,003</td>
</tr>
<tr>
<td>Radiology</td>
<td>22,062</td>
</tr>
<tr>
<td>Nursing</td>
<td>16,646</td>
</tr>
<tr>
<td>ECG</td>
<td>11,328</td>
</tr>
<tr>
<td>Physician</td>
<td>10,493</td>
</tr>
<tr>
<td>Discharge Summary</td>
<td>4,308</td>
</tr>
<tr>
<td>Echo</td>
<td>3,232</td>
</tr>
<tr>
<td>Respiratory</td>
<td>2,802</td>
</tr>
<tr>
<td>General</td>
<td>728</td>
</tr>
<tr>
<td>Nutrition</td>
<td>719</td>
</tr>
<tr>
<td>Rehab Services</td>
<td>372</td>
</tr>
<tr>
<td>Social Work</td>
<td>142</td>
</tr>
<tr>
<td>Case Management</td>
<td>92</td>
</tr>
<tr>
<td>Consult</td>
<td>5</td>
</tr>
<tr>
<td>Pharmacy</td>
<td>4</td>
</tr>
</tbody>
</table>

**Table 2.** Number of patients, admissions, and notes associated with each cohort.

<table>
<thead>
<tr>
<th>Race</th>
<th>Gender</th>
<th>Diagnosis</th>
<th>Patients</th>
<th>Admissions</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>Male</td>
<td>Septicemia</td>
<td>664</td>
<td>712</td>
<td>27,777</td>
</tr>
<tr>
<td>Black</td>
<td>Male</td>
<td>Septicemia</td>
<td>70</td>
<td>81</td>
<td>2,579</td>
</tr>
<tr>
<td>White</td>
<td>Male</td>
<td>Atherosclerosis</td>
<td>1,717</td>
<td>1,728</td>
<td>31,256</td>
</tr>
<tr>
<td>Black</td>
<td>Male</td>
<td>Atherosclerosis</td>
<td>48</td>
<td>48</td>
<td>742</td>
</tr>
<tr>
<td>White</td>
<td>Female</td>
<td>Septicemia</td>
<td>617</td>
<td>659</td>
<td>21,741</td>
</tr>
<tr>
<td>Black</td>
<td>Female</td>
<td>Septicemia</td>
<td>86</td>
<td>103</td>
<td>3,920</td>
</tr>
<tr>
<td>White</td>
<td>Female</td>
<td>Atherosclerosis</td>
<td>516</td>
<td>524</td>
<td>11,119</td>
</tr>
<tr>
<td>Black</td>
<td>Female</td>
<td>Atherosclerosis</td>
<td>46</td>
<td>48</td>
<td>802</td>
</tr>
</tbody>
</table>

**Table 3.** LIWC categories used for clinical note text analysis. The emotional tone category is a non-transparent variable calculated by aggregating over multiple categories, so it has no associated library of words.

<table>
<thead>
<tr>
<th>Category</th>
<th>Example Words</th>
<th>Words in Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotional Tone</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Positive Emotion</td>
<td>love, nice, sweet</td>
<td>620</td>
</tr>
<tr>
<td>Negative Emotion</td>
<td>hurt, ugly, nasty</td>
<td>744</td>
</tr>
<tr>
<td>Anxiety</td>
<td>worried, fearful</td>
<td>116</td>
</tr>
<tr>
<td>Anger</td>
<td>hate, kill, annoyed</td>
<td>230</td>
</tr>
<tr>
<td>Power</td>
<td>superior, bully</td>
<td>518</td>
</tr>
<tr>
<td>1st Per Plural Pron</td>
<td>we, us, our</td>
<td>12</td>
</tr>
</tbody>
</table>
Analysis of Opioid Mentions

To measure whether the relative mentions of pain medications significantly differed by patient race, we used the Apache cTAKES\textsuperscript{21} software (version 4.0.0) to identify the clinical concepts in each note and map them to the Unified Medical Language System\textsuperscript{25} (UMLS). cTAKES locates clinical concepts in a note and returns a UMLS Concept Unique Identifier (CUI) and a polarity value (negated terms have a polarity of -1). We created an opioid indicator that equaled 1 if the admission contained a note that mentioned any of the most common opioid analgesics used in the ICU, including morphine, fentanyl, hydromorphone, oxycodone, and methadone. This list was created with guidance from the clinical decision making reference, UpToDate,\textsuperscript{23} and opioids were identified in the data by tagging CUIs for which the preferred name contained the string opioid, morphine, fentanyl, hydromorphone, remifentanil (which had no matches), oxycodone, or methadone. We then ran a logistic regression model with robust standard errors (which account for any heteroskedasticity in the sample) to test whether race is significantly correlated with the likelihood of an admission containing an opioid mention in a note (Equation 1). Because negative opioid mentions in the notes were less likely, compared to positive mentions, to be reflections of opioids administered or prescribed, a second version of the regression was run. Here, the independent variable equaled 1 if the admission contained a note mentioning an opioid and the polarity equaled 1, unlike the original regression which had no polarity restriction.

**Equation 1.** Logistic regression of opioid on admission characteristics. The indicator opioid equaled 1 if the clinical note mentioned an opioid, 0 else; black equaled 1 if the patient was Black, 0 if White; female equaled 1 if the patient was female, 0 if male; septicemia equaled 1 if the primary diagnosis of the admission was sepsis, 0 if atherosclerosis. Model coefficients are represented by $\beta_1$ to $\beta_3$ and $\beta_0$ is the intercept.

$$opioid = \beta_0 + \beta_1\text{black} + \beta_2\text{female} + \beta_3\text{septicemia} + \varepsilon$$

### Results

**Vocabulary Analysis**

From our exploratory vocabulary analysis, we found there were many words with large frequency differences between racially paired cohorts. Some of these were indicative of health differences (e.g., diabetes and renal were much more common in notes for Black patients, reflecting the higher rates of diabetes and kidney disease in the Black population), and others seemed non-meaningful (e.g., and for were more common in notes for White patients). Anger-related words such as rude, belligerent, uncooperative, and aggressive stood out as possible indicators of racial bias (Table 4).

In notes for male sepsis patients, the word rude appeared 36 times in 81 admissions for Black patients, but only once in 712 White admissions (a ~300-fold difference). Among the same sample, belligerent was mentioned 36 times in 81 Black admissions, and 16 times in 712 White admissions (a ~20-fold difference); uncooperative

<table>
<thead>
<tr>
<th>Word</th>
<th>Race</th>
<th>Frequency</th>
<th>Cohort Admits</th>
<th>Freq/Cohort Admits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rude</td>
<td>White</td>
<td>1</td>
<td>712</td>
<td>0.0014</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>36</td>
<td>81</td>
<td>0.44</td>
</tr>
<tr>
<td>Belligerent</td>
<td>White</td>
<td>16</td>
<td>712</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>36</td>
<td>81</td>
<td>0.44</td>
</tr>
<tr>
<td>Uncooperative</td>
<td>White</td>
<td>38</td>
<td>712</td>
<td>0.053</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>25</td>
<td>81</td>
<td>0.31</td>
</tr>
<tr>
<td>Aggressive</td>
<td>White</td>
<td>1409</td>
<td>712</td>
<td>1.98</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>244</td>
<td>81</td>
<td>3.01</td>
</tr>
</tbody>
</table>

**Table 5. Descriptions of the use of the word rude in clinical notes.**

<table>
<thead>
<tr>
<th>Theme</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Possible negative attitude towards patient</td>
<td>&quot;Patient has been totally appropriate tonight, only rude and stubborn. He wants things when he wants them, doesn’t t wait, gets [out of bed] without assistance even though he is told not to.&quot;</td>
</tr>
<tr>
<td>Description of non-patient</td>
<td>“One son esp, not following rules ie calling before entering unit, is loud, angry and rude to [name] and MD’s”</td>
</tr>
<tr>
<td>Reporting what someone else said</td>
<td>“Floor nurse reported that pt. was very anxious and verbally rude due to annoyance with alarms and inability to get sleep.”</td>
</tr>
<tr>
<td>Describing a disease with mental/emotional symptoms</td>
<td>“At times can be rude with nurses. ...Probably a combination of uremic/hepatic encephalopathy in the of sepsis.”</td>
</tr>
<tr>
<td>Copy + paste</td>
<td>&quot;Mental status: Pt was seen to be occasionally confused, saying odd things, and sometimes belligerent, making aggressive/rude comments to staff, and other times non-compliant, taking off BP cuff, pressing the pump buttons, not staying in bed, etc. Could be related to hepatic encephalopathy vs personality disorder.&quot; (repeated in 36 notes)</td>
</tr>
</tbody>
</table>
was mentioned 25 times in 81 Black admissions, and 38 times in 712 White admissions (a ~6-fold difference); *aggressive* was mentioned 244 times in 81 Black admissions, and 1409 in 712 White admissions (a ~1.5-fold difference).

To evaluate the validity of our vocabulary analysis, we conducted a manual review of all notes containing the word *rude*. This manual review provided several possible contributing factors for the use of the word *rude* in addition to possible note writer bias (Table 5). In particular, the copy-and-pasting of notes had a major influence, as the entirety of the differences seen in our exploratory vocabulary analysis were the result of a single sentence in the Black, male septicemia cohort getting copied across 36 notes. Furthermore, this same sentence accounted for all uses of *belligerent* and 36 of the 244 uses of *aggressive* in that cohort.

**LIWC category analysis**

We observed several types of statistically significant differences in the prevalence of LIWC categories between clinical notes for our paired cohorts, displayed in Table 6. (1) Notes for Black patients had more negative overall emotional tone than notes for White patients (male septicemia: p = 0.000, female septicemia: p = 0.000) and, separately, fewer positive emotion words (male septicemia: p = 0.000, female septicemia: p = 0.046, female atherosclerosis: p = 0.009) and more negative emotion words (male septicemia: p = 0.000, male atherosclerosis: p = 0.039, female septicemia: p = 0.001). (2) There were correlations between patient race and anxiety language (male septicemia: p = 0.000, male atherosclerosis: p = 0.009), and weak correlations between race and anger language (female septicemia: p = 0.017). However, these results were opposite of the expected direction, as all significant findings showed greater anxiety and anger in notes for White patients. (3) Notes for Black patients had higher levels of social dominance-related language, as measured by LIWC’s *power* category (male septicemia: p = 0.000, female septicemia: p = 0.000) and first person plural pronouns in the male septicemia cohorts (p = 0.000). However, the results for the female cohorts showed greater use of first person plural pronouns in notes for White patients (female septicemia: p = 0.013, female atherosclerosis: p = 0.026).

**Table 6.** Differences in cohort-level means for LIWC categories. The *B* columns represent the mean percentage of the given LIWC category in the relevant Black cohort’s notes, and the *W* columns represent the mean percentage of the given LIWC category in the relevant White cohort’s notes. The category of *emotional tone* is the exception, as that category is not reported as a percentage of note, but rather a number from 0 to 100, with 0 being totally negative in tone, 50 being neutral, and 100 being totally positive. The *B-W* columns represent the value of the relevant *B* column minus the paired *W* column. Significance stars represent the *p*-values from Mann-Whitney U Tests. ***p<0.001, ** p<0.01, * p<0.05.

<table>
<thead>
<tr>
<th>LIWC Category</th>
<th>Male Septicemia</th>
<th>Female Septicemia</th>
<th>Male Atherosclerosis</th>
<th>Female Atherosclerosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Emotion</td>
<td>1.486 1.556 -</td>
<td>1.400 1.457 -</td>
<td>1.475 1.581 -</td>
<td>1.383 1.461 -</td>
</tr>
<tr>
<td>Negative Emotion</td>
<td>1.721 1.553 +</td>
<td>1.568 1.469 +</td>
<td>1.588 1.507 +</td>
<td>1.477 1.535 -</td>
</tr>
<tr>
<td>Anxiety</td>
<td>0.209 0.235 -</td>
<td>0.168 0.191 -</td>
<td>0.214 0.245 -</td>
<td>0.187 0.204 -</td>
</tr>
<tr>
<td>Anger</td>
<td>0.074 0.067 +</td>
<td>0.046 0.055 -</td>
<td>0.043 0.063 -</td>
<td>0.047 0.044 +</td>
</tr>
<tr>
<td>Power</td>
<td>2.269 2.096 +</td>
<td>2.214 2.136 +</td>
<td>2.233 2.139 +</td>
<td>2.044 2.169 -</td>
</tr>
<tr>
<td>1st Per Plural Pron</td>
<td>0.048 0.030 +</td>
<td>0.006 0.010 -</td>
<td>0.032 0.035 -</td>
<td>0.008 0.013 -</td>
</tr>
</tbody>
</table>

There was a great deal of variance among the individual clinical notes in terms of LIWC values. As an example of this variance, the distribution of LIWC categories for three notes in the same admission (*hadm_id* = 100009; from the White, male atherosclerosis cohort) is shown in Figure 1. No two example notes share the same set of observed categories, and the frequency of each category varies widely.

**Figure 1.** The distribution of LIWC categories in three notes from a single admission (*hadm_id* = 100009; from the White, male atherosclerosis cohort). The *emotional tone* category was excluded because it is not reported as a percentage of note text.
Each category is only observed in a subset of the notes, although all of our selected categories were observed in each cohort. Some categories, like negative emotion, were found in nearly all notes, while others, such as anger, were observed in less than 20% of notes (Figure 2). To understand the trends in LIWC values when they are observed, we graphed the nonzero values for each LIWC category by cohort (results for negative emotion and anger shown in Figure 3; other categories exhibited similar patterns) and measured significant differences between cohorts using Mann-Whitney U tests. Means and medians remained similar between the paired cohorts, but the spread of the values varied considerably, particularly in the first and fourth quartiles. Thus, small sets of notes with unusually high or low category frequencies were the primary factors distinguishing the cohorts, rather than systematic trends.

Figure 2. The percentage of notes in each cohort for which a nonzero proportion of the words were tagged within each LIWC category.

Figure 3. Distribution of LIWC categories negative emotion and anger, by cohort. Box-and-whisker plots were created using nonzero values only and extreme values (outside 1.5 times the interquartile range) are excluded. Significance stars are p-values of Mann-Whitney U Tests on nonzero observations. *** p<0.001, ** p<0.01, * p<0.05.
Analysis of Opioid Mentions

Fourteen CUIs representing opioids were identified in the clinical notes. The most mentioned were fentanyl (N = 13,648), morphine (N = 10,925), and oxycodone (N = 3,996). Nearly all mentions (99.8%) had a positive polarity. The full list of identified CUIs and their frequencies are listed in Table 7.

The results of the logistic regression (Equation 1) indicated that notes for Black patients were less likely to mention opioids relative to notes for White patients (odds ratio = 0.685, p = 0.003). The results of the second regression, in which the dependent variable equaled 1 if an opioid was mentioned and the polarity was positive, were nearly identical to the original regression (odds ratio = 0.687, p = 0.004). Opioids (as listed in Table 7) were commonly mentioned, occurring in over 50% of the notes in each cohort (Figure 4). Opioid mentions were more common for White patients in all cohorts except female septicemia patients.

Table 7. Opioid-related CUIs identified in the clinical note text and their frequencies.

<table>
<thead>
<tr>
<th>CUI</th>
<th>Name</th>
<th>Total</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0015846</td>
<td>fentanyl</td>
<td>13,648</td>
<td>13,624</td>
<td>24</td>
</tr>
<tr>
<td>C0026549</td>
<td>morphine</td>
<td>10,925</td>
<td>10,908</td>
<td>17</td>
</tr>
<tr>
<td>C0030049</td>
<td>oxycodone</td>
<td>3,996</td>
<td>3,987</td>
<td>9</td>
</tr>
<tr>
<td>C0066814</td>
<td>morphine sulfate</td>
<td>2,882</td>
<td>2,882</td>
<td>0</td>
</tr>
<tr>
<td>C0717368</td>
<td>acetaminophen / oxycodone</td>
<td>1,760</td>
<td>1,760</td>
<td>0</td>
</tr>
<tr>
<td>C0012306</td>
<td>hydromorphone</td>
<td>1,487</td>
<td>1,478</td>
<td>9</td>
</tr>
<tr>
<td>C0025605</td>
<td>methadone</td>
<td>903</td>
<td>903</td>
<td>0</td>
</tr>
<tr>
<td>C0546864</td>
<td>fentanyl citrate</td>
<td>522</td>
<td>520</td>
<td>2</td>
</tr>
<tr>
<td>C0282274</td>
<td>oxycodone hydrochloride</td>
<td>239</td>
<td>239</td>
<td>0</td>
</tr>
<tr>
<td>C0242402</td>
<td>opioids</td>
<td>85</td>
<td>76</td>
<td>9</td>
</tr>
<tr>
<td>C0721688</td>
<td>methadone hydrochloride</td>
<td>70</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>C0700533</td>
<td>hydromorphone hydrochloride</td>
<td>70</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>C0360457</td>
<td>morphine oral product</td>
<td>16</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>C1874397</td>
<td>atropine / morphine</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Discussion

Vocabulary and LIWC Analyses

The results of our analyses show significant differences in language used for Black and White patients in clinical notes. Notes for Black patients appeared to have a more negative emotional tone than notes for White patients, as reflected by differences in overall emotional tone as well as differences in positive and negative emotion. This could be a reflection of clinician bias, given Park et al.’s finding that negative emotional language is a form of stigmatizing language in clinical notes.9

We also observed higher levels of social dominance-related language in notes for Black patients. Like Hagiwara et al.,8 we did find significant differences in first person plural pronoun use between Black and White cohorts. However, the direction was not always consistent across cohort pairs or with previous findings. We found that for female patients with a septicemia or atherosclerosis diagnosis, there was greater use of first person plural pronouns in notes for White patients compared to those for Black patients. For male patients, we found significantly fewer mentions of such pronouns for White septicemia patients and no significant difference for male atherosclerosis patients.

Our other measure of social dominance, LIWC’s power category, also represented a higher percentage of each note for Black patients, compared to White. Results for anxiety- and anger-related language were opposite of what we expected based on previous work and our vocabulary analysis, as all significant findings showed lower levels of anxiety or anger in notes for Black patients compared to White. However, the results for anger were only statistically significant for female septicemia patients.

Validity of LIWC

To dig deeper into these findings and assess the degree to which they indicate potential provider bias, we reviewed the top words tagged in each LIWC category. We found that for some categories, LIWC was unable to differentiate between clinical language and words that may reflect bias. For example, in the male septicemia cohorts, the top words in the negative emotion category included pain, failure, low, lower, and shock. These words are typically used in the clinical context to describe patient health status, rather than reflections of provider attitude.
Similarly, top words in the power category included failure, status, up, low, and doctor—each of which is again in common clinical usage without indications of social dominance expected in non-healthcare contexts. Specific categories, such as anger, seemed to be more accurate at capturing provider attitude, but more context is required to determine whether anger language is present in the note due to provider attitude towards the patient, or due to some other reason, such as a description of a patient’s family member or visitor. While a word frequency analysis necessarily elides contextual details, an LIWC-like tool tailored for clinical notes would nonetheless be a valuable technology for researchers studying implicit bias in clinical notes. This would require adapting categorical dictionaries to distinguish between words likely related to health, such as pain, and words more clearly related to affect, such as belligerent.

Analysis of Opioid Mentions

Beyond stylistic findings, we also found an information content difference in that notes for Black patients are significantly less likely to mention an opioid. This could reflect racial bias, as it is an outcome consistent with the idea that health professionals underestimate and undertreat the pain Black patients experience. There are several possible explanations for why providers underestimate and undertreat Black patients’ pain, including incorrect provider beliefs about Black patients’ nerve endings and skin,6 or suspicion that Black patients are more likely to abuse opioids than White patients.24,25 It is also possible that these results do not indicate bias, but instead represent racial trends in opioid abuse, since during this time period the opioid crisis disproportionately affected White Americans.26 Both of these factors are likely to be intertwined in practice. In order to gain more nuance in this analysis, including distinguishing between opioid mentions in the patient history and ICU administration of opioids, these results could be checked against the medication administration data in MIMIC-III.

Implications

Overall, we find that as a method of investigating implicit bias in healthcare, applying computational analysis to clinical notes allows for faster analysis and the utilization of much larger datasets compared to hand analysis, but introduces additional challenges in accounting for the context and pragmatic understanding behind quantitatively observed differences. Developing more nuanced methods for computational analysis will be key to achieving the potential of computational techniques to gain insight into data of interest. For example, as computational techniques can be straightforwardly applied to any dataset without resource-intensive data curation, they can be used to evaluate and gain insight into implicit bias levels and mechanisms for teams, departments, or entire institutions. This could inform targeted interventions to combat implicit bias at multiple levels of health professions training and practice. For example, if there are racialized differences in the use of negative emotional language in clinical notes, medical trainees may be taught about this difference, which could act as awareness intervention27,28 that reduces bias in note writing. This would likely have positive downstream effects, as a reduction in biased notes would reduce the probability that a health professional would read a biased note and perceive a patient differently.

Limitations

This study had several limitations which can inform future work on developing computational methods for analyzing evidence of bias at scale.

Sample Limitations

The patient sample in MIMIC is strongly skewed White, and the dataset represents one well-resourced medical center in a major city in the Northeastern U.S. Additionally, all patients were critical care patients, who are sicker and likely more socioeconomically vulnerable than the overall population. Furthermore, patients in the ICU may be completely incapacitated and have limited interaction with providers. This may have a significant influence on provider attitude, by limiting interactions in which providers can form an opinion on the patient; conversely, this lack of conversation may make clinicians more likely to stereotype patients based on observable characteristics like skin color. We also limited our analysis to two primary diagnoses, representing a relatively small subset of the overall patient population. Diagnosis is likely to affect how clinicians interact with patients, as some diagnoses are more incapacitating than others (e.g., septicemia patients are usually far sicker than atherosclerosis patients), and some diagnoses have more behavioral manifestations than others. The racial skew of the data also limited the power of our analysis. A larger share of Black patients in the dataset would allow for further cohort stratification, controlling for factors like insurance provider or creating multiple age brackets. Additionally, the small size of the Black cohorts compared to the White cohorts reduced both statistical power and the diversity of data available to draw on. We tended to find more significant results in the male and female septicemia cohorts compared to the atherosclerosis cohorts, the latter of which had notably smaller Black populations both by absolute admission counts and relative to their White populations. Several
of these limitations are inherent in the MIMIC dataset itself, highlighting important considerations of under-representation in NLP and other informatics work based on MIMIC data. Future research on characterizing—and potentially mitigating—implicit bias in clinical documentation will thus best be served by sampling datasets with explicit criteria for diverse representation of patient demographics, including race, age, and gender identity. In addition, techniques such as propensity score matching (in cohort construction) and structured equation modeling (in data analysis) can help to reduce the influence of confounding variables.

**Missing Note Writer Demographics**

Information on note writer demographics can be important predictors for levels of implicit bias\(^29\) and would also be valuable to incorporate into comparative analyses. However, this information is included unreliably or not at all in MIMIC-III. Future study could apply similar methods to ours to identifiable data within a healthcare system, which would allow for the controlling of clinician characteristics.

**Copy-and-Pasting of Note Text**

As highlighted by our manual review, the copy-and-pasting of note text across multiple notes has the potential to substantially distort analyses relying on word counts. A recent study by Rule et al\(^30\) describes a potential method for identifying these occurrences which can be employed in future work. Additionally, the diversity of note lengths and content types in MIMIC-III notes—ranging from a few dozen words noting an encounter to extensive documentation of history and physical findings—affects LIWC values and may overweight some notes. Various strategies may be employed in future work for reducing the impact of these factors, including length-sensitive weighting and focused analyses of specific note types. For example, notes focused on objective measures and tests, such as radiology notes, may be less likely to reflect bias given the limited interaction between the note author and the patient.

**Conclusion**

This study investigated the use of computational methods to study racial bias in a large, heterogenous dataset of clinical note text. Computational analysis identified significant differences in note style and content between Black and White patients, including that notes for Black patients had more negative emotional tone, greater use of social dominance language, and fewer mentions of opioid medications. We identified multiple potential factors contributing to these differences in addition to implicit bias, including mis-categorization of healthcare words as emotional in tone. Our findings do not suggest that the impact of implicit bias in healthcare is overestimated—rather, they illuminate the complexity and importance of effective measurement and detailed analysis of evidence of bias in healthcare practice. Our study showed that computational text analysis methods have significant potential for characterizing racial differences in clinical documentation, and identified key design considerations for future research into the mechanisms of racial disparities in health documentation.

**Acknowledgements**

This research was performed during the first author’s participation in the Short-Term Trainee Program of the University of Pittsburgh Biomedical Informatics Training Program. Both authors were supported by the National Library of Medicine of the National Institutes of Health under award number T15LM007059.

**References**


Advancing Interoperability of Patient-level Social Determinants of Health Data to Support COVID-19 Research

Jimmy Phuong¹,², Stephanie Hong BS³, Matvey B. Palchuk⁴, Juan Espinoza⁵, Daniella Meeker⁶, David A. Dorr⁷, Galina Lozinski⁸, Charisse Madlock-Brown⁹, William G. Adams⁸

¹Division of Biomedical and Health Informatics, UW Medicine, Seattle, Washington; ²University of Washington Medicine Research IT, Seattle, Washington; ³Section of Biomedical Informatics and Data Science, Johns Hopkins University School of Medicine, Baltimore, Maryland; ⁴TriNetX, LLC, Cambridge, MA; ⁵Department of Pediatrics, Children’s Hospital Los Angeles, Los Angeles, CA; ⁶Department of Preventive Medicine, University of Southern California, Los Angeles, California; ⁷Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, OR; ⁸Department of Pediatrics, Boston Medical Center/Boston University School of Medicine; ⁹Tennessee Clinical and Translational Science Institute, University of Tennessee Health Science Center, Memphis, Tennessee

Abstract
Including social determinants of health (SDoH) data in health outcomes research is essential for studying the sources of healthcare disparities and developing strategies to mitigate stressors. In this report, we describe a pragmatic design and approach to explore the encoding needs for transmitting SDoH screening tool responses from a large safety-net hospital into the National Covid Cohort Collaborative (N3C) OMOP dataset. We provide a stepwise account of designing data mapping and ingestion for patient-level SDoH and summarize the results of screening. Our approach demonstrates that sharing of these important data - typically stored as non-standard, EHR vendor specific codes - is feasible. As SDoH screening gains broader use nationally, the approach described in this paper could be used for other screening instruments and improve the interoperability of these important data.
Introduction

Including social determinants of health (SDoH) data in health outcomes research is essential for studying the sources of healthcare disparities, identifying exposure and behavioral risk factors, and developing strategies to mitigate stressors. SDoH are the risk factors related to how people live, grow, and learn and they may confer advantages or disadvantages towards development of health outcomes. These psychosocial complexities (e.g., housing instability, food insecurity, social isolation, chronic stress, financial insecurity) are the social needs that contribute powerfully to racial/ethnic disparities in outcomes, which are highly influential factors towards worsening quality of life, increased disability, exacerbations of illness, and premature mortality. Screening for these factors have been crucial for understanding the complex hardships that patients experience, especially during the COVID-19 pandemic.

Despite their importance, routine collection of patient-level SDoH and facilitated uses of such information from electronic health records (EHRs) have significant data sharing barriers in obtaining and maintaining information through the data life-cycle. First, few health systems have policies to incentivize consistent screening and collection of patient-level SDoH. More and more, these policies are mandated from State or Federal funding agencies, like for Medicaid populations. Even without policy mandates, some health systems have a rapidly advancing mission to improve quality reporting, health equity, and to detect and address patient social needs. Although policies and incentives help to drive change, the institutional mission is also a key driver to collect patient-level SDoH. Health systems would need time and resource investments to modify operational workflows, develop staff capacities, and have documentation plans in place for information storage and retrieval. Studies have reported that observations about patient social determinants of health learned during clinical encounters may be documented within disparate areas within EHRs, often written from the provider perspective as unstructured clinical notes, which require significant extraction capacities for reuse. It is unclear if these observations are limited to certain clinical visit scenarios or occur under non-random conditions, or encompass what aspects of social need. More importantly, while cross-referencing between the patient and provider perspective would add context to understand the social needs, datasets describing social needs that a patient experiences should be reported from the patient and maintained as close to the source as possible. Equitable care is only possible with clear and active participation from the patients.

A growing number of institutions, particularly those serving traditionally marginalized populations, are routinely screening and supporting patients with unmet SDoH needs using structured screening instruments like PRAPARE, WellRx, WeCare, and BMC-THRIVE (or THRIVE). These screening tools inquire about housing instability, food insecurity, education attainment, access to care issues, and various other social risk factors as a short questionnaire. Structure screening instruments provide a flexible, low-barrier method for patients and providers to engage in dialogue on social needs with the care needs. Screening tools can be low-fidelity paper questionnaires or electronic survey instruments completed in the waiting-room or prior to the clinical encounter. Once collected, providers can record the instance of diagnostic screening or findings of social need for further referrals. In addition, national efforts such as the Gravity Project and UCSF Siren are developing national platforms to systematically organize and represent the data in these structured screening tools and other EHR data using data standards. The groups are also working to fill gaps in mapping when identified, providing a pathway for representing patient-level SDoH in research sourced from the patient responses to the screening tools.

While consensus mappings of standard concepts are emerging, the way to implement them across the data life-cycle is unclear. Specifically, screening tools are works-in-progress and may evolve over time with different questions and answer options. The studies have generally reported that screening tool data may be predominantly stored within EHR FlowSheet tables. Patient-level SDoH in FlowSheets means that the data are compartmentalized within a tabular data structure, relationally mapped to associate the survey, each question, the responses to each question, and the instance of data collections. Common data models (CDMs) are often the conduits for making clinical knowledge portable and accessible for research applications and secondary data use in Public Health. As FlowSheet records are localized implementations and may contain text-based responses, additional ETL (extract-transform-load) logic would need to engineer the translation between what each FlowSheet measurement means and the corresponding standardized concept representation.

Finally, with data harmonization, CDMs may transform between models allowing for data sets to be compared across standards that were previously limited by choice of CDM. One of the main hurdles to constructing FlowSheet mappings is the lack of personnel with the technical access and domain expertise. As such, services like TriNetX and consortia like Observational Health Data Sciences Institute (OHDSI) provide technical capacities that facilitate...
adoption and creation of de-identified or limited datasets. However, to analyze across multiple institutions operating with different CDMs, either the source institutions may invest to change CDM instances or midstream data harmonization may be applied to get the data into comparable format. In the National COVID Cohort Collaborative (N3C), data submission may occur using any of a select set of CDMs with the intention to harmonize data into Observational Medical Outcomes Partnership (OMOP) during the data ingestion process. There remains a paucity of research that describes this process of engineering patient-level SDoH and harmonizing the information across CDMs. Harmonization may increase the potential for data loss during translation, and loss of question set relationships. The process of translating patient-level SDoH across CDMs would benefit from intentional design during the screening tool FlowSheet mapping development.

In this report, we describe a pragmatic design and approach that leverages previous work related to SDoH Data Engineering. In collaboration with Boston Medical Center (BMC), TriNetX, and N3C, we explored the encoding needs for transmitting THRIVE screening tool responses and the technical hurdles experienced in data ingestion and harmonization (DI&H) into the N3C OMOP dataset.

Methods

Objective
We provide the step-wise account of designing data mapping and ingestion for patient-level SDoH. We describe 1) the conceptual model of information flow from clinical encounter to submission to the N3C, 2) the upstream data mapping needed for i2b2 to TriNetX, and 3) the midstream data harmonization considerations for TriNetX to OMOP. We generated concept sets reflective of the screening tool questions and answers to support inspection of the data records. The record counts were returned to the upstream data stewards as confirmatory validation of patient-level SDoH record counts in the data flow through TriNetX dataset and the N3C OMOP limited dataset.

Data source
Boston Medical Center (BMC) is the largest safety-net hospital in New England and began institution-wide screening for SDoH in 1999. As of July 1, 2021 over 200,000 patients had been screened for at least one SDoH domain. BMC has implemented screening for social needs using the THRIVE screening tool and stored the information in FlowSheet tables. The FlowSheet includes some questions and responses that are incorporated into standard instruments that are represented by LOINC, including PRAPARE, but is not itself a standardized instrument. Data is stored within Flowsheets in the Epic EHR, supporting the referral and routing of social needs detected from patient reporting.

Research instruments
The THRIVE screening tool, displayed in Figure 1, is a one-page, 8 question screening instrument with questions related to homelessness, food insecurity, trouble paying for utilities, trouble paying for medication, transportation difficulties, child/elder care challenges, and desire for additional education. Each category is screened using two different questions (Figure 1). The questions used in THRIVE were a subset of questions from national survey tools for use in routine clinical practice. Minor changes were made during the first three years of use. Multiple versions of the THRIVE screening tool were released (2017, 2018, and 2020), so patients may have responded to different versions of the screening. In most settings, the screening was administered via paper and pencil on a clipboard and the data entered by Medical Assistants within EHR FlowSheet measurement records. At BMC, routine use of THRIVE was preceded by a single question related to homelessness.
At the start of the engineering process, multiple collaborative discussions within the N3C SDoH domain team were aimed to conceptualize the process of getting patient-level SDoH data into the N3C OMOP dataset (Figure 2). The overall workflow to get data released can be simplified into four phases: 1) Clinical encounter, 2) Data entry (Transcription into EHR FlowSheet tables), 3) Primary transformations (EHR to CDM), and 4) Secondary transformations (CDM to CDM).
Clinical encounter: A patient completes the SDoH screening tool, either on their own, or administered by a healthcare provider, and per institutional protocol. The screening tool itself may be a paper form, an electronic form, or entered directly into the EHR. Data entry: The patient’s responses are entered into the EHR, storing FlowSheet records for the responses to the screening. Primary Transformation: Before transformations may happen, subject matter experts mapped the THRIVE questions and possible answers to LOINC and SNOMED codes. The EHR data goes through an ETL to be converted into an i2b2 data instance and LOINC and SNOMED codes are incorporated during this transformation. The i2b2 data instance is ingested by TriNetX, one of the acceptable N3C CDMs for data submission. Secondary Transformation: TriNetX generates the dataset for upload to N3C data ingestion pipeline, which gets parsed, applied to CDM schema, then mapped to OMOP. A final quality check provides feedback to the contributing sites, prior to being published to the cloud-based FedRamp N3C Enclave for use by the research community.

During clinical encounters, BMC clinical information as well as responses to the THRIVE screening are collected then transcribed into the EHR systems. The clinical information and screening responses were first ETL’d into an ACT/i2b2 CDM. To do this, we initially reviewed the THRIVE questions (circa 2020) and generated standardized OMOP concept mappings for the questions and answers using LOINC and SNOMED. Since the data were not directly transformed from the EHR to OMOP, the mapping encodings needed to refer to the vocabulary and concept code to be transmitted to i2b2. Once we compared with the BMC FlowSheet measurements for the THRIVE questions, questions-answers from the prior versions of THRIVE were included and for further mapping. We used PDFs of the THRIVE screening tool to cross-reference the interpretation of the questions and answer options. Most questions-answers were not represented verbatim within LOINC, so we mapped to the nearest LOINC concept with minimal loss of information or change in interpretation. If the concept was not found, we would map to the closest SNOMED (using prefix “SOMD:”) term. After three iterations of review, the mappings represented the closest semantic representation of the question-answer using LOINC and SNOMED concepts. Mappings are shown below in Table 1.

BMC electronic data warehouse creates an i2b2 data instance containing the mapped FlowSheet information for THRIVE, which gets ingested by TriNetX. Thereafter, TriNetX filters for patients that meet the N3C COVID Phenotype criteria and generates an N3C-compliant payload for BMC to submit. The LOINC and SNOMED concept codes are used to form concept sets used in the data quality checks.
Table 1. Mappings for SDoH questions related to homelessness, food insecurity, trouble paying for medication, transportation or utilities. For each THRIVE question, the original text is presented, along with the selected concept code, the code description if different than the question, and the source (the instrument or vocabulary that originated the concept code). For each THRIVE answer, the original text is presented along with the corresponding standard concept code in LOINC or SNOMED.

Data ingestion and midstream harmonization transformations
The N3C DI&H pipeline has been developed in Python and SQL and implemented in the NIH’s instance of the Palantir Foundry Platform. N3C ETL jobs are added as stepwise pipeline tasks and managed using a unified framework within Foundry to ingest and harmonize all incoming COVID-19 EHR data from the participating data partners. Participating data partners can submit their data in one of the five known Common Data Model used by the Clinical and Translational Science Award (CTSA) hubs; OMOP, ACT, TriNetX, PCORnet or PEDSnet. Data partners submit their datasets through a highly secured sFTP location. The N3C DI&H pipeline transforms the submitted data to the OMOP model before merging the dataset to cloud based FedRamp N3C Enclave. The transformation pipeline steps include parsing flat data files, data conformance checks against the native CDM format, primary key checks, domain mapping, and semantic vocabulary translations for all terminologies that exist in the source data, i.e. ICD-10-CM, ICD-10-PCS, LOINC, RxNorm, HCPCS and CPT4.

The THRIVE dataset is submitted in TriNetX CDM format. This dataset can be utilized in N3C to crossmap SDoH observational codes in LOINC or SNOMED to OMOP concept_ids. More specifically, all of the distinct SDoH observational codes submitted in BMC datasets are added to the value set mapping table such that this enriched crosswalk mapping table can be used to translate all other incoming SDoH related codes from other N3C participating data partners. The SDoH concept can be either mapped to the OMOP Observation domain or the Measurement domain. This "map to" information is specified in the domain_id column of the OMOP vocabulary tables. Therefore, based on the domain id, the SDoH observational data are either inserted in the Observation domain or the Measurement domain with translated OMOP codes as observation_concept_id or the measurement_concept_id, respectively. The string value result or the answer to the SDoH observation codes are
mapped to corresponding concept ids and inserted in the value_as_concept_id field along with the value_source_value, the verbatim string value from the source data representing the result/answer of the SDoH concept. By convention, concepts that do not correspond to an existing term in the OMOP vocabulary are added to an instance of the OMOP CDM and assigned a concept ID larger than 2000000.

**Results**

As of July 1, 2021, 76,900 patients including their prior clinical findings since 2018-01-01 were included in the BMC N3C data extract submission. All data were successfully ingested into the N3C enclave. 50,400 (65.5%) had at least one THRIVE SDoH assessment, 49,880 (64.9 %) assessed about homelessness, 21,790 (28.3%) assessed about food insecurity, 20,440 (26.6%) assessed about trouble paying for utilities, and 19,120 (24.9%) assessed about trouble paying for medications. Among respondents, 13.5% were homeless or had unstable housing, 26.4% were experiencing food insecurity, 15.2% reported having trouble paying for utilities, and 13.1% reported having trouble paying for medications.

The N3C enclave is downstream from a number of harmonization and mapping steps, which increases the complexity of provenance tracking, but allows N3C to leverage earlier data cleaning and mapping steps. Data was initially shared as i2b2 data with TriNetX and submitted to N3C in TriNetX CDM format. One of the N3C DI&H pipeline steps is semantic translation from source value sets to OMOP concept ids. The semantic translation utilizes an N3C value set mapping table (analogous to OMOP SOURCE_TO_CONCEPT_MAP) to translate all of the source column field values for a given domain table from the source data to a corresponding OMOP concept_id.

In the TriNetX CDM, the SDoH concept questions are captured in the lab_code column field and the SDoH concept answers are captured in the text_result_value column field. Due to the fact that data partners may collect and submit various enumerations of the answer string for any given SDoH answer concepts in the text_result_val field, N3C opted to extend the OMOP source to concept mapping tables with custom entries in order to harmonize these variations. For example, the text values transferred to N3C included “LOINC:LA30189-7 (I have a steady place to live)” and “LOINC:LA30189-7 (I have housing)”. These LOINC codes were assigned to the parenthetical strings before data were transferred to N3C. While the response “I have housing” is an exact match to the LOINC code LA30189-7, “I have a steady place to live” does not have an assigned LOINC code. When harmonizing data N3C mapped both strings to OMOP concept_id 37079501 (i.e. LOINC:LA30189-7). An alternative strategy would involve late-binding for mappings “I have a steady place to live”, e.g. extending the OMOP Vocabulary to create a custom concept id for “I have a steady place to live” and apply the CONCEPT RELATIONSHIP table at the time of analysis to group these into equivalent analysis concepts. Using either approach, provenance and the original answer text is retained.

In the 2020 version of THRIVE, the screening tool introduced a checkbox for the patient to choose “I do not want to answer the questions” (i.e., “SOMD:31021000119100 (Screening declined)”). In total, 7,530 of the 76,900 patients had elected that they do not voluntarily want to answer the questions. Separately, BMC had incorporated FlowSheet measurement to indicate whether the patient had acknowledged they could not answer the screening tool due to language barrier. Over 1,370 patients had indications of language barriers annotated during transcription. These two types of questions are not about the patient’s SDoH per se, but psychometric assessment markers to provide evidence of validity and reliability of the respondent results to all questions in that screening instance. Through inclusion of these FlowSheet entries, the survey responses of “No” can be clarified as a 1) Patient Endorsement of Answer option “No”, 2) “I don’t understand” as inferred based on the Provider observation about linguistic barrier, 3) a question unintentionally left blank, or 4) “No, I don’t want to answer.” Only in the first option can the responses be taken for immediate value.

**Discussion**

In this report, we describe a pragmatic approach to mapping structured SDoH screening data to standard vocabularies and demonstrate that sharing of these important data - typically stored as non-standard, EHR vendor specific codes - is feasible. National efforts to organize and share SDoH data like the Gravity Project and UCSF Siren are progressing quickly; however, in our review of the literature, we only found a few published studies on standards-based representation of SDoH screening data and none that demonstrated that these standards could be used with real-world data for data sharing between sites using different CDMs, highlighting the novel nature of this work.
There is broad acceptance in clinical domains that screening and intervention related to SDoH should be part of routine clinical care. At present, the practice has not become widespread. Some health systems are driven by an institutional mission to improve care, equity, and address the non-clinical social needs that impact the patients’ well-being. Others have adopted screening methods earlier as part of State or Federal mandates. With the COVID pandemic, patient data on SDoH and data to inform of patient cohorts already experiencing the burden of social needs was in high demand.

As SDoH screening gains broader use nationally, the approach described in this paper could be used for other screening instruments and improve the interoperability of these important data. For sites that serve traditionally marginalized or under-resourced populations such as the health system in this project, capturing these data and being able to include them in research has the potential to more fully describe the life experience and health determinants for their patients. When mapped to standard terminologies, these data can be shared on a national scale to help ensure that the broadest possible array of people and their data are represented in national collaborations, like N3C.

Several important challenges were encountered during this project. First, the THRIVE screener evolved over the first three years of use. Versioning metadata was not available in the submitted data set. In several cases, only the answer values changed even though the question text had been edited and the variable identifier remained the same. Correctly mapping values was possible, but required meeting with the developers and clinical teams to better understand the changes, and selecting among inexact matches. In the example shared above, the THRIVE instrument includes a single item that is almost identical to the PRAPARE question which has a pre-assigned LOINC/OMOP code (“What is your living situation today?”). However, rather than presenting responses with three possible values for housing status in THRIVE, PRAPARE has two sequential questions and LOINC codes. The second question asks separately in a Yes/No format “Are you worried about losing your housing?” (https://forms.loinc.org/93025-5), whereas THRIVE simply presents “I have a place today, but I am worried…”. While this PRAPARE question and their “Yes” and “No” responses have standard LOINC codes, answers are not meaningful without the context of a question. In other words, the multiple-choice option in THRIVE is represented as two separate questions in PRAPARE, thus in order to correctly analyze and harmonize data, both question and answer values must be considered in the mapping process. This changes the standard architecture of SOURCE_TO_CONCEPT_MAP (or CONCEPT_RELATIONSHIP) to require joining two fields in the source data (question and response values) to properly assign the adjudicated value of “at risk” record. Notably, LOINC now includes richer representation of instruments, including information about the instrument, the questions, and the answers for survey data. The OMOP common data model includes “Standardized Derived Data Elements” where there is consensus on complex but consistent logical algorithms for deriving data elements from standardized facts. A similar approach for logical representations may be useful to derive common data elements from heterogeneous survey-based data elements measuring the same or similar concepts using distinct questions and answers.

One limitation of this study is that it reflects one site's approach to collecting the information. Several steps may be unique to the site, though the insights learned may benefit other sites addressing data silos and patient-level SDoH data representations. The ability to map to LOINC codes improved the continuity of the information and it didn't appear affected by the multiple transformation steps. This case study applied concept mappings post hoc to extract data that has already been collected into FlowSheets, where the mapping decisions may vary by site and require local technical expertise. FlowSheet records that produce derived summary scores and panel indices, assessments requiring clinical context to understand their non-random occurrence or missingness, and assessment observations based on conditional responses (e.g., probing questions) represent limitations to the FlowSheet extraction presented. As the source vocabularies change, updating the concept mappings will be a sustainability and versioning hurdle. To reduce site variability in FlowSheet mapping decisions, a better approach would be to have metadata maps designed, vetted, and maintained by the instrument developers, such that downstream instrument adopters can incorporate the instruments and compare across sites with fewer technical and interpretation barriers. Our workflow focused on data harmonization into OMOP where other CDM endpoints may be desired. We acknowledge that instrument registration, secondary data-use consideration, and harmonization as part of the data life-cycle may be out-of-scope for instrument developers. However, this remains a major challenge in summarizing data across instruments as many survey instruments do not have controlled vocabulary representations. We encourage instrument developers towards this early strategic planning as it enables downstream comparisons and analyses.

In summary, there is a pressing need to better understand and include SDoH in research and clinical data repositories. There are heterogeneous mechanisms for data capture and standardization, which may result in
duplicate efforts where data are captured with commonly used instruments that are not standardized at the time of data capture or ETL. While this case-study was a targeted, pragmatic intervention to extract patient-level SDoH from FlowSheet records, this approach was intended to be scalable and reusable, though not comprehensive. The vast majority of EHR implementations have their own bespoke survey implementations in flowsheets, yet extraction remains a laborious and technical endeavor. Within the well-described domain of structured screeners like PRAPARE, AHC, and THRIVE, mapping locally represented data to standards is feasible and can be used within a variety of CDMs like OMOP, i2b2, and TriNetX. And, here, we’ve shown the utility to translate across these CDMs. As routine screening increases and more data become available, the approach described in this paper as well as the work of many others should be used to ensure that these important data can be shared within and between systems. Understanding the value of undertaking this effort is a critical next step to convincing health systems to invest time and effort in doing so.

**Funding Acknowledgements**

Jimmy Phuong was partially funded by the National Center for Data to Health (CD2H) grant [NIH/NCATS U24TR002306], supplemental funding from the National COVID Cohort Collaborative [NIH/NCATS U24TR002306-04S3], Juan Espinoza and Daniella Meeker have received funding from grants UL1TR001855 and UL1TR000130, and a subaward from grant U24TR002306 from the National Center for Advancing Translational Science (NCATS) of the U.S. National Institutes of Health. William Adams has received funding from UL1TR001430. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

**Bibliography**


28. UCSF Social Interventions Research & Evaluation Network (SIREN) [Internet]. Social Needs Screening Tool Comparison Table [Internet]. 2020 [cited 2021 Mar 10]. Available from: https://sirenetwork.ucsf.edu/SocialNeedsScreeningToolComparisonTable
Testing a filtering strategy for systematic reviews: evaluating work savings and recall

Randi Proescholdt¹, Tzu-Kun Hsiao¹, Jodi Schneider¹, Aaron M. Cohen², Marian S. McDonagh², Neil R. Smalheiser³
¹University of Illinois at Urbana-Champaign, Champaign, IL; ²Oregon Health & Science University, Portland, OR; ³University of Illinois at Chicago, Chicago, IL

Abstract

Systematic reviews are extremely time-consuming. The goal of this work is to assess work savings and recall for a publication type filtering strategy that uses the output of two machine learning models, Multi-Tagger and web RCT Tagger, applied retrospectively to 10 systematic reviews on drug effectiveness. Our filtering strategy resulted in mean work savings of 33.6% and recall of 98.3%. Of 363 articles finally included in any of the systematic reviews, 7 were filtered out by our strategy, but 1 “error” was actually an article using a publication type that the SR team had not pre-specified as relevant for inclusion. Our analysis suggests that automated publication type filtering can potentially provide substantial work savings with minimal loss of included articles. Publication type filtering should be personalized for each systematic review and might be combined with other filtering or ranking methods to provide additional work savings for manual triage.

Introduction

Systematic reviews (SRs) are extremely time-consuming; an average SR takes 67 weeks¹ and costs about $141,000² in staff time. A variety of machine learning approaches are being examined to assist SR teams, often focused on prioritizing the records retrieved³,⁴ or reducing the need or extent of dual screening⁵,⁶. The time required is correlated with the number of records requiring manual triage of titles and abstracts for apparent relevance⁷. Hence, a key goal is to reduce the number of records that a SR team initially needs to examine while preserving recall, as close to 100% as possible. The goal of this work is to evaluate the potential of a particular strategy, using publication type and study design filters for automatic filtering of articles for contributing to automation of SRs.

In this paper, we tested our filtering strategy retrospectively against 10 previously completed SRs about comparative drug effectiveness. Our strategy uses two machine learning models, Multi-Tagger⁸ and web RCT Tagger⁹, in combination with National Library of Medicine (NLM)’s MeSH terms and publication types in order to retain as many relevant articles as possible, while reducing the number of articles needing manual screening. The models have previously been evaluated using information retrieval measures, but need further evaluation in order to gain the trust of systematic reviewers¹⁰ and to estimate the potential work savings in real-life situations. The 10 SRs used to evaluate this strategy came from the Drug Effectiveness Review Project (DERP). DERP is a collaboration of state Medicaid agencies that commission SRs aimed to help inform decisions about the drugs that would be available to Medicaid recipients in each state.

Methods

We included a series of SRs from 2003-2018 conducted for DERP by the Pacific Northwest Evidence-based Practice Center at Oregon Health & Science University (OHSU)¹¹. For each of the 10 DERP reports used in this analysis, we received information, such as the study designs each review planned to include and reference libraries containing records for the citations screened, including decisions on inclusion in the SR. For each of the 10 SRs being studied, our evaluation calculated the work savings (i.e., the number of articles in the initial retrieval set that were filtered out by our strategy, divided by the number of articles in the initial retrieval set) and the recall (the number of articles finally included in the SR that passed by our filtering strategy, divided by the total number finally included in the SR). The filtering strategy is shown in Table 1. We retained abstracts if any of rules 1-3 applied. For rules 1 and 2, we checked automated publication type predictive scores from the Multi-Tagger⁸ and web RCT Tagger⁹ against designated thresholds which optimally balanced precision and recall (i.e., rather than optimizing recall alone, we chose the threshold which gave the highest F1: any article receiving a score below the threshold was filtered out, and any article equal to or above the threshold was retained). For rule 3, we retrieved NLM’s MeSH indexing. If an included design was found in the MeSH terms or publication types, the article was retained.
Table 1: Abstract filtering rules. Items were retained if any of rules 1-3 applied, and filtered out otherwise.

<table>
<thead>
<tr>
<th>Rule Number</th>
<th>Source of Publication Type or Study Design Information</th>
<th>Tags and MeSH terms relevant to study designs over the 10 reviews</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 1</td>
<td>Multi-Tagger</td>
<td>Case-Control Studies</td>
<td>Above threshold that gave the optimal F1(^2); or item not processed by Multi-Tagger (i.e., article not in English or lacking abstract)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clinical Study</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cohort Studies</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Meta-analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Practice Guideline</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Prospective Studies</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Randomized Controlled Trial</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Retrospective Studies</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Systematic Review</td>
<td></td>
</tr>
<tr>
<td>Rule 2</td>
<td>Web RCT Tagger</td>
<td>Randomized Controlled Trial</td>
<td>Above 0.01 threshold for RCTs(^3); or item not processed by Web RCT Tagger (i.e., article not in English or lacking abstract)</td>
</tr>
<tr>
<td>Rule 3</td>
<td>MeSH Terms and Publication Types</td>
<td>Case-Control Studies</td>
<td>One or more relevant study design terms were applied in NLM MeSH indexing.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clinical Study</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cohort Studies</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clinical Trial</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Meta-Analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Observational Study</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Practice Guideline</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Prospective Studies</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Randomized Controlled Trial</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Retrospective Studies</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Systematic Review</td>
<td></td>
</tr>
</tbody>
</table>

From each SR (summarized in Table 2), we used the following data: the list of study designs that the review stated as relevant for inclusion; PMIDs titles and abstracts screened manually (i.e., triage), PMIDs full-text screened, and PMIDs included in the final review. We customized the list of relevant study designs from Table 1 to each review, as shown in Table 2.

Not all of the study designs each SR listed as relevant for inclusion had direct matches to tags in Multi-Tagger. In such cases, we applied related tags we deemed likely to be relevant. Since there is no single Multi-Tagger score encompassing all comparative observational studies, we applied the following tags and MeSH terms for SRs that listed observational studies as relevant for inclusion: Cohort Studies, Case-Control Studies, Retrospective Studies, Prospective Studies, and Clinical Study. To ensure high recall, we also applied Observational Study as a MeSH term for all reviews that included observational studies and Clinical Trial as a MeSH term for reviews that included randomized controlled trials. Additionally, we applied Systematic Review, Meta-Analysis, and Practice Guideline as tags and MeSH terms for all 10 reviews, since the DERP team sought and reviewed the full-text of articles with these designs in order to help identify any articles missed by the original search.
Table 2: The 10 reviews from DERP, their included study designs, and the corresponding Multi-Tagger study designs.

<table>
<thead>
<tr>
<th>#</th>
<th>SR Name</th>
<th>Study Designs Eligible for Inclusion in SR</th>
<th>Multi-Tagger Tags Applied</th>
</tr>
</thead>
</table>
| 1 | Anticoagulants-Original-Report | 1. Head-to-head or active-controlled randomized trials  
2. Systematic reviews  
3. Cohort or case-control observational studies | -Case-Control Studies  
- Clinical Study  
- Cohort Studies  
- Meta-analysis  
- Practice Guideline  
- Prospective Studies  
- Randomized Controlled Trial  
- Retrospective Studies  
- Systematic Review |
| 2 | Asthma-COPD              | 1. Head-to-head randomized controlled clinical trials  
2. Comparative systematic reviews  
3. Comparative observational studies | -Case-Control Studies  
- Clinical Study  
- Cohort Studies  
- Meta-analysis  
- Practice Guideline  
- Prospective Studies  
- Randomized Controlled Trial  
- Retrospective Studies  
- Systematic Review |
| 3 | Benzodiazepines-Summary-Review | 1. Systematic reviews | -Meta-analysis  
- Practice Guideline  
- Systematic Review |
| 4 | Hepatitis-C-Update-2     | Best evidence available from:  
1. Head-to-head randomized controlled trials  
2. Observational studies  
3. Systematic reviews  
4. Other designs (e.g., pooled analyses) | -Case-Control Studies  
- Clinical Study  
- Cohort Studies  
- Meta-analysis  
- Practice Guideline  
- Prospective Studies  
- Randomized Controlled Trial  
- Retrospective Studies  
- Systematic Review |
| 5 | Long-Acting-Insulins     | 1. Head-to-head randomized controlled trials  
2. Comparative observational studies  
3. Systematic reviews | -Case-Control Studies  
- Clinical Study  
- Cohort Studies  
- Meta-analysis  
- Practice Guideline  
- Prospective Studies  
- Randomized Controlled Trial  
- Retrospective Studies  
- Systematic Review |
| 6 | Long-Acting-Opioids-Update-7 | 1. Head-to-head controlled clinical trials  
2. Comparative systematic reviews  
3. Comparative observational studies | -Case-Control Studies  
- Clinical Study  
- Cohort Studies  
- Meta-analysis  
- Practice Guideline  
- Prospective Studies  
- Randomized Controlled Trial  
- Retrospective Studies  
- Systematic Review |
| 7 | MS-Drugs-Update-3        | 1. Head-to-head controlled clinical trials  
2. Placebo-controlled trials  
3. Comparative observational studies  
4. Comparative systematic reviews | -Case-Control Studies  
- Clinical Study  
- Cohort Studies  
- Meta-analysis |
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
</table>
| 8 | Newer-Diabetes-Meds-Update-2 | 1. Head-to-head randomized controlled trials  
2. Head-to-head prospective cohort studies  
3. Case-control studies |
| 9 | PCSK9 | 1. Controlled clinical trials  
2. Systematic reviews  
3. Comparative observational studies |
| 10 | Second-Generation-Antipsychotics-Update-5 | 1. Head-to-head randomized controlled trials  
2. Placebo-controlled trials  
3. Comparative systematic reviews  
3. Comparative observational studies with a concurrent control group |

For each SR, we tabulated: a) the number of articles in the initial retrieval set (i.e., the actual search retrieved by the SR team in preparing their report); b) the number of articles filtered out using the strategy just described (i.e., the number of articles that the SR team actually screened but would not have if they had used our strategy); c) the percent work savings; d) the number of articles that DERP finally included in their final SR report (i.e., the number of articles actually included in the SR, based on our assumption that the SR team’s actual results included the ideal set of articles); e) the number of finally included articles that were lost using the strategy just described; f) the percentage recall. These statistics are shown in Table 3. (In our analysis, we did not analyze whether our filtering strategy could have resulted in additional relevant articles for final inclusion in the SR reports.)

Our error analysis examined each article that was filtered out by our PT strategy but included in the final SR. We examined the model predictive scores and MeSH terms to understand why the article was filtered out, as well as an assessment of its publication type based on documentation in the DERP reference library, the article’s metadata, and the article’s full-text. We also assessed whether it met the SR’s original inclusion criteria in terms of study design.
Results

Table 3. Summary statistics.

<table>
<thead>
<tr>
<th>DERP Report</th>
<th># in initial retrieval set</th>
<th># filtered out by our strategy</th>
<th>% work savings</th>
<th># of included articles</th>
<th># of included articles removed by our strategy</th>
<th>% recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anticoagulants-Original-Report</td>
<td>1766</td>
<td>659</td>
<td>37.32</td>
<td>82</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Asthma-COPD</td>
<td>1964</td>
<td>497</td>
<td>25.31</td>
<td>28</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Benzodiazepines-Summary-Review</td>
<td>581</td>
<td>302</td>
<td>51.98</td>
<td>12</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Hepatitis-C-Update-2</td>
<td>4917</td>
<td>1417</td>
<td>28.82</td>
<td>75</td>
<td>2</td>
<td>97.33</td>
</tr>
<tr>
<td>Long-Acting-Insulins</td>
<td>1086</td>
<td>301</td>
<td>27.72</td>
<td>37</td>
<td>1</td>
<td>97.3</td>
</tr>
<tr>
<td>Long-Acting-Opioids-Update-7</td>
<td>503</td>
<td>60</td>
<td>11.93</td>
<td>13</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>MS-Drugs-Update-3</td>
<td>1849</td>
<td>825</td>
<td>44.62</td>
<td>45</td>
<td>3</td>
<td>93.33</td>
</tr>
<tr>
<td>Newer-Diabetes-Meds-Update-2</td>
<td>1065</td>
<td>400</td>
<td>37.56</td>
<td>21</td>
<td>1</td>
<td>95.24</td>
</tr>
<tr>
<td>PCSK9</td>
<td>75</td>
<td>32</td>
<td>42.67</td>
<td>13</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Second-Generation-Antipsychotics-Update-5</td>
<td>1110</td>
<td>314</td>
<td>28.29</td>
<td>37</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 4. List of included articles filtered out.

<table>
<thead>
<tr>
<th>DERP Report Name</th>
<th>PMID</th>
<th>Title</th>
<th>Actual study design</th>
<th>Reason filtered out</th>
<th>Error or exception?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hepatitis-C-Update-2</td>
<td>16267758</td>
<td>Risk factors for perinatal transmission of hepatitis C virus (HCV) and the natural history of HCV infection acquired in infancy.</td>
<td>Cohort Study</td>
<td>Cohort studies predictive score below threshold</td>
<td>Error</td>
</tr>
<tr>
<td>Hepatitis-C-Update-2</td>
<td>22813094</td>
<td>Comparison of current US risk strategy to screen for hepatitis C virus with a hypothetical targeted birth cohort strategy.</td>
<td>Comparative study; birth cohort strategy</td>
<td>Cohort studies predictive score below threshold</td>
<td>Error</td>
</tr>
<tr>
<td>Long-Acting-Insulins</td>
<td>22966091</td>
<td>Does insulin glargine increase the risk of cancer compared with other basal insulins?: A French nationwide cohort study based on national administrative databases.</td>
<td>Cohort Study</td>
<td>Cohort studies predictive score below threshold</td>
<td>Error</td>
</tr>
<tr>
<td>MS-Drugs-Update-3</td>
<td>19936821</td>
<td>Parenthood and immunomodulation in patients with multiple sclerosis.</td>
<td>Cohort Study</td>
<td>Cohort studies predictive score below threshold</td>
<td>Error</td>
</tr>
</tbody>
</table>
Acting on these initial assumptions, we found that our filtering strategy resulted in work savings ranging from 11.9% to 52.0% (mean 33.6%) and recall of 93.3% to 100% (mean 98.3%). We examined the 7 articles finally included in any of the SRs but filtered out by our strategy, and found that 1 of these “errors” was actually an article whose publication type that the SR team had not pre-specified as relevant for inclusion (Table 6). Because this exclusion is not a true error, we recalculated recall by dividing the number of true errors in each SR by the number of total included articles. The recalculated recall of our filtering strategy ranges from 95.24% to 100% (mean 98.5%). Five of the remaining six errors occurred because the score for one particular article type, Cohort Studies, was below our chosen threshold. Had we adjusted the threshold for Cohort Studies down to 0.02, we would have achieved slightly better recall (i.e., range of 93.3% to 100%; mean 99.1%), with minimal loss of work savings (range of 11.1% to 52.0%; mean 31.6%). The recalculated recall using only the true errors AND using the lowered threshold for Cohort Studies results in recall ranging from 95.6% to 100% (mean 99.3%).

**Discussion**

In the present study, we tested the hypothesis that one can achieve substantial work savings and near-perfect recall using a publication type filtering strategy for automated triage that was applied retrospectively to 10 Drug Effectiveness Review Project SRs. Using predictive scores from Multi-Tagger, we initially set thresholds for filtering out articles based on an optimal balance between precision and recall; future work could also consider personalized thresholds to optimize for high recall for particular study designs. The results are very encouraging and will inform our plans to implement publication type filtering prospectively during the creation of new SRs by a variety of teams.

Another opportunity for future work is to analyze whether this filtering strategy may potentially identify additional articles not found by the SR team. Because our focus was work savings and recall based on the actual SR results, we did not consider whether this filtering strategy may be better in some ways than human searching. For example, the strategy could enable an SR team to start from a larger initial set of articles that they would not have had the resources to screen manually. Additional analysis is needed to understand whether such a strategy would result in additional articles relevant for inclusion while still resulting in work savings. In the present study, we assumed that the actual SR results were the ideal set of articles.

Our study has several limitations. First, we only included articles that have PMIDs in our analysis due to the availability of MeSH terms and publication types in PubMed metadata. Additionally, because the study was retrospective, we were limited in our understanding of the SR process, such as the context of the inclusion of some articles. Past research on reducing workload in reviews has noted that different topics exhibit different work savings and recall. We did not examine the role of topic other than to note that different SRs varied in the number and kind of article types that they deemed relevant, which could certainly impact on the performance of our strategy. Additional

<table>
<thead>
<tr>
<th>Article ID</th>
<th>Title</th>
<th>Type</th>
<th>Tagger</th>
<th>Exception</th>
</tr>
</thead>
<tbody>
<tr>
<td>24463630</td>
<td>Pregnancy outcomes in the clinical development program of fingolimod in multiple sclerosis.</td>
<td>Clinical Study</td>
<td>Clinical Study predictive score below threshold</td>
<td>Error</td>
</tr>
<tr>
<td>25300980</td>
<td>Drug utilization, safety, and effectiveness of exenatide, sitagliptin, and vildagliptin for type 2 diabetes in the real world: data from the Italian AIFA Anti-diabetics Monitoring Registry.</td>
<td>Retrospective cohort study</td>
<td>Cohort studies predictive score below threshold</td>
<td>Error</td>
</tr>
</tbody>
</table>
SRs conducted by a variety of teams need to be analyzed in order to ascertain the most appropriate predictive score thresholds that will ensure maximal recall while still providing substantial work savings. Prioritizing screening based on the tagger score, rather than pre-specifying a threshold, should also be tested in future work. Some studies designs often include a variety of types, and those with less common designs can be missed by the tagger. For example, two database studies (PMID 22966091; PMID 25300980) and a birth cohort study (PMID 22813094) that were filtered out lacked the MeSH Cohort Studies term and may have received low Cohort Study Multi-Tagger predictions because their characteristics are not typical of Cohort Studies. One of the DERP reports used a “Best Evidence” approach, in which some designs (e.g., Randomized Controlled Trials) were prioritized over others, and some designs (including some not explicitly stated in the inclusion criteria) were considered if and when articles using the prioritized designs were not found. Our initial strategy did not account for this approach. Our choices described here, regarding which observational study designs to include, were somewhat arbitrary and across-the-board; however, our findings suggest that the list of included study designs should be expanded or refined to optimize results for individual SRs.

**Conclusion**

In order to apply the Multi-Tagger tool realistically in the workflow of a SR team, we suggest careful consideration of what article types might potentially be relevant but are often omitted from explicit inclusion. Automated publication type filtering may also be useful for other types of evidence syntheses such as rapid reviews\(^1\) and scoping reviews\(^1\). As well, publication type filtering should optimally be combined with other filtering or ranking methods\(^1\) that may provide additional work savings at the manual triage stage. In the future, we plan to provide web-based tools for anyone to obtain predictive publication type scores for articles not indexed by PubMed (i.e., indexed in databases such as EMBASE or PsycINFO).

**Data Availability**

Data is publicly available at http://doi.org/10.13012/B2IDB-9257002_V1

**Acknowledgements**

Supported by NIH grant 5R01LM010817. Thanks to Halil Kilicoglu for comments on the manuscript.

**References**


Multi-Site Testing of an Opioid Prescribing Electronic Clinical Quality Measure Following Elective Primary Total Hip and/or Total Knee Arthroplasties

Avery Pullman, BS\(^1\), Mica Curtin-Bowen, BA\(^1\), Ania Syrowatka, PhD\(^{1,2}\), Alexandra Businger, MPH\(^1\), Michael Sainlaire, MS\(^1\), Stuart Lipsitz, ScD\(^{1,2}\), Tien Thai, BS\(^1\), Troy Li, BS\(^1\), David W. Bates, MD, MSc,\(^{1,2}\) Patricia C. Dykes, RN, PhD\(^{1,2}\)

\(^1\)Brigham and Women’s Hospital, Boston, MA; \(^2\)Harvard Medical School, Boston, MA

Abstract

As the United States faces the third wave of the ongoing opioid epidemic, development of measures which report on prolonged opioid prescribing (POP) rates, specifically following orthopedic surgeries, are needed to better understand and improve prescribing practices at the clinician group level. Brigham and Women’s Hospital (BWH) has been contracted by the Centers for Medicare and Medicaid Services (CMS) to create a novel electronic clinical quality measure (eCQM) to quantify the prolonged opioid prescribing rate of opioid episodes lasting > 42 days in patients aged 18+ years following elective primary total hip arthroplasties (THA) and/or total knee arthroplasties (TKA) for use in the Merit-Based Incentive Payment System (MIPS). When this measure was tested on two geographically distinct sites, it was found that the THA rate was 3.80% and 16.07% at sites 1 and 2, respectively, and that the TKA rate is 7.65% and 24.15% at sites 1 and 2, respectively. This manuscript reports on the testing of this eCQM between these two sites, highlighting differences in state and organizational level policies regarding opioid prescribing and documentation practices.

Background and Significance

Total hip arthroplasty (THA) and total knee arthroplasty (TKA) are the two most common joint replacement procedures in the United States; in 2019 alone, 91,814 THAs and 139,991 TKAs were performed in the U.S.\(^1\) Opioids are frequently prescribed for pain management following joint arthroplasty procedures, recent literature has shown that in 2017, 89.7% of THA and 91.5% of TKA patients received an opioid prescription within 60 days of discharge.\(^2\)

Prescription opioid use following total joint arthroplasty (TJA), although common and helpful for relieving short-term pain, also carries risks, including opioid poisoning (overdose), adverse interactions with other substances, respiratory depression, and death.\(^3\) As the United States continues to experience the wide-ranging impacts of the opioid epidemic, concern about prescription opioid use has evolved beyond the medical field and is an issue in law enforcement and the public. The Center for Disease Control (CDC) estimates that in 2017, healthcare providers wrote 58.5 opioid prescriptions per 100 persons.\(^3\)

The American Academy of Orthopedic Surgeons has released clinical practice guidelines to standardize the post-operative prescription regiment for total joint arthroplasties (TJA).\(^4\) Although the strict follow through of these guidelines has been shown to result in lower post-operative opioid consumption with no significant changes in postoperative pain control,\(^5\) variation in TJA-related opioid prescribing practices remains high.\(^6\)

While prescribing a higher quantity of opioids after orthopedic surgeries may not afford significantly more pain relief than a lower quantity, orthopedic surgeons routinely prescribe more post-operative opioids than most other surgical specialties.\(^5,7\) The objective of this measure is to provide a tool to assess the prolonged opioid prescribing (POP) rate in patients without previous opioid exposure following elective primary THA/TKA using routinely collected electronic health record (EHR) data to help drive high quality, evidence-based care. This electronic clinical quality measure (eCQM) leverages opioid prescription and patient demographic information from EHR systems to compile a risk-adjusted rate of patients who are prescribed opioids for > 42 days following elective primary THA and/or TKA procedures.
The rate, measured at the clinician group level, is expressed as a percentage where a lower POP percentage is indicative of higher quality care. This measure is intended for use in the Merit-Based Incentive Payment System (MIPS) within the Centers for Medicare and Medicaid Services (CMS) which provides physicians with financial incentives for improvements in quality care. This manuscript builds upon previous work which developed and tested this measure at a single site, by now testing the measure at two geographically distant healthcare systems with distinct patient populations.

Materials and Methods

Alpha and beta testing were conducted at two geographically distant large healthcare systems (referred to as ‘Site 1’ and ‘Site 2’) using EHR data. Alpha testing was comprised of reliability and validity testing to assess the feasibility of implementing the eCQM into EHR systems, and beta testing was conducted to assess the unadjusted and risk-adjusted POP rate across both sites. Site 1 used the EHR vendor ‘Epic’ from six clinician groups from 2016-2019, and Site 2 used the EHR vendor ‘Cerner’ from eight clinician groups from 2017-2019. This measure is stratified, meaning that separate POP rates were calculated for THA and TKA patient populations to reflect the differences in postoperative prescribing practices by procedure type. The same methodology is used for both THA and TKA POP rates.

During measure development, the POP eCQM was defined using the following criteria:

**Numerator:** The subset of patients from the denominator who were prescribed post-operative opioids for > 42 days after surgical discharge following an elective primary THA or TKA.

**Denominator:** All patients, aged 18 years or older, who received an elective primary THA or TKA procedure and do not meet any exclusion criteria.

**This eCQM excludes patients if:**

- The patient was prescribed opioids within the 90 days prior to the index admission
- The patient received a diagnosis of Opioid Use Disorder within the 365 days prior to the index admission
- The patient had a Cancer diagnosis within the 365 days prior to the index admission or 90 days following discharge
- The patient had a diagnosis of Sickle Cell Disease within the 365 days prior to the index admission or 90 days following discharge
- The patient received hospice or palliative care within the 365 days prior to the index admission or 90 days following discharge
- The patient was discharged against medical advice (AMA)
- The patient received a separate THA- or TKA-related procedure within the 90 days prior to the index admission or 90 days after hospital discharge
- The patient had more than two THA or TKA procedure codes were documented during the index hospitalization
- The patient had an additional surgery within 90 days following discharge from their THA/TKA

**Alpha Testing:**

Alpha testing involved analysis of data from both sites and consisted of reliability, feasibility, and validity testing. Reliability and feasibility were tested by examining the availability of the data elements required for both the eCQM implementation and risk-adjustment across EHR systems. Using the National Quality Forum (NQF) feasibility scorecard, the data element’s reliability and feasibility were quantified.

Validity was examined at both sites by performing multiple rounds of chart reviews, considered the “gold standard”, on random samples of patients who met the inclusion and exclusion criteria of the measure. The gold standard results were then compared to those produced by the eCQM and a Kappa value was calculated to quantify the level of agreement. Any instances of disagreement were documented and analyzed by the development team.
Beta Testing:

Beta testing at both sites examined patient demographics and determined the unadjusted and risk-adjusted results of the eCQM at the clinician group level. The risk-adjustment model for this measure is harmonized with the NQF1550: Risk Standardized Complication Rate and risk adjusts for age, sex, number of procedures (1-2), comorbid conditions, language, race, zip code, smoking status, and Body Mass Index (BMI). To risk-adjust, data was first randomly split into a test and validation sample (50% in each sample) and fit to the same generalized linear mixed model (with same covariates) in both samples. The demographics of the test and validation samples were compared using P-values to explore similarities between the groups.

The predicted over expected (P/E) ratio, which compares the “predicted” number of patients in the numerator to the “expected” number of patients at a given clinician group was then calculated for all test and validation samples. Using a hierarchical logistic regression, group-specific intercepts were modeled arising from a normal distribution. The clinician group intercept represents the likelihood of a numerator event for procedures performed by the given group after accounting for patient risk factors present in the model.

A Spearman correlation coefficient was used to assess the agreement in POP rates between clinician groups in test and validation samples. The risk-adjusted POP rate at the clinician group level was then determined by multiplying the predicted/expected ratio by the unadjusted POP rate from the corresponding clinician group. 95% confidence intervals were calculated for each clinician group.

Calculation of a C-statistic, and a Hosmer-Lemeshow test were used to assess the adequacy of the model. The C-statistic was used to determine similarities between test and validation samples. The Hosmer-Lemeshow test assessed the goodness of fit for the logistic regression model used for risk adjustment.

Results:

Reliability Testing:

Following NQF Feasibility Scorecard calculation, it was determined that 100% of the data elements used to define and calculate the measure were feasible, receiving the highest scores for each section of the scorecard. Total elements needed for risk-adjustment were available for 99.26% of variables in Site 1, and 86.99% of variables in Site 2. Site 1 patient, clinician, and procedure information can be found in table 1, and the complete frequency of data elements for risk adjustment in both samples can be found in Table 2 below.

Table 1: Site 1 and 2 patient demographic information (n = 17,372 (site 1), n = 4,593 (site 2))

<table>
<thead>
<tr>
<th>Variable</th>
<th>Site 1: 2016-2019</th>
<th>Site 2: 2016-2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Admissions</td>
<td>17,372</td>
<td>4,593</td>
</tr>
<tr>
<td>Number of Patients Included in the Measure</td>
<td>11,636</td>
<td>2,126</td>
</tr>
<tr>
<td>Number of Patients Excluded from Measure</td>
<td>5,698</td>
<td>2,446</td>
</tr>
<tr>
<td>Number of Clinician Groups</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Number of Numerator Patients</td>
<td>1,006</td>
<td>422</td>
</tr>
<tr>
<td>THA procedure</td>
<td>42.79%</td>
<td>42.29%</td>
</tr>
<tr>
<td>TKA procedure</td>
<td>57.19%</td>
<td>57.71%</td>
</tr>
</tbody>
</table>
Smoking status was missing for the vast majority of Site 2 patients, with the information only being available in 1.05% (n=48) of the total sample. As a result, smoking status was excluded from the risk-adjustment model for this measure. All other risk-adjustment variables were available for use in the model (e.g., less than 10% missing). The impact of missing smoking status is elaborated upon in the discussion.

In addition to the availability of risk-adjustment variables, the eCQM development team analyzed the frequency of data element availability for opioid prescription information in the included (denominator) sample. Opioid prescription information analysis was limited to the included sample due to the data provided from site 2, which did not include opioid information on excluded patients. This information was used to calculate the total opioid episode; e.g., the time from the start of the first prescription until the end of the final prescription. The total opioid episode is based on the days’ supply (DS) of opioids, which can be entered in the EHR, or calculated using the prescription’s dosage (D), quantity (Q), and frequency (F) information. If any of the D, Q, or F information is missing, the DS cannot be calculated for a patient. In Site 1, the DS as well as the D, Q, and F of opioid prescriptions were consistently populated across the included sample with 100% (n=11636) of the information available to calculate the total opioid episode for each patient. In comparison, opioid prescription information was less complete in Site 2. The DS information was available for 66.37% of patients (n=1411). The opioid prescription D, Q, and F elements had a higher availability than the DS (72.39%, 79.77%, 78.22%, respectively). In total, PNP rates were able to be calculated for 100% of the denominator sample in Site 1, and 68% of the denominator sample at Site 2, due to the less frequent standardized documentation recorded at this site. The impact of missing data, as well as the analysis of missing data, is expanded upon in the discussion.

**Validity testing:**

Chart reviews using a random sample of Site 1 patients from 2016-2018 were conducted to assess the data agreement between manual chart reviews and the EHR calculation from the eCQM. Three rounds of chart reviews, examining 50 patients each round, were conducted and produced a final Kappa value of 1.0, with 96.67% agreement (95% CI = 94.08, 100). A fourth round of chart reviews were conducted on a random sample of patients who met the exclusion criteria from Site 1, resulting in 100% agreement (95% CI = 88.28, 100) and a Kappa coefficient of 1.0. A round of chart reviews was conducted by Cerner personnel on Site 2 patients. The chart review of 30 patients (10 numerator, 10 denominator, 10 exclusion) produced 100% agreement with a final kappa coefficient of 1.0. A Kappa coefficient of 1.0 demonstrates that this measure has excellent agreement between manual and EHR review, indicative of strong validity.

**Beta testing:**

Site 1 was comprised of six clinician groups with a total sample of 17,372 THA and/or TKA patients. Site 2 was comprised of eight clinician groups with a total sample of 4,593 THA and/or TKA patients. Patients had a median

---

**Table 2: Site 1 & 2 data element availability**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Site 1 Frequency</th>
<th>Site 1 Missing</th>
<th>Site 1 Total</th>
<th>Site 1 Percentage</th>
<th>Site 2 Frequency</th>
<th>Site 2 Missing</th>
<th>Site 2 Total</th>
<th>Site 2 Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Risk Adjustment Information</strong> (n = 17,332 (Site 1), n = 4,553 (Site 2))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insurance Type</td>
<td>17.305</td>
<td>1.7</td>
<td>17.322</td>
<td>99.9%</td>
<td>4,553</td>
<td>0</td>
<td>4,553</td>
<td>100%</td>
</tr>
<tr>
<td>Body Mass Index (BMI)</td>
<td>17.202</td>
<td>12.0</td>
<td>17.322</td>
<td>99.31%</td>
<td>4,170</td>
<td>383</td>
<td>4,553</td>
<td>91.59%</td>
</tr>
<tr>
<td>Primary Language</td>
<td>17.149</td>
<td>17.3</td>
<td>17.322</td>
<td>99%</td>
<td>4,526</td>
<td>27</td>
<td>4,553</td>
<td>99.41%</td>
</tr>
<tr>
<td>Smoking Status</td>
<td>17.068</td>
<td>25.4</td>
<td>17.322</td>
<td>98.53%</td>
<td>48</td>
<td>4505</td>
<td>4,553</td>
<td>1.05%</td>
</tr>
<tr>
<td>Zip code</td>
<td>17.304</td>
<td>18</td>
<td>17.322</td>
<td>99.9%</td>
<td>4,137</td>
<td>416</td>
<td>4,553</td>
<td>90.86%</td>
</tr>
<tr>
<td>Sex</td>
<td>17.322</td>
<td>0</td>
<td>17.322</td>
<td>100%</td>
<td>4,553</td>
<td>0</td>
<td>4,553</td>
<td>100%</td>
</tr>
<tr>
<td>Race</td>
<td>16.928</td>
<td>39.4</td>
<td>17.322</td>
<td>97.73%</td>
<td>4,552</td>
<td>1</td>
<td>4,553</td>
<td>99.98%</td>
</tr>
<tr>
<td>Admit Age</td>
<td>17.322</td>
<td>0</td>
<td>17.322</td>
<td>100%</td>
<td>4,553</td>
<td>0</td>
<td>4,553</td>
<td>100%</td>
</tr>
<tr>
<td>Condition</td>
<td>17.144</td>
<td>17.8</td>
<td>17.322</td>
<td>98.97%</td>
<td>4,552</td>
<td>1</td>
<td>4,553</td>
<td>99.98%</td>
</tr>
<tr>
<td>Total Element Availability</td>
<td></td>
<td>99.26%</td>
<td>86.99%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Opioid Prescription Information</strong> (included sample only: n = 11,636 (Site 1), n = 2,126 (Site 2))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Days Supplied</td>
<td>11,636</td>
<td>0</td>
<td>11,636</td>
<td>100%</td>
<td>1,411</td>
<td>715</td>
<td>2,126</td>
<td>66.37%</td>
</tr>
<tr>
<td>Dosage</td>
<td>11,636</td>
<td>0</td>
<td>11,636</td>
<td>100%</td>
<td>1,539</td>
<td>587</td>
<td>2,126</td>
<td>72.39%</td>
</tr>
<tr>
<td>Quantity</td>
<td>11,636</td>
<td>0</td>
<td>11,636</td>
<td>100%</td>
<td>1,696</td>
<td>430</td>
<td>2,126</td>
<td>79.77%</td>
</tr>
<tr>
<td>Frequency</td>
<td>11,636</td>
<td>0</td>
<td>11,636</td>
<td>100%</td>
<td>1,663</td>
<td>463</td>
<td>2,126</td>
<td>78.22%</td>
</tr>
</tbody>
</table>

---

417
age of 65.74 years (Site 1) and 64.30 years (Site 2). In both sites, TKA procedures were more common than THA procedures, where 57.19% and 57.71% of procedures at Site 1 and Site 2, respectively, were TKA. The sample largely identified as white (89.75%, 80.95%, respectively). Demographic information is included in Table 3 below. From the total samples, 66.98% (n=11,636) of patients from Site 1 and 46.68% (n=2,126) from site 2 did not meet any of the exclusion criteria and were included in the measure denominator.

Table 3: Patient demographic information (n = 17,372 (Site 1), n = 4,593 (Site 2))

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Site 1</th>
<th>Site 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Age (SD)</td>
<td>65.73</td>
<td>64.30</td>
</tr>
<tr>
<td>Age ≥ 65 years</td>
<td>56.01%</td>
<td>52.63%</td>
</tr>
<tr>
<td>18 ≤ Age ≤ 65 years</td>
<td>43.99%</td>
<td>47.37%</td>
</tr>
<tr>
<td>Male (%)</td>
<td>42.76%</td>
<td>36.88%</td>
</tr>
<tr>
<td>White</td>
<td>89.75%</td>
<td>80.95%</td>
</tr>
<tr>
<td>Black/African American</td>
<td>3.59%</td>
<td>17.31%</td>
</tr>
<tr>
<td>Other</td>
<td>6.66%</td>
<td>1.74%</td>
</tr>
<tr>
<td>English as first language</td>
<td>95.26%</td>
<td>98.45%</td>
</tr>
</tbody>
</table>

Unadjusted POP rates, defined as the percentage of patients receiving opioids for > 42 days following surgery, varied widely by both procedure type and by healthcare system. The overall unadjusted THA POP rate was 3.80% (Site 1) and 16.07% (Site 2). The unadjusted TKA POP rates were 12.63% (Site 1) and 25.68% (Site 2). See Figure 1 below for complete opioid episode duration by site and surgery type.

Figure 1: Opioid episode duration by site and surgery
Risk-adjusted rates at Site 1 ranged from 2.47%-8.01% (THA) and 8.96%-17.32% (TKA), producing overall site rates of 4.33% and 12.96%, respectively. Risk-adjusted rates from Site 2 ranged from 5.72%-11.61% (THA), and 10.28%-30.68% (TKA), producing overall site rates of 7.65% and 24.15% respectively. Adjusted and unadjusted POP rates at the clinician group level can be found in Table 4 below.

Table 4: Risk adjusted data by site and clinician group (n = 11,636 (Site 1), n = 1,766 (Site 2))

| Clinician Group | Site | Stratified Hip Rates | | Stratified Knee Rates | |
|-----------------|------|----------------------|----------------------|----------------------|
|                 |      | P/E                  | 95% CI               | Unadjusted | Adjusted | P/E                  | 95% CI               | Unadjusted | Adjusted |
| A               | 1    | 1.01                 | 0.62-1.64            | 4.92       | 5.15     | 0.87                 | 0.57-1.33            | 10.26      | 12.47    |
| B               | 1    | 0.71                 | 0.27-1.92            | 2.82       | 3.65     | 0.62                 | 0.31-1.25            | 7.14       | 8.91     |
| C               | 1    | 0.65                 | 0.33-1.30            | 4.91       | 3.35     | 0.69                 | 0.42-1.15            | 15.39      | 10.00    |
| D               | 1    | 0.66                 | 0.21-2.03            | 3.19       | 3.38     | 0.66                 | 0.29-1.50            | 10.82      | 9.44     |
| E               | 1    | 1.57                 | 1.00-2.49            | 8.01       | 8.01     | 1.21                 | 0.82-1.78            | 18.38      | 17.32    |
| F               | 1    | 0.48                 | 0.21-1.13            | 2.48       | 2.47     | 0.74                 | 0.46-1.23            | 11.55      | 10.70    |
| Overall Rate:   |      | 4.33 (Adjusted)      |                       |            |          | 12.96 (Adjusted)     |                       |            |          |
| B               | 2    | 1.12                 | 0.50-2.52            | 5.68       | 5.72     | 0.71                 | 0.28-1.88            | 5.38       | 10.28    |
| D               | 2    | 2.27                 | 1.58-3.28            | 23.53      | 11.61    | 2.13                 | 1.60-2.84            | 35.78      | 30.50    |
| F               | 2    | 1.28                 | 0.73-2.52            | 6.49       | 6.54     | 1.00                 | 0.64-1.59            | 15.22      | 14.35    |
| I               | 2    | 1.31                 | 0.73-2.38            | 15.0       | 6.68     | 2.14                 | 1.64-2.80            | 36.07      | 30.68    |
| Overall Rate:   |      | 7.62 (Adjusted)      |                       |            |          | 24.15 (Adjusted)     |                       |            |          |

Clinician groups with <50 procedures in the sample were excluded from risk-adjustment due to low sample size. P/E represents the Predicted/Expected ratios.

We ranked the predicted/expected ratios in the test and validation samples and then estimated the Spearman rank correlation coefficient to correlate the ranking between samples. Table 5 below shows the predicted/expected ratios for the six clinician groups. These test and validation samples provide a Spearman rank correlation of 0.81 for THA procedures and 0.89 for TKA procedures, which is considered strong.

Table 5: Predicted/Expected test and validation samples by site and surgery type (n = 11,636 (Site 1), n = 1,766 (Site 2))

| Site | Clinician | Predicted/Expected Ratios-Hip | | Predicted/Expected Ratios-Knee | |
|------|-----------|-------------------------------|----------------------|----------------------|
|      |           | Test                          | Validation           | Test                          | Validation           |
| 1    | A         | 1.10                          | 0.92                 | 0.90                 | 0.84                 |
| 1    | B         | 0.82                          | 0.61                 | 0.60                 | 0.65                 |
| 1    | C         | 0.70                          | 0.62                 | 0.71                 | 0.69                 |
| 1    | D         | 0.68                          | 0.64                 | 0.60                 | 0.72                 |
| 1    | E         | 1.65                          | 1.49                 | 1.24                 | 1.18                 |
| 1    | F         | 0.52                          | 0.45                 | 0.74                 | 0.75                 |
| 2    | B         | 0.85                          | 1.39                 | 0.75                 | 0.69                 |
| 2    | D         | 1.92                          | 2.63                 | 2.03                 | 2.23                 |
| 2    | I         | 1.44                          | 1.13                 | 1.03                 | 0.97                 |
| 2    | F         | 1.01                          | 1.62                 | 2.16                 | 2.13                 |
| Overall Spearman |       | 0.81                          |                      | 0.89                 |

Clinician groups with <50 procedures were not included in risk-adjustment due to low sample size.
We compared sociodemographic characteristics of patients included in our test and validation samples and found there were no differences at the patient level (p = 0.12-0.68) or between clinician groups (p = 0.99). Moreover, patient level characteristics (within effects) are predictive of the outcome with a C-statistic of 0.71 (THA) and 0.69 (TKA). The C-statistic was analyzed with the socio-economic status (SES) variables included. A C-statistic of 0.5 indicates that the predictive model is no better than random chance, and a C-statistic of 0.7 or above indicates a good model\textsuperscript{12}, indicating that our model is effective at predicting outcomes, with some room for improvement.

A Hosmer-Lemeshow statistic was performed to compare predicted and expected events. These calibration approaches had a P value of 0.62 (THA), and 0.64 (TKA). A P value > 0.1 demonstrates a good fit between observed and expected values.

Discussion

This manuscript outlines the testing of an eCQM intended to assess opioid prescribing practices at two geographically distant sites with distinct patient populations. This study highlights the differences in both data availability and prescribing practices between the two sites. Site 2 recorded notably less data regarding their patient’s opioid episodes (68% in site 2 compared to 100% in site 1) and, in the data recorded, demonstrated patients with a markedly higher rate of prolonged opioid use following surgery. According to the literature review, Site 2’s patients are representative of common postoperative prescribing durations following THA and TKA.\textsuperscript{15} Site 1, however, maintained complete documentation of their opioid prescribing information and contains patients who are prescribed opioids for a shorter duration following their surgery compared to both Site 2 and the rates seen in the literature.\textsuperscript{15} Therefore, such a relationship may demonstrate the importance of maintaining accurate post-operative opioid prescribing data to address prolonged use following surgery.

Measure developers aim to develop eCQMs to minimize provider burden and do not have the authority to require clinician groups to record the D/F/Q/DS of an opioid prescription within EHR systems. Differences in rates across sites were to be expected based on the lack of federal-level policies regarding opioid documentation. It is currently up to the state and clinical group/healthcare system’s discretion to determine what elements require documentation, and currently, few states have any such requirements.\textsuperscript{15} Considering the potential for opioid misuse following joint arthroplasty procedures, it is recommended that sites document relevant opioid prescription information, as seen in Site 1, to understand their own clinician group level prescribing practices, and ultimately to address the ongoing opioid epidemic.

There are several limitations to this research. As mentioned above, missing data availability in Site 2 is noted as a limitation which impacted this measure in both numerator/denominator assignment and in the risk-adjustment model. Many Site 2 patients (n=715) fit the denominator criteria and were included in the measure but lacked enough opioid prescription information needed to assess their inclusion in the numerator. This means that there are likely patients from Site 2 who experienced a prolonged opioid use episode but were not included in the POP rate, artificially lowering the rate of both the clinician group and the overall sample. It is likely that the POP rates in Site 2 are higher than presented in this analysis. However, due to the lack of documentation at this site, we are unable to say anything definitive. In order to ensure that the missing data from Site 2 was random, our team analyzed random samples of patients with missing data and complete data by patient demographic (age, BMI, race, sex, primary language) as well as by other possible factors including clinician group, surgeon, time of year, and differences in medication type. Ultimately, we found no significant trends in the missing data, indicating that the missing data is not biased towards a particular patient cohort.

In addition to this shortcoming in data element availability, smoking status was removed from the measure as a risk-adjustment variable due to the lack of documentation in Site 2 (where smoking status information was available for 1.05% of patients). Studies have shown that smoking status is related to increased opioid use following THA compared to non-smokers with similar patterns seen following TKA.\textsuperscript{16} However, using site 1 data, we examined the effect of the risk-adjustment model when both including and excluding smoking status and found no significant difference between the two groups. We maintain the belief, however, that like proper opioid documentation, in order for this public health epidemic to be addressed, a key step is diligent documentation of all relevant information.

With testing complete, this measure has now been for the Measure under consideration (MUC) list for 2021. We hope that if the measure is approved, the POP eCQM will aid providers in improving their prescribing practices and enable hospital systems to track the relevant opioid information needed to curb the opioid epidemic.
Neighborhood Characteristics and COVID-19 Incidence and Mortality in Southeastern Pennsylvania

Alexandra A. Rizaldi, Sherrie Xie, PhD, Rebecca A. Hubbard, PhD, Blanca E. Himes, PhD
Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, United States

Abstract

The COVID-19 pandemic has differentially impacted people according to their race/ethnicity, socioeconomic status, and preexisting conditions. Public health surveillance efforts, especially those occurring early in the pandemic, did not gather nor report adequate individual-level demographic information to identify these differences, and thus, neighborhood-level characteristics were used to note striking disparities in the US. We sought to determine whether risk factors associated with COVID-19 incidence and mortality in five Southeastern Pennsylvania counties could be better understood by using neighborhood-level demographic data augmented with health, socioeconomic, and environmental characteristics derived from publicly available sources. Although we found that education level and age of residents were the most salient predictors of COVID-19 incidence and mortality, respectively, neighborhoods exhibited a high degree of segregation with multiple correlated factors, which limits the ability of neighborhood-level analysis to identify actionable factors underlying COVID-19 disparities.

Introduction

The Coronavirus Disease 2019 (COVID-19) pandemic is a public health emergency that continues to devastate the global population, having resulted in over 29.1 million infections and 534,733 deaths in the US alone as of March 12, 2021. Striking disparities by race/ethnicity in severe COVID-19 outcomes and death in the US have been noted since the onset of the pandemic. Severe morbidity and death also occur more frequently among COVID-19 patients of older age and those with preexisting conditions, including chronic respiratory disease, cardiovascular disease, diabetes, obesity, and hypertension. Exposure to ambient air pollution, including fine particulate matter (PM2.5), increases risk for many of these conditions and also for COVID-19 infection. Structural racism and income inequality have contributed to stark and persistent social disparities: racial/ethnic minorities and the poor are more likely to live and work in crowded environments with increased pollution and have more chronic diseases, tenuous employment, sustained toxic stress, and inequitable access to health care systems. These factors contribute to racial/ethnic minorities and the poor being at peak risk for COVID-19 and its dire outcomes.

The Southeast Pennsylvania (SEPA) region consists of urban Philadelphia County and the surrounding suburban counties of Bucks, Chester, Delaware, and Montgomery. SEPA has been heavily affected by COVID-19, with cumulative, county-level incidence as of March 12, 2021 ranging from 5.25 (Bucks) to 7.59 (Philadelphia) per 100 residents, and mortality ranging from 0.14 (Chester) to 0.23 (Delaware) per 100 residents. Individual-level data reporting was limited at the start of the pandemic, and public health efforts began by focusing on neighborhoods reporting the highest rates of COVID-19 infection and death. Most notably in Philadelphia, which has a high proportion of Black residents, overcoming testing barriers in neighborhoods with majority Black residents was a priority because such neighborhoods had the highest rates of COVID-19. More broadly, over half of Delaware and Philadelphia counties consist of environmental justice areas—census tracts where 20 percent or more of individuals live in poverty and/or 30 percent or more of the population identifies as non-white minority. Residents of these areas have historically suffered a disproportionate burden of pollutant exposures from sources that include oil refineries, trash incinerator plants, and vehicular exhaust from major roadways, putting them at increased risk for various diseases, including COVID-19. Between March 2020 and March 12, 2021, SEPA experienced two distinct pandemic surges: an initial peak in April 2020 and a resurgence in early winter of that year. Characteristics of COVID-19 cases and deaths during these two waves have been found to vary around the world in terms of age distribution, race/ethnicity, and the socioeconomic status and occupations of those affected.

Although it was known that the COVID-19 pandemic disproportionately impacted racial/ethnic minority and socioeconomically disadvantaged groups in the US including at-risk populations within Southeastern Pennsylvania, we sought to determine whether in the absence of individual-level data, a richer compilation of neighborhood-level health, demographic, environmental, and socioeconomic characteristics would facilitate the identification of COVID-19 risk factors.
Methods

COVID-19 Data. Data on COVID-19 cases and deaths was obtained from the public health departments of five counties in Southeast Pennsylvania, and that of individual long-term care facilities (LTCFs) was obtained from the Pennsylvania Department of Health website. Data was aggregated by county subdivision geographical units according to the smallest one provided by local public health departments: zip codes in Philadelphia County and municipalities in Bucks, Chester, Delaware, and Montgomery Counties. With the exception of Bucks County data, which was provided directly by its health department, data was downloaded from county health department websites at two time points that distinguished SEPA COVID-19 peaks: August 18, 2020 and March 12, 2021. Based on these dates, we defined three study periods: Wave 1 (March 1 to August 18, 2020), Wave 2 (August 19, 2020 to March 12, 2021), and Total (March 1, 2020 to March 12, 2021). Data for Wave 2 was determined by finding the difference between Total and Wave 1 counts. LTCFs were geocoded using the R package ggmap and assigned to a zip code or municipality via OGIS; COVID-19 outcome counts attributed to LTCFs were subsequently aggregated by zip code or municipality and excluded from the public health department total outcome counts.

American Community Survey Variables. Demographic and socioeconomic status variables were obtained from the 2015-2019 American Community Survey (ACS) 5-year estimates using the R package tidyCensus24 for geographic areas corresponding to COVID-19 data. Population density was calculated by dividing population by land area in square kilometers, which was derived from the Census Bureau’s cartographic boundary files from the MAF/TIGER geographic database. Sex was expressed as the proportion of males in the population. Race was re-leveled into four groups: proportion of non-Hispanic/Latino White, non-Hispanic/Latino Black, Asian, and Hispanic/Latino. Age was re-leveled into four groups: proportion of residents 18 to 34, 35 to 49, 50 to 64, and greater than 65 years old. Household size was re-leveled into three groups: proportion of 1-person, 2-person, and 3 or more-person households. Education level was re-leveled into four groups based on the proportion of people aged 25 years or older with: less than 9th grade education, at least a high school diploma, at least a bachelor’s degree, and at least a graduate degree (master’s, professional school, or doctorate). Median household income in the past 12 months was expressed in 2019 inflation-adjusted US dollars (USD). Three housing characteristics were expressed as a proportion of all households: owner-occupied households, single-parent households (male or female householder with no spouse present and children under 18 years), and households with no vehicle available.

Household Health Survey Variables. Data on the health status and behaviors of local residents was obtained from the Southeast Pennsylvania Household Health Survey (SEPA-HHS), conducted in 2012, 2015, and 2018 by the Public Health Management Corporation (PHMC). This community survey of Philadelphia and four surrounding counties interviewed residents aged 18 years or older by telephone to measure their health status and healthcare experiences. Responses from the three survey years were aggregated by geographical subdivision corresponding to COVID-19 data. Sampling bias was addressed by stratifying the sample by geographic subareas and by including population weights based on Census estimates of race, age, sex, ethnicity, household size, and income derived using Claritas, Inc. to give more weight to underrepresented and less weight to overrepresented segments of the sample. Empirical Bayes estimation of small area prevalence was used to smooth prevalence estimates of zip codes or municipalities with small sample sizes towards the county-level value.

Health outcomes considered were the proportion of people with: 1) Diabetes, based on an affirmative response to “Have you EVER been told by a doctor or other health professional that you have or had diabetes?” and excluding diabetes occurring only during pregnancy; 2) Asthma, based on an affirmative response to “Have you EVER been told by a doctor or other health professional that you have or had asthma?”; 3) Hypertension, based on an affirmative response to “Have you EVER been told by a doctor or other health professional that you have high blood pressure or hypertension?” and excluding hypertension only during pregnancy; 4) Obesity, defined as BMI ≥ 30, calculated from each respondent’s height in meters and weight in kilograms based on the equation weight/squared-height; and 5) Mental health condition, based on an affirmative response to “Have you ever been diagnosed with any mental health condition, including clinical depression, anxiety disorder or bipolar disorder?”.

Lifestyle factors were the proportion of people: 1) Now smoking, based on a dichotomous re-leveled question in which respondents answered either “Everyday” or “Some days” to “Do you NOW smoke cigarettes every day, some days, or not at all?”; 2) Exposed to smoke at home, based on an affirmative response to “Does anyone living in your household smoke cigarettes, cigars or pipes INSIDE your home?”; 3) Eating 3+ fruits/vegetables per day, based on a dichotomous re-leveled self-reported question in which respondents answered “How many servings of fruits and vegetables do you eat on a typical day?” with a number three or greater; and 4) Exercising, based on a dichotomous re-leveled self-reported question in which respondents answered “Thinking about the past month, how many times
per week did you participate in any physical activities for exercise that lasted for at least one half-hour, such as walking, basketball, dance, rollerblading or gardening?” with a number three or greater.

Socioeconomic status factors related to health were the proportion of people who: 1) had forgone food due to cost, based on an affirmative response to “In the last 12 months, did you ever cut the size of meals or skip meals because there was not enough money in the budget for food?”; and 2) Have health insurance, based on an affirmative response to “Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare?”. Insurance status was further classified according to variables representing the proportion of people with a specific insurance type: 1) Employer-sponsored, defined as health insurance through work, school, or union; 2) Private/Personal, defined as health insurance that the respondent or a family member buys on their own; 3) Medicare, defined as health insurance through “Medicare A” for 2012 and 2015 surveys and “Medicare” for the 2018 survey; 4) Medicaid, defined as health insurance through Medicaid or another state program such as Medical Assistance (M.A.) or HealthChoices; and 5) Military, defined as health insurance through TRICARE, CHAMPUS, VA, or Military.

Social capital was captured as a 3-level variable based on an additive point score calculated from three survey variables: 1) Improve, based on an affirmative response to “Have people in your neighborhood ever worked together to improve the neighborhood?” which represented 1 point; 2) Belong, based on the response to “Please tell me if you strongly agree, agree, disagree, or strongly disagree with the following statement: I feel that I belong and am a part of my neighborhood.”, with “Strongly agree” representing 2 points, “Agree” representing 1 point, and “Disagree” and “Strongly disagree” representing 0 points; and 3) Participate, based on the question “How many local groups or organizations in your neighborhood do you currently participate in such as social, political, religious, school-related, or athletic organizations?”, with the number of organizations corresponding to the number of points, up to 12 points maximum. Based on the distribution of respondents’ point scores, low social capital was defined as the proportion of people with 0 to 1 point, representing the 0 to 25th percentile; medium social capital was defined as the proportion of people with 2 to 3 points, representing the 25th to 72nd percentile, and high social capital was defined as the proportion of people with 4 to 14 points, representing the 73rd to 100th percentile.

Environmental Protection Agency (EPA) Variables. Air pollution was based on mean estimates of PM$_{2.5}$ by geographical subdivision corresponding to COVID-19 data, derived from 2010 to 2019 rasterized yearly estimates computed from data sourced from EPA Air Data.\textsuperscript{30}

Statistical Analysis. Statistical analyses were conducted in R.\textsuperscript{31} Bivariate analyses were conducted using quasi-Poisson generalized linear models with COVID-19 incidence or mortality during each of the three time periods as the outcome and each of 23 variables representing neighborhood race, age, household size, income, education level, insurance billing class, social capital, and health characteristics of its residents as a predictor, using log(population) as an offset term. Incidence rate ratios (IRRs) were obtained by exponentiating model coefficients of each variable. Multivariable models with COVID-19 incidence or mortality during each of the three time periods as the outcome were created via a two-stage process: 1) feature selection was conducted using LASSO regression analysis with the quasi-Poisson family and a population offset term. Variables were selected if they had IRRs $>1.1$ or $<0.9$ in the final models; 2) multivariable quasi-Poisson regression models were created using selected variables for each outcome (i.e., COVID-19 incidence and mortality for each of the three time periods), including offsets for population.

Results

SEPA COVID-19 incidence and mortality rates. COVID-19 outcomes data was available for 284 zip codes and municipalities. The total (i.e., through March 12, 2021) LTCF-excluded COVID-19 incidence and mortality per 100 residents in these geographic areas ranged from 1.13 to 12.06 and 0 to 0.87, respectively (Figure 1). In Wave 1, incidence and mortality per 100 residents ranged from 0 to 4.35 and 0 to 0.58, respectively (Figure 1).
respectively, while in Wave 2, they ranged from 0.25 to 9.01 to 0 to 0.87, respectively. The area with greatest total incidence was the 19136 zip code in Philadelphia, where four correctional facilities are located. The area with greatest total mortality was Richlandtown Borough in Bucks County.

**Bivariate Associations Between Neighborhood Characteristics and COVID-19 Outcomes.** Characteristics of the neighborhood-level variables for the 265 geographical subdivisions are shown in Table 1. Of 23 variables considered, COVID-19 incidence was associated with all but sex and household size across the total period, with all variables during Wave 1, and with all but sex and asthma prevalence in Wave 2. COVID-19 mortality was associated with all variables but education and exercise level across the total period, with sex, race, age, household size, median household income, owner-occupied and without vehicle household prevalence, diabetes, asthma, hypertension, mental health condition, home smoke exposure, and Medicaid insurance prevalence during Wave 1, and with all but population density and sex prevalence in Wave 2. The directionality of significant relationships remained consistent across the three time periods studied (Figure 2). Several neighborhood-level variables were highly correlated with one another (Figure 3). For example, having a greater percentage of Black residents was positively correlated with population density, owner-occupied and without vehicle household prevalence, chronic health condition prevalence (diabetes, asthma, hypertension, and obesity), the percent of residents who now smoke or are exposed to smoke at home, the percent of residents who have forgone food due to cost, and the percent of residents who have Medicaid insurance. Meanwhile, a greater percentage of Black residents was negatively correlated with median household income, the percent of residents with a high school diploma, bachelor’s degree, or graduate degree, the percent of households that are owner-occupied, the percent of residents eating three or more fruits/vegetables per day, and the percent of residents with employer-sponsored insurance. Additionally, a greater percentage of residents aged 65 or older was positively correlated with the percent of residents who are White and negatively correlated with the percent of households with three or more residents, the percent of residents with less than 9th grade education, and the percent households that are single parent. For education level, a greater percentage of residents with less than 9th grade education was positively correlated with the percent of residents who are Latino and the percent of households without a vehicle, and negatively correlated with the percent of residents with health insurance.

**Figure 2.** Neighborhood-level factors associated with COVID-19 A) incidence and B) mortality for three time periods: Total (March 2020 – March 12, 2021), Wave 1 (March 2020 – August 18, 2020), and Wave 2 (August 18, 2020 – March 12, 2021). Shown are incidence rate ratios (IRRs) with 95% confidence intervals corresponding to bivariate quasi-Poisson generalized linear models created with COVID-19 incidence or mortality as the outcome and individual variables listed as predictors.
Table 1. Average prevalence estimates for zip codes (Philadelphia) and municipalities (Bucks, Chester, Delaware, Montgomery), expressed per 100 residents unless otherwise stated.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean (SD) or Median (IQR)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population density (people/km²)</td>
<td>930.3 (2075.3)*</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>48.7 (2.7)</td>
</tr>
<tr>
<td>Race</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>86.3 (20.0)*</td>
</tr>
<tr>
<td>Black</td>
<td>4.8 (11.7)*</td>
</tr>
<tr>
<td>Asian</td>
<td>3.0 (4.9)*</td>
</tr>
<tr>
<td>Latino</td>
<td>3.9 (4.4)*</td>
</tr>
<tr>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>18-34</td>
<td>22.2 (8.0)</td>
</tr>
<tr>
<td>35-49</td>
<td>18.6 (2.9)</td>
</tr>
<tr>
<td>50-64</td>
<td>21.3 (4.0)</td>
</tr>
<tr>
<td>65+</td>
<td>16.1 (5.5)</td>
</tr>
<tr>
<td>Household Size</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>26.8 (9.2)</td>
</tr>
<tr>
<td>2</td>
<td>33.8 (6.0)</td>
</tr>
<tr>
<td>3+</td>
<td>40.6 (9.3)</td>
</tr>
<tr>
<td>Education Level</td>
<td></td>
</tr>
<tr>
<td>Less than 9th grade</td>
<td>1.8 (2.5)*</td>
</tr>
<tr>
<td>High school diploma</td>
<td>93.8 (6.6)*</td>
</tr>
<tr>
<td>Bachelor’s degree</td>
<td>39.0 (30.4)*</td>
</tr>
<tr>
<td>Graduate degree</td>
<td>14.1 (14.7)*</td>
</tr>
<tr>
<td>Socioeconomic Status</td>
<td></td>
</tr>
<tr>
<td>Median household income (USD)</td>
<td>85055.5 (31332.4)</td>
</tr>
<tr>
<td>Housing (% of all households)</td>
<td></td>
</tr>
<tr>
<td>Owner-occupied</td>
<td>72.2 (26.7)*</td>
</tr>
<tr>
<td>Single-parent</td>
<td>9.3 (11.4)*</td>
</tr>
<tr>
<td>Without vehicle</td>
<td>5.4 (8.9)*</td>
</tr>
<tr>
<td>Health</td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td>11.9 (4.0)</td>
</tr>
<tr>
<td>Asthma</td>
<td>15.7 (3.7)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>30.7 (6.3)</td>
</tr>
<tr>
<td>Obesity</td>
<td>28.3 (7.2)</td>
</tr>
<tr>
<td>Mental health condition</td>
<td>17.5 (4.6)</td>
</tr>
<tr>
<td>Lifestyle</td>
<td></td>
</tr>
<tr>
<td>Now smoking</td>
<td>16.0 (6.3)</td>
</tr>
<tr>
<td>Exposed to smoke at home</td>
<td>10.8 (5.6)</td>
</tr>
<tr>
<td>3+ fruits/vegetables per day</td>
<td>48.6 (8.4)</td>
</tr>
<tr>
<td>Exercise</td>
<td>54.6 (5.3)</td>
</tr>
<tr>
<td>Socioeconomic Status</td>
<td></td>
</tr>
<tr>
<td>Forgone food due to cost</td>
<td>9.7 (5.5)</td>
</tr>
<tr>
<td>Have health insurance</td>
<td>93.0 (4.8)</td>
</tr>
<tr>
<td>Insurance Billing Class</td>
<td></td>
</tr>
<tr>
<td>Employer-sponsored</td>
<td>59.6 (8.4)</td>
</tr>
<tr>
<td>Private/Personal</td>
<td>36.2 (4.4)</td>
</tr>
<tr>
<td>Medicare</td>
<td>27.8 (5.0)</td>
</tr>
<tr>
<td>Medicaid</td>
<td>9.2 (7.1)</td>
</tr>
<tr>
<td>Military</td>
<td>1.9 (1.0)</td>
</tr>
<tr>
<td>Social Capital</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>29.5 (7.7)</td>
</tr>
<tr>
<td>Medium</td>
<td>47.9 (6.2)</td>
</tr>
<tr>
<td>High</td>
<td>25.1 (6.8)</td>
</tr>
<tr>
<td>Pollution</td>
<td></td>
</tr>
<tr>
<td>PM 2.5 (µg/m³)</td>
<td>9.4 (0.3)</td>
</tr>
</tbody>
</table>

*Variables with approximately normal distributions were described with mean and standard deviation (SD), while the rest were described with median and interquartile range (IQR).
Neighborhood characteristics associated with COVID-19 incidence and mortality according to multivariable analysis. The lasso feature selection step found that education level, age, and sex were associated with COVID-19 incidence while age, exercise level, and prevalence of asthma, diabetes, and mental health conditions were associated with COVID-19 mortality in at least one time period considered (Table 2). However, only a subset of these variables remained significant in the final quasi-Poisson generalized linear models, as indicated by the significance thresholds shown in Table 2. For the three time periods considered, the only features consistently selected were education level for COVID-19 incidence and age 65+ years for COVID-19 mortality. In the final quasi-Poisson generalized linear models, the proportion of residents aged 25 years and older with less than 9th grade education was found to be the only significant (p<0.05) predictor of COVID-19 incidence across all three time periods. When considering COVID-19 mortality, the proportion of residents aged 65 years and older was found to be the only significant predictor of COVID-19 mortality across all three time periods. In both COVID-19 incidence and mortality, the strength of the relationship as measured by IRR magnitude with education level and age, respectively, was greater in Wave 1 compared to Wave 2. IRRs for COVID-19 incidence and education level were 1.88 for Wave 1 and 1.34 for Wave 2, and IRRs for COVID-19 mortality and age were 2.56 for Wave 1 and 1.65 for Wave 2.

<table>
<thead>
<tr>
<th>COVID-19 Outcomes</th>
<th>Total Incidence</th>
<th>Wave 1 Incidence</th>
<th>Wave 2 Incidence</th>
<th>Wave 1 Mortality</th>
<th>Wave 2 Mortality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education Level</td>
<td>Less than 9th grade</td>
<td>High school diploma</td>
<td>Graduate degree</td>
<td>Less than 9th grade</td>
<td>High school diploma</td>
</tr>
<tr>
<td>Health</td>
<td>Diabetes</td>
<td>Asthma</td>
<td>Hypertension</td>
<td>Obesity</td>
<td>Mental health condition</td>
</tr>
<tr>
<td>Lifestyle</td>
<td>Exposed to smoke at home</td>
<td>3+ fruits/vegetables per day</td>
<td>Exercise</td>
<td>Exposed to smoke at home</td>
<td>3+ fruits/vegetables per day</td>
</tr>
<tr>
<td>Socioeconomic Status</td>
<td>Felt food due to cost</td>
<td>Felt food due to cost</td>
<td>Felt food due to cost</td>
<td>Felt food due to cost</td>
<td>Felt food due to cost</td>
</tr>
<tr>
<td>Insurance Billing Class</td>
<td>Employer-sponsored</td>
<td>Private/Patient</td>
<td>Medicare</td>
<td>Medicaid</td>
<td>Military</td>
</tr>
<tr>
<td>Social Capital</td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
<td>Low</td>
<td>Medium</td>
</tr>
<tr>
<td>Air Pollution</td>
<td>PM2.5 (μg/m³)</td>
<td>PM2.5 (μg/m³)</td>
<td>PM2.5 (μg/m³)</td>
<td>PM2.5 (μg/m³)</td>
<td>PM2.5 (μg/m³)</td>
</tr>
</tbody>
</table>

Figure 3. Correlation plot of COVID-19 outcomes and neighborhood-level characteristics considered. Color scale denotes Pearson’s correlation measure.

427
Table 2. Factors associated with COVID-19 incidence and mortality in multivariable analyses. Shown are incidence rate ratios (IRRs) and 95% confidence intervals reflecting the risk associated with a 10% increase in neighborhood prevalence of a variable for all features selected by lasso. *p<0.05; **p<0.001

<table>
<thead>
<tr>
<th></th>
<th>COVID-19 Incidence</th>
<th>COVID-19 Mortality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Wave 1</td>
</tr>
<tr>
<td>Sex: Male</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Age: 65+</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>&lt; 9th grade education</td>
<td>1.47</td>
<td>(1.38, 1.56)**</td>
</tr>
<tr>
<td>Diabetes</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Asthma</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Mental health condition</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Exercise</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Medicaid</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Discussion

Our study found that in Southeastern Pennsylvania, education level was the most salient predictor of COVID-19 incidence at the zip code and municipality level. For all three time periods studied, a higher proportion of residents aged 25 or older with less than 9th grade education was associated with an increased risk of COVID-19 incidence. Previous studies have found a relationship between education level and COVID-19 incidence. In national, county-level risk-adjusted analysis, the percentage of adults without a high school degree had the strongest association with an increased risk of COVID-19 incidence, which was consistent with our results. Individuals with lower levels of education are less likely to work remotely, which suggests that they are less able to physically distance from other people and have to choose between staying at home and foregoing wages, or going to work and increasing their risk of COVID-19 infection. Higher education levels have also been associated with higher personal knowledge of COVID-19, which may offer greater understanding of health information and prevention measures to reduce the risk of COVID-19 infection. The decreased strength in the relationship between education level and COVID-19 incidence between the first and second waves is consistent with a national, county-level analysis that found that the percentage of adults with less than high school education who had COVID-19 decreased as the pandemic progressed. Hypotheses for this shift include that mitigation efforts affected people with lower levels of education more strongly and that access to testing increased by a greater amount among the highly educated.

In terms of COVID-19 mortality, the proportion of residents aged 65 or older was the most salient risk factor. This relationship held after the exclusion of state-reported cases and deaths attributed to LTCFs, which serve older adults and suffered a disproportionate burden of COVID-19 infection, especially in the early stages of the pandemic. The relationship between age and COVID-19 mortality has also been well reported. In a UK-based study, current age was exponentially associated with COVID-19 mortality, and older adults with additional risk factors (e.g., comorbid conditions) were at an even greater risk of death. Another study analyzing age-specific patterns of COVID-19 in 45 countries estimated a log-linear increase of a population’s infection fatality ratio by age in adults older than 30 years. In older individuals, severe COVID-19 outcomes have been attributed to aging innate and adaptive immune systems, increased inflammation and cytokine profiles, and chronic disease comorbidities, all of which can reduce the body’s ability to prevent the development of cytokine storm, an inflammatory response that can lead to tissue damage and multi-system organ failure. Our finding that the relationship between older age and COVID-19 mortality decreased in strength from Wave 1 to Wave 2 may be due to the availability of vaccinations and increased testing, as well as a decreased burden of COVID-19 in congregate living facilities (e.g., LTCFs and prisons) following the implementation of more effective mitigation strategies. Our multivariable models identified two salient risk factors, but other significant relationships identified in at least one time period included education level, age, and sex with COVID-19 incidence, and age, exercise level, and prevalence of asthma, diabetes, and mental health with COVID-19 mortality. Our neighborhood-level analyses were...
limited due to the high correlation among many neighborhood variables, which hampered our ability to understand which variables may have driven the relationships with COVID-19 outcomes. Although lasso is a technique that can efficiently select predictors while minimizing overfitting of data, it is unable to differentiate between factors that have a similar statistical relationship with an outcome. For example, education level was correlated with other measures of socioeconomic status, race, insurance status, and health outcomes, making it difficult to discern which of these predictors truly underlies the observed relationship between education level and COVID-19 incidence at the neighborhood level. Despite this limitation, lasso has been used in COVID-19 studies to select features for final models of COVID-19 prevalence, and in other studies investigating chronic disease prevalence in small geographic areas to identify variables associated with disease outcomes in a similar manner to this study.

Our study is subject to additional limitations. The PHMC Household Health Survey, one of the three sources used to represent neighborhood characteristics of the Southeastern Pennsylvania region, consisted of adult respondents whose mean age is in the mid-50s for the years considered. This limitation does not substantially affect our results given that we used weighted survey results and that the COVID-19 disease burden largely affected adults. Additionally, our COVID-19 outcome data was limited in that we excluded LTCF case and death counts from our outcome variables, not all LTCFs reported data to the Pennsylvania Department of Health, and data from other congregate living facilities such as prisons and personal care homes was not available to us. Because all of our variable estimates were at the municipality or zip code level, data granularity also limits the specificity of results; data available at a more granular geographical subdivision such as census tract would have allowed for more precise conclusions about the region.

In summary, we found that education level and the age of residents were the most salient neighborhood-level predictors of COVID-19 incidence and mortality, respectively, with the strength of the relationship decreasing as the pandemic progressed. Due to the high level of collinearity among geographic area variables that reflects the cumulative burden of lifetime adversities such as systemic racism experienced by Philadelphia area residents, clarifying the role of individual variables with COVID-19 was made difficult. While neighborhood-level analyses can be useful in determining the specific needs of vulnerable populations and in informing policies to address health disparities related to COVID-19, our results underscore the importance of gathering individual-level data as a pandemic emerges and progresses.

References


Implementation of a Maternal Child Knowledgebase

Donna A. Santillan, PhD¹, Mark K. Santillan, MD, PhD¹, Heather A. Davis, MS¹, Mark Crooks, MS¹, Patrick J. Flanagan¹, Chris E. Ortman¹, Elissa Z. Faro, PhD¹, Stephen K. Hunter, MD, PhD,¹ Boyd K. Knosp, MS, FAMIA¹
¹University of Iowa, Iowa City, IA

Abstract

To advance the application of clinical data to address maternal health we developed and implemented a Maternal Child Knowledgebase (MCK). The MCK integrates data from every pregnancy that received care at the University of Iowa Hospitals & Clinics (UIHC) and links information from the pregnancy episode to the delivery episode and between the mother and child. This knowledgebase contains integrated information regarding diagnoses, medications, mother and child vitals, hospital admissions, depression screenings, laboratory value results, and procedure information. It also collates information from the electronic health record (EPIC), the Social Security Death Index, and the Medication Administration Record into one knowledgebase. To enhance usability, we designed a custom viewer with several pre-designed queries and reports that eliminates the need for users to be proficient in SQL coding. The recent implementation of the MCK has supported multiple projects and reduced the number of Obstetrics-related data queries to the Biomedical Informatics group.

Introduction

Many years of applying the results from males in clinical trials to women have done a disservice to the health of women and challenges the quality of our general medical knowledge in the science and medicine of women’s health. Only in the past decade are we beginning to identify key sex differences in the signs and symptoms of diseases, such as heart disease, and in differences in responses to pharmaceutics. While women’s health, in general, has been largely understudied, pregnancy-related work is even more elusive. This work is especially vital for rural states and for reducing disparities in healthcare. Kozhimannil et al. reported that rural residents have a 9 percent greater probability of maternal morbidity and mortality than urban residents based on data from the National Inpatient Sample.[1] Women of color experience a disproportionate burden of maternal mortality in the U.S. Notably, Non-Hispanic black women and Native American/Native women who are 4-5 times more likely to die.[2] Our objective was to utilize a technology-based approach to facilitate maternal health surveillance and research, including work to identify health disparities.

Because pregnancy involves more than one patient, that is the mother and child(ren), data extraction and organization are complicated. First, the correct mother-child dyad must be identified and linked. When data are viewed, there needs to be a clear system to identify the pairs. When a woman has given birth to more than one child in a pregnancy or has had more than one pregnancy at an institution, linking the correct mother and child(ren) becomes more complex. Furthermore, identifying the true source of data can be challenging. Important pregnancy outcomes data can be found in either the mother’s or child(ren)’s charts or in both locations. Challenges occur when data are located in multiple locations and do not agree.

One technological approach to address these issues is to develop an enterprise data warehouse for research (EDW4R) populated with data from pregnancy episodes. The use of EDW4R, with guidance from National Center for Advancing Translational Sciences (NCATS) sponsored Clinical Translational Science Award (CTSA) award hubs, has grown nationally.[3] While the adoption of electronic health records (EHRs)[4 5] has accelerated data acquisition, problems of data integration are further magnified as data sources are varied [6 7]; universal data structures and data harmonization techniques may not apply to all clinical subspecialties. It is recognized that EHRs are patient-centric and can enhance clinical practice; EDW4R integrate data from multiple sources on all patients within the enterprise and “focuses on enhancing the well-being of current patients, the purpose of an [EDW4R] is to produce generalized knowledge that can be extended to future patients.”[8] Timely use of EDW4R data can rapidly be deployed to answer contemporary questions. The impact of COVID-19 disease and predicting COVID-19 outcomes has been a significant challenge over the past year. Using an EDW4R of 11586 patients, Sun H. et al. have developed an acute outpatient
clinical prediction model that has shown excellent validated performance in the prediction of COVID-19 related hospitalization (AUC=0.76), critical illness (AUC=0.79), and death (AUC=0.93).[9] While these efforts are more pervasive in general medicine services, such harmonization and integration are not as prevalent in Obstetrics and Gynecology (OB/Gyn). The National Institutes of Health has published recommendations to encourage such data harmonization and a unified knowledgebase in Obstetrics and Gynecology to further the study of many perinatal conditions such as the hypertensive pregnancy disease, preeclampsia.[10] Herein, we describe the development and implementation of one such OB/Gyn-focused enterprise data warehouse for research, termed the Maternal-Child Knowledgebase (MCK), at the University of Iowa.

Materials and Methods:

Stakeholders/Team Members: The operations team, which oversees components of the MCK, is composed of a clinical data team and the biomedical informatics group. The clinical data team (CDT) consists of the principal investigator of the Women’s Health Tissue Repository (WHTR) and the Clinical Research Director of the Maternal-Fetal Tissue Bank (MFTB). Both CDT members are also Associate Directors in the Institute for Clinical and Translational Science as well as clinical-translational scientists in perinatal biology. This biobanking and clinical data infrastructure provides rigor in patient consent, sample handling and processing, and comprehensive coverage of clinical annotations of samples from EHRs. With over 9000 participants in MFTB, the CDT members have significant experience in data acquisition, management, harmonization, and analysis. The processing and storage of biological samples and clinical data methods of this team have previously been published.[11] Success of the data warehousing has been due to collaboration with the Institute for Clinical and Translational Science Biomedical Informatics (ICTS-BMI) group. The BMI group contains members with expertise in data governance, project management, EHR-data extraction, and data architecture.

Development of the Maternal-Child Knowledgebase: The development of the MCK is an iterative process involving multiple steps: 1) CDT members identify and locate clinically significant distinct clinical variables in the front-facing EHR pertaining to obstetrics and pediatrics. 2) The ICTS-BMI group locates these variables in the transactional database of the EHR. 3) The CDT members validate the veracity of the extracted variables. If the observed extracted data (i.e., gestational age at delivery) is inconsistent with what is expected from the EHR, the CDT and the ICTS-BMI group meet to review the data and potential causes of invalid data in the EHR. 4) With confirmed data variables, the ICTS-BMI group develops novel data tables and user-friendly SQL queries aimed for independent use by investigators with little-to-no SQL experience. The resultant OBstetrics Data Integration Architecture (OBDIA) structure has allowed for an integration of multiple data sources within the institution which includes but is not limited to imaging data, diagnoses, vital sign information, medications, medication administration, and procedures for the maternal-fetal dyad (Figure 1). Information from state immunization registry flow into the EHR and can also be imported into the MCK. The Social Security Death Index is fed directly to the EDW4R. Crucial to the structure of the MCK is the joining of all data in the EHR related to the pregnancy. In our EHR, this is facilitated using the pregnancy episode identification number (preg_episode_id) linking maternal data and neonatal data under a singular pregnancy instance. The pregnancy episode is used to link the maternal and child data and to identify any data related to that particular pregnancy. Each pregnancy, even from the same birthing person, has a unique pregnancy episode identification number. If there are multiple children from the same pregnancy such as in the case of twins, then both children are linked to the mother through the same pregnancy episode identification number; therefore, the architecture of the MCK had to be designed for “one to many” relationships. All pregnancies from the same mother can be found based on a query using a patient identifier for the mother, such as her medical record number or a synthetic identification number. Additionally, the flexibility of the design of the OBDIA allows for multiple sources, even from different EHR vendors to feed into the MCK.

Figure 1. Schematic of Obstetrics Data Integration Architecture.

![Data Integration Architecture (DIA)](image-url)
**OBDIA structure:** Powering this OBDIA architecture is a common semantic layer that makes use of established standardized vocabularies and value sets. For example, as seen in Figure 2, sex can be defined in different ways in different systems. In this example, there are 3 entries in the Ultrasound imaging system and 5 in the Epic system. Using standard vocabularies, each system value will be mapped to corresponding values as the semantic layer. Clearly, this will include storing patient demographic data, patient diagnoses using ICD9/ICD10 codes, storing laboratory results using LOINC codes, and problems using ICD9/ICD10/SNOMED-CT codes. As new data sources arise, the ICTS-BMI group can add them into the current architecture.

**EDW4R viewer:** The ICTS-BMI group has also developed a secure web-based Knowledgebase viewer to allow approved individuals the ability to perform queries of the EDW4R without expert SQL knowledge. Predefined queries retrieve the data and search fields allow for limiting returned results. Query results can also be limited by greater than and less than symbols. The framework is flexible and allows new queries to be added or existing queries to be edited as new data sources become available.

**Data Security:** The ICTS-BMI group relies on the infrastructure provided by the UI and UIHC IT Offices for all physical and virtual servers in Federal Information Security Modernization Act (FISMA) compliant data centers.[12] Faculty, trainees, and staff undergo initial and refresher data security and privacy training via online courses; the UI Human Resources Office logs course completion. ICTS-BMI accesses such logs to ensure that only individuals who completed required courses can query the MCK. ICTS-BMI prohibits efforts to re-identify de-identified data, with audit trails of requests and usage. All users must complete annual security awareness training that includes protection of confidential information system. Administrators must complete technical security standards training. Compliance for security training is tracked.

**Data Governance and Access:** Access to all data is legislated by the CDT and the ICTS-BMI. Access to the clinical data may be obtained through the CDT who act as the honest broker of the data. A formal data request is made by meeting with the CDT to discuss goals of the study and requested data elements. The data dictionary is discussed with the requestor. Once these data elements are agreed upon, the data requestor files an external data sharing request form detailing the study and the data elements requested. Once Institutional Review Board Approval is provided, the CDT provides the relevant data through a REDCap file repository [13] or another secure system.
Results:

The entire MCK data architecture includes over 60,700 pregnancies within the EHR since the OB/GYN module for the EHR was implemented at the institution. A web-based interface has been developed to allow the CDT access to a large amount of distinct data elements within the EHR knowledgebase structure. Figure 3 depicts an example of a simple query in which a specific ICD-10 diagnosis code can be entered in the query box and in real-time all the patients in the MCK with the corresponding diagnosis code are displayed with other corresponding identifying information. Predefined queries were built in the novel viewer which are focused around: maternal and baby diagnoses, procedures, clinical laboratory results, vitals, delivery summary information, and maternal medications and social history. The query ability was expanded to allow for the use of > or < for results. These data then can be exported to .csv or .xls files. Importantly, queries can be performed on mother and/or child data to identify the pregnancy episode in which the outcome occurred. A full set of data can then be extracted based on the pregnancy episode.

The MCK also provides strong support for our biobanking efforts. A field was added that identifies studies in which the patient is enrolled. This can be used to extract the relevant clinical information from women who consented to share data and biological specimens.[11] Therefore, it then also supports all projects utilizing the Maternal Fetal Tissue Bank. For example. The MCK was used to collect the clinical data for a study measuring whether a specific protein can predict successful induction of labor. The MCK facilitates the data collection for the 64 currently signed user agreements for the Maternal Fetal Tissue Bank.

The use of the MCK has accelerated many publications, collaborations, health surveillance projects, and the obtaining of large NIH and foundation grants. It expedites de-identified data requests for project feasibility that are frequently requested for grant applications. Over the first two years of its implementation, it has supported projects ranging from geospatial mapping of obstetric patient outcomes to outcomes of gestational diabetics treated with a specific insulin to the Iowa Maternal Health Innovation Program to reduce rural severe maternal morbidity and mortality. It has also been able to support our collaborations with other institutions,[14] There have been 37 requests for data. Because this data extraction is performed using the predefined queries of the knowledgebase, it can be performed by a domain expert as opposed to a data architect. As a result, this reduces the need to clarify the intent of the data request and reduces the number of simple requests to data architects. Consequently, they have more time to focus on complex requests and/or to work on developing new knowledgebases or other data innovations.

Discussion:

We have created a robust content-specific knowledgebase architecture in maternal and child health that accelerates the acquisition and utilization of large amounts of data. An OB/Gyn focused enterprise data warehouse for research has proven to be a novel crucial tool in the acceleration of maternal and neonatal health based clinical/translational research. While the MCK data architecture is reliant on harmonized and verified clinical data, its development and expansion continue in order to address the evolving needs of investigator-driven studies. Development of a robust common OBDIA underlying the MCK will allow for a robust integration of clinical data sets from disparate EMR systems. This is particularly useful for rural hospital systems with limited resources to devote to bioinformatics. In any research-resource limited setting regardless of rurality, a tool such this will be widely applicable and useful to all states as OBDIA and other derived tools like it will facilitate clinical data acquisition in these areas. Enabling clinical data acquisition in these areas will broaden the ability to compare health outcomes.
The need for validated data requires engaged content experts. From our experience in building and deploying the MCK, it can take significant time to identify the “true” variables within the EHR when the same information is documented in multiple locations by different providers. It is important to determine which piece of data can and will be considered the truth by most investigators who would be the end users of that data. In addition, there can be data points from the same visit (such as a blood pressure being taken twice at an appointment to check the first measurement) or the same inpatient day (such as vitals being taken every 1-2 hours). For multiple points of data, it is necessary to display data in a format that is not overwhelming and can be easily interpreted. We overcame these challenges by having dedicated content experts work with the technical team and iteratively test the results and potential display/report options.

The MCK in conjunction with the novel data viewer has enabled broader access to data, without the requirement of dual-data entry or extensive data training. Previously, the current flow of obtaining and collating clinical data for analysis involved the 1) manual extraction of the information from medical records in the EHR by a trained local nurse, care provider, or research team members; 2) dual entry of this clinical data into a database, such as REDCap[13], by the trained staff; and 3) regular data veracity audits by the study PI and data veracity team. Clearly, there are many levels at which this system may break down from data entry to difference in understanding the data dictionary. A standard architecture minimizes ambiguity and provides a solid foundation for accurate research and data analyses. The MCK has increased our department’s data awareness, encouraging further development of data-driven research and standardization of clinical data entries by adding more discreet fields in the EHR. It has also provided a model on our campus for other domains (cancer, infectious disease, neurodevelopment) in developing their own data knowledgebases.

Prior to the MCK, much EHR data could be retrieved by the ICTS-BMI team. However, it resulted in repeated queries for the same information for multiple requests. With the implementation of the MCK in conjunction with the novel data viewer, more requests can be fulfilled without involving an ICTS-BMI team member. As a result, the ICTS-BMI can work to expand the MCK contents and viewer architecture. Thus, the design and implementation of the MCK has availed personnel resources to continue innovating and making better use of their expertise. Future work in designing graphical overviews of data will also assist in increased use of the data with less input required by the ICTS-BMI team.

The concept of harmonized datasets and databases is not a new one in obstetrics and gynecology. For example, the international perinatal research group, Global Pregnancy Collaboration (CoLab), has focused on stimulating collaboration to understand the causes of and develop new interventions for adverse pregnancy outcomes throughout the world by sharing biological samples and clinical data. This is in part achieved through published data harmonization techniques.[15] The group further developed a harmonized shared database (term COLLECT) that could be used in any international perinatal research setting such that the data, if given appropriate permissions, could be shared among the CoLab members to increase power and diversity for studies.[16] Further, countless multicenter NIH clinical trials and studies in obstetrics and gynecology rely on harmonized data collection. Yet, the data collection largely relies on “hand extraction” of data from medical records or one-time uploads of data to a database. To our knowledge, the Maternal Child Knowledgebase and Datamart Viewer is the first published data architecture and user-friendly viewer of its kind in maternal and neonatal health focused on automatically extracting data from the EHR that is refreshed without manual updates.

Current MCK limitations include being limited to EHR data only, so no social determinants of health data and other relevant datasets are currently included. Yet, given its ability to be modified, as these social determinants of health data are more distinctly defined in the EHR, they may be added to the data structure. Additionally, subsets of patients are being surveyed regarding these variables and these results can be implemented into the OBDIA. MCK would also benefit from having a FHIR interface that would enable using it as a canonical data source to feed research data networks[10] Another limitation is that paternal health is not expressly linked within our data structure but may be a potential area of expansion. Currently, data is limited to what is housed within our singular institution’s EHR. The data is also limited by a direct link existing between the clinical data of the mother and child; there is no direct connection that is already established between the child’s and father’s medical records. This important addition will require future work. While there are clear legal and ethical considerations to be addressed when working with healthcare information, we are currently working with other institutions in Iowa to develop a shared knowledgebase. Finally, while the MCK is constantly being updated, there is a data lag of 2-7 days. Yet the ability to perform retrospective analyses using 27000+ pregnancies over the course of 10 years, makes this data lag insignificant.
The future challenges of this knowledgebase lie in bringing research outcomes back into the clinic to drive clinical decision support. While common current data requests may include up to hundreds of variables, the MCK is built to access almost all discreet data in the EHR. Since the implementation of the MCK, we have successfully supported numerous groups who utilize our harmonized data model. The benefits of this system have been clear: there are an increased number of collaborations, positive grant reviews, and an increased ability to rapidly provide validated data related to maternal health. The flexible design of the system has also allowed us to quickly add variables needed for one principal investigator to be available to the larger community. For example, studies related to rural health resulted in the addition of the rural-urban commuting code (RUCA) for each subject’s residence.

Our long-term goal is to expand data collection to cover the lifespan of women to facilitate addressing the major shortcomings in healthcare surveillance and research and to support work addressing the disparities that exist in research and healthcare delivery. The adaptability of our system combined with its disease-agnostic nature support its widespread application across the clinical and translational spectrum. However, delays in the refresh of data and the addition of outside data source limits the clinical usefulness of this tool. Rather, the tool is a unique resource for long-term planning to improve health and for translational research.

Mother-child data represents a unique circumstance where one patient becomes two and addresses the need to link these two patients’ information. The MCK facilitates exploration of this data and enables future work particularly in the in the realm of the developmental origins of disease as in utero exposures may affect not only the long-term health of the mother but also the offspring. Further because every pregnancy episode from a mother who received prenatal care at UIHC is included, not only can health surveillance throughout one pregnancy can occur, but surveillance across multiple pregnancies is also possible. This may provide useful information about the effects of different exposures during pregnancies and their effects on health outcomes.

Currently, there are over 560 customers with approximately 2400 hospitals in the US, representing 225 million patients in the US, that utilize the same EHR system as our institution.[17] Other medical record systems also employ pregnancy summary sheets that join medical information from throughout pregnancy and briefly into the postpartum period or algorithms can be employed to identify medical events during pregnancy.[18 19] While the exact architecture may need to be adjusted to be utilized for health systems with a different electronic medical record vendor, it is possible to employ a similar structure to capture medical events related to pregnancy. Therefore, our approach has broad applicability. Ultimately, the goal is to combine these datasets for improved health surveillance and translational science. With time, generational health analyses will be achievable with more granularity than has been previously possible.

**Conclusion:**
The development and implementation of a Maternal Child Knowledgebase has significantly improved our capacity to conduct pregnancy-related research and health surveillance. The novel data viewer has eliminated the need for SQL coding knowledge in order to query MCK data. Thus, this powerful tool has been vital for expanding data access to investigators to and reducing the need for data queries to be fulfilled by ICTS-BMI. The ICTS-BMI now has more bandwidth to innovate data delivery.

**References:**


Integrating Patient-generated Digital Health Data into Electronic Health Records (EHRs) in Ambulatory Care Settings: EHR Vendor Survey and Interviews

Ryan J. Shaw, PhD, RN1,2, Mina Boazak, MD, Victoria Tiase, PhD, RN, Gloria Porter, MSN, RN, Jedrek Wosik, MD, Sarah Bumatay, LeAnn Michaels, Julie Stone, MPA, Deborah Cohen, PhD, Rowena Dolor, MD, MHS

1Duke Mobile App Gateway, Clinical & Translational Science Institute, Duke University School of Medicine, Durham, NC, USA; 2Duke University School of Nursing, Durham, NC, USA; 3Department of Psychiatry and Behavioral Sciences, Duke University School of Medicine, Durham, NC, USA; 4The Value Institute, New York-Presbyterian Hospital, New York, NY, USA; 5Division of Cardiology, Department of Medicine, Duke University School of Medicine, Durham, NC, USA; 6Oregon Rural Practice-based Research Network, Oregon Health & Science University, Portland, OR, USA; 7OCHIN, Portland, OR, USA; 8Department of Family Medicine, Oregon Health & Science University, Portland, OR, USA; 9Duke Primary Care Research Consortium, Duke Clinical & Translational Science Institute, Durham, NC, USA; 10Department of Medicine, Duke University School of Medicine, Durham, NC, USA

Abstract

Data traditionally collected in a clinic or hospital setting is now collected electronically in everyday environments from patients, known as patient-generated health data (PGHD). We conducted informal interviews and collected survey data from major ambulatory care EHR vendors that serve the majority of the U.S. market to collect information on how their clients are integrating PGHD into EHRs. Of the 9 EHR vendors contacted, 6 completed the survey and 5 participated in a 45-minute interview. Feedback from the vendors included how PGHD use has steadily risen over the past decade and how the COVID-19 pandemic accelerated PGHD use. Pathways for data from devices or surveys to be brought securely into the EHR are increasing. While promising, adoption of health IT systems has its challenges. There are disparities in EHRs, devices, and applications. We concluded that more supportive policies are needed to advance PGHD integration.

Introduction

The U.S. healthcare system is in a transitional period. Data traditionally collected in a clinic or hospital setting is now able to be collected electronically from patients. These data are known as patient-generated health data (PGHD). PGHD are “health-related data created, recorded, or gathered by or from patients (or family members or other caregivers) to help address a health concern.” The potential for PGHD to impact health is significant. By providing insights into the day-to-day health of an individual, patients and clinicians can employ better strategies to prevent and manage acute and chronic conditions, and clinicians and scientists can use these data to generate and apply analytical techniques to improve risk prediction and diagnoses. The benefits of PGHD reach across care facilities and diverse geographic locations through web-based interoperable data exchange to deliver more precise treatment and self-management assistance to broad populations.

As healthcare moves beyond EHR implementation, the integration of PGHD from connected devices, including mobile health technologies, is gaining speed. Companies like Apple Inc. have the ability for patients to aggregate their health records from multiple sources on an iPhone and integrate data via authentication into health system patient portals such as Epic’s MyChart. It is also possible to integrate third-party data, such as patient-generated blood glucose levels, into the EHR via Apple HealthKit. PHD integration capability is quickly expanding into Android platforms with Google Fit and through data aggregation companies such as Validic and Xealth.

Ambulatory care practices with access to PGHD in their electronic health records (EHRs) may be able to improve patient outcomes, care coordination, quality, and cost-effectiveness. The opportunity and need for PGHD became apparent.
in 2020 when the novel coronavirus pandemic abruptly replaced many in-person primary and specialty care visits with telehealth—such as eVisits and telephone calls. Yet identifying which data are needed and supporting patients and clinicians through data capture and transfer into EHRs is highly complex. Effective use of PGHD in clinics poses many challenges, including clinician and patient burden, poor usability, workflow integration challenges, and the potential to exacerbate health inequities. The Office of the National Coordinator for Health Information Technology (ONC) outlines multiple technical challenges related to accuracy of measurements, data provenance, interoperability, implementation, and privacy and security concerns in the data lifecycle (i.e., collection, transmission, storage, and analysis). In addition, selecting valid devices from an increasing number of options, integration into new care delivery models, costs for patients, equitable access to technology, and inadequate information technology (IT) literacy are among the many challenges facing adoption of PGHD nationwide.

The Agency for Healthcare Research and Quality (AHRQ) commissioned an environmental scan to inform the development of a practical guide that ambulatory care settings can use as they approach the use of PGHD for patient care. As part of the environmental scan, from November 2020 to February 2021, we collected survey data and interviewed major ambulatory care EHR vendors that serve the majority of the U.S. market to collect information on how their clients are integrating PGHD into EHRs.

Methods and Approach

Survey and Interview Guide Development and Approach. We contacted ambulatory EHR vendors (N=9) that serve 95% of the U.S. market to collect information on how their clients are integrating PGHD into EHRs. We assessed what PGHD their clients are using or plan to integrate into their EHR, including PGHD type (e.g., biometric, patient activity, questionnaires, health history), PGHD transfer (e.g., active, passive), and technical approaches (e.g., HL7, APIs, Bluetooth). We asked about interoperability standards (e.g., SMART, FHIR, HL7v2, web services, extensible markup language [XML], and consolidated-clinical document architecture [C-CDA]); whether design schemas such as Open mHealth and standardized medical coding terms are leveraged (e.g., SNOMED, LOINC, RxNORM); and what developer platforms (e.g., Apple HealthKit, Google Fit) and which tools, products, and 3rd-party companies (e.g., Fitbit, Garmin) integrate data into their EHRs.

To collect this information, we invited vendor representatives (from November 2020 to February 2021) to complete an online survey, and then participate in a follow-up interview. Vendor representatives were familiar with the processes involved and the state of PGHD integration into their company's EHR. The survey and interview guide were developed by our team with feedback and expert opinion from a pre-identified technology expert panel (TEP). Through iteration we refined the question set with approval from AHRQ (See Appendices C and D). We contacted vendors at least twice in attempt to have them complete an online survey in Qualtrics. Data were analyzed using descriptive statistics.

Following completion of the survey, we invited vendors to a 45-minute video interview. Vendors were asked 10 questions exploring factors contributing to the successes and challenges of integrating PGHD into EHRs. Interviewers took notes during the interview and asked permission to record interviews. Data regarding the successes, challenges, and resources of PGHD were analyzed using content analysis to identify recurring themes in the interviews. We described other feedback from vendors as a summary narrative.

Results

Of the 9 EHR vendors contacted, 6 completed the survey and 5 participated in a 45-minute interview. Vendors interviewed serve approximately 80% of the US ambulatory care market. EHR vendors described factors that contribute to integration of PGHD into EHRs. Table 1 displays results from the survey. Table 2 describes themes that arose around success, challenges, and resources needed.

EHR vendor survey. Of the six vendors who responded, nearly all (n=5, 83%) stated they allow for PGHD to be integrated (ingested). The timeline since integration began ranges between two to ten or more years. The five that allow for PGHD to be ingested are further described. All (n=5, 100%) provide pre-built and custom-built functionality to process and manage PGHD, and PGHD is part of the original contract for some (n=3, 60%) and an add-on for others (n=2, 40%). Nearly all (n=4, 80%) allow for a "Bring your own device" (BYOD) model, and most (n=3, 60%) allow for PGHD to be received outside out of the patient portal.
Some described functionality to notify providers and patients (n=3, 60%) if PGHD need action or are out of range. Vendors provided further details as comments to these questions. It is an implementation decision between vendors and clients to select what type of notifications the client (clinic) may want based on the PGHD. PGHD may appear in a dashboard for clinicians to review to identify patients with high risk in need for outreach. Triggers can automatically send messages to providers or page a nurse pool based on incoming data. Notifications can also be developed that remind patients to complete patient reported outcomes, submit data, or perform action based on data received.

All (n=5, 100%) stated their EHR has the capability to send patient data from the EHR to mobile health apps. All use iOS HealthKit (n=5, 100%), some use Android’s Google Fit (n=2, 40%) and other partner platforms (n=2, 40%) to integrate PGHD. HealthKit is easier for vendors to leverage due to the maturity of the Health App which provides data and security standardization. Three vendors (60%) explained their tool provided these data from mobile health apps in graphical format for clinicians within the EHR. The variety of data that could be ingested varied by the aggregator source. iOS HealthKit allows for a variety of data to be tethered to the Apple Health App. Thus, any data integrated with HealthKit could be pulled. Similarly, partnering with a data aggregator company like Validic or Raziel Health allows for additional types of data to be pulled into an EHR. Other partner vendors reported included Livongo, TytoCare, IdealLife, Fitbit, Garmin, Omron, Qardio, iHealth, Welch-Allen, and Withings. While it may be technically possible to pull in dozens if not hundreds of data types from remote health monitors and surveys, the need or value to do so must be tied to a care delivery model.

With regard to the transfer of data, all (n=5, 100%) allow for passive transfer while most (n=4, 80%) allow for push, active, and pull. The technical approach by vendors varies with FHIR (n=4, 80%) standard APIs (n=3, 60%), and web services (n=3, 60%). The use of standardized medical coding terminologies (e.g., SNOMED, LOINC) varied across vendors. Leveraging FHIR was recommended and encouraged to create data standards across the industry to facilitate integrating mobile health app data into EHRs. None of the vendors use design schemas such as Open mHealth or IEEE P1752 standards process.

One vendor (20%) stated they were able to consume or translate incoming PGHD into another language, Spanish, although it was unclear if anyone has done this yet. Three (60%) provide PGHD resources through their patient portal. These resources may include instructions for connecting devices, collecting and uploading data, and what to do if results are out of normal range. Some vendor systems allow patients to connect to supported devices without practice assistance or tech support (n=3, 60%), and some require a clinic to activate prescribed devices (n=3, 60%).

Table 1. EHR vendor survey responses (N=6)

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
<th>Number (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does your EHR allow for PGHD to be ingested? (N=6)</td>
<td>Yes</td>
<td>5 (83)</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>1 (17)</td>
</tr>
<tr>
<td>n=5 below</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Do you allow for a bring your own device (BYOD) model?</td>
<td>Yes</td>
<td>4 (80)</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>1 (20)</td>
</tr>
<tr>
<td>To process and manage PGHD does your EHR require custom-built or pre-built functionality, or both?</td>
<td>Both</td>
<td>5 (100)</td>
</tr>
<tr>
<td>Is PGHD inclusion part of the original contract with clients or an add-on?</td>
<td>Original contract</td>
<td>3 (60)</td>
</tr>
<tr>
<td></td>
<td>Add-on</td>
<td>2 (40)</td>
</tr>
<tr>
<td>Can PGHD be received outside of the patient portal?</td>
<td>Yes</td>
<td>3 (60)</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>2 (40)</td>
</tr>
<tr>
<td>Is PGHD accessible by providers/health system to intervene?</td>
<td>Yes</td>
<td>4 (80)</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>1 (20)</td>
</tr>
<tr>
<td>Does your EHR have functionality to notify providers regarding PGHD (e.g., exists, needs action, or is out of range)?</td>
<td>Yes</td>
<td>3 (60)</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>2 (40)</td>
</tr>
<tr>
<td>Does your EHR have functionality to notify patients regarding PGHD (e.g., exists, needs action, or is out of range)?</td>
<td>Yes</td>
<td>3 (60)</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>2 (40)</td>
</tr>
<tr>
<td>Does your EHR have the capability to send patient data from the EHR to mobile health apps?</td>
<td>Yes</td>
<td>5 (100)</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>
Clinical focus and integration actionable health outcomes and create value. This includes clinical champions, a patient advocate, and resources needed to make the integration of PGHD actionable. Organizational support and readiness to use PGHD in a meaningful way are needed for success. A variety of factors influence the success of PGHD to improve health outcomes and create value. This includes clinical champions, a patient-focused approach, and data and device governance in which PGHD are part of a targeted care delivery model. Other factors include interoperability and economic viability. Vendors described that challenges arise when a well-resourced plan with all stakeholders is not the approach. Resources such as educational support and technical support are key.

**EHR vendor interviews.** Table 2 describes themes and descriptions from EHR vendors on the successes, challenges, and resources needed to make the integration of PGHD actionable. Organizational support and readiness to use PGHD in a meaningful way are needed for success. A variety of factors influence the success of PGHD to improve health outcomes and create value. This includes clinical champions, a patient-focused approach, and data and device governance in which PGHD are part of a targeted care delivery model. Other factors include interoperability and economic viability. Vendors described that challenges arise when a well-resourced plan with all stakeholders is not the approach. Resources such as educational support and technical support are key.

**Table 2.** EHR vendor interview themes: Factors that contribute to the success and challenges of making PGHD integration actionable and resources needed to support PGHD

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
<th>Number (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does your EHR allow for the push/active or pull/passive transfer of PGHD?</td>
<td>Push, active and pull&lt;br&gt;Passive</td>
<td>4 (80)&lt;br&gt;5 (100)</td>
</tr>
<tr>
<td>What technical approach to PGHD integration does your EHR support?</td>
<td>HL7&lt;br&gt;EHPR&lt;br&gt;Standardized APIs&lt;br&gt;Web services</td>
<td>4 (80)&lt;br&gt;4 (80)&lt;br&gt;3 (60)&lt;br&gt;4 (80)</td>
</tr>
<tr>
<td>Does your EHR use design schemas such as Open mHealth? (OmH; IEEE P1752 standards process)?</td>
<td>Yes&lt;br&gt;Not sure&lt;br&gt;No</td>
<td>0 (0)&lt;br&gt;4 (60)&lt;br&gt;2 (40)</td>
</tr>
<tr>
<td>What platforms does your EHR partner with to integrate PGHD?</td>
<td>iOS HealthKit&lt;br&gt;Android Google Fit&lt;br&gt;Other</td>
<td>5 (100)&lt;br&gt;2 (40)&lt;br&gt;3 (66)</td>
</tr>
<tr>
<td>Does your EHR have the ability to translate PGHD in different languages?</td>
<td>Yes&lt;br&gt;Not sure&lt;br&gt;No</td>
<td>1 (20)&lt;br&gt;1 (20)&lt;br&gt;3 (60)</td>
</tr>
<tr>
<td>Does your EHR have the ability to consume PGHD in different languages?</td>
<td>Yes&lt;br&gt;Not sure&lt;br&gt;No</td>
<td>1 (20)&lt;br&gt;1 (20)&lt;br&gt;3 (60)</td>
</tr>
<tr>
<td>Are there readily available resources through your patient portal for patients about PGHD?</td>
<td>Yes&lt;br&gt;Not sure&lt;br&gt;No</td>
<td>3 (60)&lt;br&gt;1 (20)&lt;br&gt;1 (20)</td>
</tr>
<tr>
<td>Patients have the ability to easily connect to supported devices without practice assistance or tech support&lt;br&gt;Patients require a clinic to activate prescribed devices</td>
<td>Yes&lt;br&gt;Yes</td>
<td>3 (60)&lt;br&gt;3 (60)</td>
</tr>
</tbody>
</table>

**Organizational support and readiness.** Organizations need to invest and prepare for the use of PGHD. This requires consistent organization-wide processes on how to leverage PGHD, marketing representation to create value for patients, and buy-in across the enterprise.

**Clinical champions.** Physicians, nurses, case managers, social workers, and other stakeholders need to advocate and champion the use of PGHD for patient care. Champions must be maintained over time, or interest may fade.

**Robust care delivery model.** PGHD need to be tied to a clinical focus (e.g., congestive heart failure [CHF], hypertension). This allows for data and devices to be selected that are appropriate for specific clinical outcomes and targeted to the care delivery model allowing patient self-management and clinical decision-making. This permits data governance and protocol development to understand how to act upon the data by patients and providers. Processes must be used to create value and not increase burden on the care team.

**Data governance.** PGHD need to be valid, accurate, and well managed with rules to make them useful, timely, interpretable, and effective. Collection and interpretation must be tailored to the clinical focus and population. Protocols and triggers need to be incorporated into the EHR to encourage patient self-management and clinician decision making. Data analytics are needed to discern signal from noise. Decisions need to be made as to how data will be analyzed over time and in a tailored time window. Data density, when data are missing or too frequent, requires protocols for how to deal with changes in data frequency over time. Data need to be aggregated across sources and visualized in a dashboard with clinical decision support tools.
• **Device governance.** Device management could be a BYOD model, managed by clinics or by a vendor. Devices could be delivered as a kit that collects data specific to the clinical focus. Multiple devices may be needed but can contribute to complexity. The approach may be influenced by the type of EHR vendor the clinic or health system contracts with.

• **Interoperability.** Data need to be exchanged seamlessly across geographic boundaries between disparate organizations, systems, and sources. Consistent standards are critical and may include HL7 and FHIR.

• **Patient-focused approach.** Data and devices need to be useable and appropriate for the target population. The demographics of the patient population need to be considered (e.g., use of iOS or Android, technical literacy, broadband access, physical dexterity).

• **Technical support.** A fundamental requirement is to provide support for patients across the lifespan in diverse environments. Support ought to be provided by a technical person from the clinic or organization, the device manufacturer, or outsourced. Clinical staff, such as RNs, are not the best fit for this role.

• **Economic viability.** The use of PGHD needs to be incorporated into the business model of the organization to demonstrate revenue generation or cost savings.

**Challenges: Factors that contribute to challenges of making integration of PGHD actionable**

• **Lack of regulations and industry standards.** Data need to be standardized across the industry. There are disparities in EHRs, devices, and applications. Not all EHR vendors use consistent standards, such as FHIR. Standards for some EHR vendor platforms are not as mature as others.

• **Poor data governance.** Protocols are needed to create value from PGHD. Analysis of disparate data sources and determining how and when to act upon data are critical. Organizations may struggle with the legality of PGHD.

• **Patient technology hurdles.** Technical and data literacy must be considered for the target population. Access to broadband internet, particularly in rural locations, may be a hurdle. Patients need to be proficient with how to use devices, particularly multiple devices, which can be amplified for patients who have multiple chronic illnesses and who are often older.

• **Manual data entry and lack of analytics.** Automated data entry is needed when possible. Resources should be dedicated so that data are programmed to be automatically ingested by software to create meaning and value for patients and clinicians.

• **No care delivery model.** Responsibility for the data is needed and it needs to be tied to health outcomes to select the most appropriate data type and device in order to create value. There remains a lack of national standards around care models for PGHD.

**Resources needed to support PGHD**

• **Clinical application and data processes.** Organizational investment is needed to develop use cases for PGHD in a variety of care models. Data procedures include governance, protocols, and processes that guide the use of PGHD for clinical decision making and patient self-monitoring that meet patient outcome, regulatory, and legality needs.

• **Clinical workflow capacity.** Clinicians need dedicated time to incorporate PGHD into their clinical workflow.

• **Educational support for patients and providers.** Education and training need to be provided to all end users to understand the benefits and limitations to PGHD.

• **Technical support.** Technical support to patients across the lifespan in diverse environments provided by a technical person.

Vendors described that most clients collect PGHD through surveys that are offered through their patient portal. This may include information being collected before or between clinic visits. COVID-19 has increased use of surveys to collect information on symptoms, exposure, and testing. Less common is the use of remote monitoring devices.

The use of remote monitoring devices is offered in a variety of ways, depending on the relationship between ambulatory care clinics and the EHR vendor. The vendor may offer devices in a ‘kit model,’ where the patient receives a suite of devices tethered toward a clinical target area. For example, patients with hypertension may receive a Bluetooth-enabled blood pressure monitor and in-home wireless scale. Similar kits for COVID-19 symptoms are on the rise for temperature, blood pressure, and pulse oximetry remote monitoring. The rise of the Hospital at Home model is a similar and quickly growing care delivery model that is accelerating the use of PGHD.
integration. Third-party vendors may offer device kits and provide tech support. These are negotiations between the clinic or health system, EHR vendor, and device vendor.

Vendors described that decisions to implement are driven by the clinical side, while the EHR and device vendor provide development support. Development, implementation, and testing usually takes 6–12 months, though prioritized topics can have accelerated timelines. Development and development costs are unique to the EHR vendor and their relationship with the clinics. This influences the way in which PGHD is financed, which could be by per-patient transaction.

EHR vendors stated their systems undergo full security assessments. Data are more protected once they come into the EHR ecosystem, which could be through a patient portal or via API from a device company. Security with devices and their associated apps needs to be worked out with the respective device companies. Risk is held on the patient’s side before data cross into the EHR, and data may not fall under privacy and security regulations. Patients should be encouraged to use standard security approaches, such as user authentication, and to limit health information exchange with third-party apps. Limited regulations around PGHD and consumer-based devices makes this an ongoing challenge.

Discussion

The COVID-19 pandemic accelerated the use of digital health. As social distancing measures were enforced, care providers were motivated to collect data from patients remotely. While initially focused on telehealth models of care using video visits, as the pandemic unfolded, other models of care that integrate PGHD grew. While telehealth programs expanded, EHR vendors described how COVID-19 slowed implementation of projects nationally for PGHD in the beginning of the pandemic. At the same time, the rapid shift to remote care delivery promoted pockets of innovation in the use of PGHD for monitoring COVID-19-positive patients. One such example was a partnership between Cleveland Clinic and Epic. As reported by news media, they developed a 14-day interactive care plan through the patient portal where patients can enter symptoms, temperature, and oxygen at home, while clinicians monitor them from afar.10

Feedback from the EHR vendors described how PGHD use has steadily risen over the past decade as access to the internet and smartphones have proliferated. Pathways for data from devices or surveys to be brought securely into the EHR are increasing. Geographic barriers are falling and allowing for PGHD to be transmitted in near real-time from environments that patients spend more of their lifetime in. There are a variety of models that can be adopted in the integration of PGHD. Ambulatory care clinics are able to partner with EHR vendors, and other partner vendors such as data aggregators (e.g., Validic, Raziel Health) or device companies (e.g., Omron, Qardio, Withings) to build frameworks and support for PGHD to be used for a variety of clinical needs.

Vendors reported using iOS to integrate a variety of PGHD to their EHRs and largely support the Apple ecosystem. Android was only supported by 40% of respondents. This is a concern, given that approximately 38% of the U.S. market is Android.11 Few vendors allow for multi-language support, with only one reporting to allow for data to be ingested in Spanish. These factors have the potential to contribute to disparities in healthcare access among underserved populations.

There are disparities in EHRs, devices, and applications. Vendors discussed the lack of regulations and enforced standards around PGHD. While standards such as FHIR are encouraged, they are not used exclusively. Fortunately, new interoperability rules from ONC, under the 21st Century Cures Act, will boost the exchange of data through APIs and FHIR standards.

A growing number of telehealth vendors provide technology-enabled services that integrate PGHD.12 Banner Health, for example, partners with Cerner and Xealth to simplify how clinicians prescribe digital health for telehealth and remote patient monitoring. Clinicians can prescribe digital therapeutics, smartphone, and internet apps as tools that connect with the EHR for chronic disease management, behavioral health, maternity care, and surgery preparation and post-surgical monitoring.13
Limitations. While we attempted to collect data from vendors that serve over 95% of the U.S. ambulatory care market, we were not able to collect data and conduct interviews with all vendors contacted. Questions did not focus on implementation at the vendor versus level. We did not assess issues around accuracy, reliability and utility of PGHD. Future research should address these issues and conduct similar inquiries with health systems and care providers.

Conclusion
Capturing PGHD facilitates patients and clinicians to better understand and predict illness dynamics and to develop approaches to improve health outcomes and deliver personalized care. The COVID-19 pandemic accelerated the use of PGHD, as care providers were encouraged to collect data from patients remotely. Feedback from the EHR vendors highlighted the evolution that PGHD is undergoing as tools for patient care delivery. Over the past decade, PGHD use has steadily risen as access to the internet and smartphones have proliferated. Pathways for data from devices or surveys to be brought securely into the EHR are increasing. While promising, adoption of health IT systems has many challenges. There are disparities in EHRs, devices, and applications. Vendors discussed the lack of regulations and enforced standards around PGHD. While standards such as FHIR are encouraged, they are not used exclusively. The ONC Cures Act Final Rule has provided much-needed regulation, structure, and incentives to help alleviate challenges. Nevertheless, more supportive policies are needed to support integration of PGHD into EHRs and clinical care.

Funding: This work was supported under Contract 75Q80120D00019 / 75Q80120F32001, Agency for Healthcare Research and Quality. AHRQ Publication No. 21-0031.

References
Predicting Unplanned 7-day Intensive Care Unit Readmissions with Machine Learning Models for Improved Discharge Risk Assessment

Katherine Shi, MS1,2, Vy Ho, MD1,3, Joanna J. Song, BS1, Katelyn Bechler, BS1, Jonathan H. Chen, MD PhD 4,5

1Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA; 2Northwestern University Feinberg School of Medicine, Chicago, IL; 3Division of Vascular Surgery, Department of Surgery, Stanford University School of Medicine, Stanford, CA; 4Division of Hospital Medicine, Department of Medicine, Stanford University School of Medicine, Stanford, CA; 5Stanford Center for Biomedical Informatics Research, Stanford University School of Medicine, Stanford, CA

Abstract
Unplanned readmission to the intensive care unit (ICU) confers excess morbidity and mortality. We explore whether machine learning models can outperform the current standard, the Stability and Workload Index for Transfer (SWIFT) score, in assessing 7-day ICU readmission risk at discharge. Logistic regression, random forest, support vector machine, and gradient boosting models were trained and validated on Stanford Hospital data (2009-2019), externally validated on Beth Israel Deaconess Medical Center (BIDMC) data (2008-2019) and benchmarked against SWIFT. The best performing model was gradient boosting, with AUROC of 0.85 and 0.60 and F1-score of 0.43 and 0.14 on internal and external validation, respectively. SWIFT had an AUROC of 0.67 and 0.51 and F1-score of 0.33 and 0.10 on Stanford and BIDMC data, respectively. Machine learning models predicting 7-day ICU readmission risk can improve current ICU discharge risk assessment standards, but performance may be limited without local training.

Introduction
Intensive care units (ICUs) provide the highest acuity care in inpatient medicine, with an 8-19% mortality rate translating to 500,000 deaths in the United States annually1,2. Unplanned ICU readmission, in which a patient is discharged from the ICU to a non-critical care hospital ward and returns non-electively within the same admission, occurs in 4-11% of ICU patients and is associated with a 6 to 7-fold increased risk of mortality3–6. An estimated 18-22% of unplanned readmissions are preventable, representing an opportunity for improved patient care7,8.

Existing ICU discharge decisions are largely driven by subjective clinical judgement and are thus vulnerable to human error. Discharge timing may also be influenced by a variety of non-patient-centered factors such as hospital crowding, staff resourcing, and caregiver fatigue. A multicenter study of 151 hospitals found that ICU readmission rates in teaching hospitals decreased significantly following the implementation of legislation limiting resident work hours9. Currently, the Stability and Workload Index for Transfer (SWIFT) criteria is the only validated risk score for ICU discharge but performs poorly and inconsistently on implementation10–12. Thus, there is a need for standardization in ICU discharge practices.

Implementation of a machine learning model to generate standardized probability risk scores for unplanned ICU readmission may improve discharge decision-making by reducing the cognitive burden placed on caregivers. Previous machine learning attempts to predict ICU readmission have been limited by reliance on single institution data11–17. This study expands upon prior work by including multi-site validation. Machine learning models predicting 7-day ICU readmission were trained and internally validated on Stanford Medicine Research Data Repository (STARR) data from Stanford Hospital and externally validated on the Medical Information Mart for Intensive Care (MIMIC)-IV data from Beth Israel Deaconess Medical Center (BIDMC). We hypothesized that machine learning models can outperform the current standard of ICU discharge, predicting readmission up to 7 days from the time of discharge.
Methods
Patients from Stanford Hospital (2009 - 2019) were used for model training and internal validation, and patients from BIDMC (2008 - 2019) were used for external validation. This study was approved by the Stanford Institutional Review Board. Authors who worked with primary MIMIC-IV data were credentialed PhysioNet users with approved access to the database.

Cohort definition.
Patients who were at least 18 years old and who were admitted to the ICU for at least 24 hours followed by transfer to a lower-acuity inpatient unit were eligible for inclusion. The primary outcome of interest was ICU readmission within 7 days of transfer from the ICU. Patients who fit these criteria were assigned a positive label. In patients with multiple eligible hospital or ICU admissions, only the first eligible stay was selected to optimize model generalizability. Patients discharged directly out of the hospital from the first ICU stay were excluded. Patients who were not readmitted to the ICU and underwent a surgical procedure between first ICU admission to hospital discharge were excluded to discount possible cases of elective “readmission” to the surgical ICU for post-operative care. Similarly, patients who were readmitted to the ICU and underwent a surgical procedure between first ICU admission to ICU readmission were excluded. Patients whose most recent code status was not “full code” (accepts all potential emergent and life-sustaining interventions) were also excluded to ensure all patients in the cohort were eligible for readmission into the ICU. Most recent code status was defined as the last code status update prior to hospital discharge or ICU readmission for patients who were not readmitted or were readmitted to the ICU, respectively.

Feature selection, missingness, and imputation.
Features were selected a priori based on clinical methods of diagnosing, treating, and monitoring organ dysfunction and failure as well as prior models predicting ICU readmission or mortality. From prior work, features from the SWIFT score, a gradient boosting model by Rojas et al, and recurrent neural network by Lin et al were included. Additional features were derived from three ICU mortality scores: Sequential Organ Failure Assessment (SOFA), Acute Physiology and Chronic Health Evaluation (APACHE) II, and Simplified Acute Physiology Score (SAPS) II. Feature values were extracted from time of admission to time of transfer from the ICU. Combinations of max, min, last, and counts of occurrences above or below normal physiologic range were used to capture a comprehensive view of patient status throughout the ICU stay. Additional demographic and hospital stay information were included: age, sex, ethnicity, admission directly from the emergency department, length of first ICU stay, and length of hospital stay defined as duration from hospital admission to first ICU transfer to a different inpatient unit. Continuous variables were not discretized, and categorical variables were given a discrete numerical representation. Feature normalization and pruning for correlation or data record missingness were not applied to ensure a comprehensive clinical view that captures various etiologies for broader disease categories and less prevalent disease conditions. Missing values were replaced using column mean imputation. A total of 115 features were included (see table 1 for full feature set).

Model training, tuning, and selection.
STARR data was randomly split by cases into a training set for model training and selection (60%), validation set for hyperparameter tuning (20%), and test set (20%) for internal performance measurement. Models explored included logistic regression model (LRM), support vector machine (SVM) with linear kernel, random forest, and gradient boosting machine (GBM) from the scikit-learn package v0.24.2. The average results of 5-fold cross validation were used to determine accuracy, area under the receiver operator curve (AUROC), precision, recall, and F1-score (table 2). Optimization for the F1-score was used in preference for high sensitivity and precision in predicting ICU readmission. Logistic regression and gradient boosting were the best performing classifiers and thus selected for further hyperparameter tuning on the validation set.

We performed a grid search to tune hyperparameters. Logistic regression hyperparameters included solver type (stochastic average gradient, Newton’s method, library for large linear classification, limited-memory Broyden-Fletcher-Goldfarb-Shanno), regularization type (L1 or L2) and regularization strength (λ). Manually chosen regularization strengths ranged from 0.01 to 100. Gradient boosting hyperparameters included learning rate, maximum features to consider, maximum tree depth, and number of trees. Four different parameter search grids were built with learning rate ranging from 0.001 to 1, maximum features to consider ranging from 20 to 100, maximum tree depth ranging from 4 to 20, and number of trees ranging from 100 to 1750. Search grids were manually adjusted to evaluate a greater number of hyperparameter values in a smaller range in successive iterations of hyperparameter tuning.
Table 1: Description of full feature set.

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic</td>
<td>age, sex, ethnicity</td>
</tr>
<tr>
<td>Hospital factors</td>
<td>ICU length of stay, hospital length of stay, admission direct from emergency department</td>
</tr>
<tr>
<td>Vital signs</td>
<td>Last, min, max, high, low: temperature, heart rate, mean arterial pressure&lt;br&gt;Last, max, min: SaO2, SpO2</td>
</tr>
<tr>
<td>General</td>
<td>Last, max: C-reactive protein (CRP), erythrocyte sedimentation rate (ESR)&lt;br&gt;Counts of inflammatory markers (CRP, ESR) that are higher than physiologic range</td>
</tr>
<tr>
<td>Cardiac</td>
<td>Last, high: B-natriuretic peptide&lt;br&gt;Last, max, high: troponin</td>
</tr>
<tr>
<td>Respiratory</td>
<td>Number of days intubated, number of hours between extubation and transfer out of ICU&lt;br&gt;Last: PaO2, FiO2, PaO2/FiO2 ratio&lt;br&gt;Max, min: PaO2, PaCO2</td>
</tr>
<tr>
<td>Metabolic</td>
<td>Last, max, high: lactate&lt;br&gt;Last, max, min: pH&lt;br&gt;Last, max, min, high, low: sodium, potassium, bicarbonate</td>
</tr>
<tr>
<td>Renal</td>
<td>Last, max: creatinine&lt;br&gt;Last, max, high: blood urea nitrogen (BUN)</td>
</tr>
<tr>
<td>Endocrine</td>
<td>Max, min: cortisol&lt;br&gt;Counts of glucose measurements greater than or equal to 300 mg/dL</td>
</tr>
<tr>
<td>Gastrointestinal</td>
<td>Last, max: ALT, AST, total bilirubin, direct bilirubin, lipase</td>
</tr>
<tr>
<td>Hematologic</td>
<td>Last, min, low: hemoglobin, platelet&lt;br&gt;Last, max, min, high, low: white blood cell count&lt;br&gt;Last, max, min: INR</td>
</tr>
<tr>
<td>Neurologic</td>
<td>Last, max, min: Glasgow Coma Score</td>
</tr>
<tr>
<td>Musculoskeletal</td>
<td>Last, max: creatine kinase (CK)</td>
</tr>
<tr>
<td>Imaging</td>
<td>Counts of number of electrocardiograms, x-rays, chest x-rays, MRIs, and CTs taken during the first ICU stay</td>
</tr>
<tr>
<td>Medications</td>
<td>Counts of intravenous diuretics, antihypertensives, inotropes or pressors, and sedatives given during ICU stay**</td>
</tr>
</tbody>
</table>

*Last = last value prior to ICU transfer out, max = max value during ICU stay, min = min value during ICU stay, high = count of values above normal physiologic range during ICU stay, low = count of values below normal physiologic range during ICU stay
** Diuretics include furosemide, bumetanide. Antihypertensives include labetolol, nicardipine, hydralazine, nitroglycerin, nitroprusside. Inotropes and pressors include epinephrine, norepinephrine, phenylephrine, vasopressin, dopamine. Sedatives include diazepam, lorazepam, midazolam, fentanyl, hydromorphone, morphine, dexmedetomidine, haloperidol, olanzapine, quetiapine, risperidone.
Internal and external validation.

The best performing LRM and GBM models from hyperparameter tuning were selected for internal and external validation on the STARR and MIMIC test sets, respectively. The SWIFT criteria were also applied to both test sets to benchmark model performance against the best existing standard. A SWIFT score of 15 or above was assigned a positive label as validated by Gajic et al. Primary metrics used to assess model performance on the test sets included accuracy, AUROC, and area under the precision-recall curve (AUPRC). The optimal receiver operator curve (ROC) threshold was selected based on Youden’s J statistic. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1-score were evaluated at this threshold. ROC and precision-recall (PR) curve plots were generated to visualize the tradeoffs at each decision threshold.

Results

From STARR, 286 of 3107 patients (9.2%) were readmitted within seven days of ICU discharge. From MIMIC, 802 of 13,841 patients (5.8%) had an eligible readmission.

Model selection and hyperparameter tuning.

Logistic regression, gradient boosting, support vector machine with linear kernel, and random forest models were trained on STARR. The baseline readmission prevalence in the training set was 9.2% (172/1863). Accuracy and precision were similar across all models. AUROC was highest for gradient boosting, random forest, and logistic regression. Logistic regression had the highest F1-score and recall followed by gradient boosting. See table 2 for details. Logistic regression and gradient boosting were selected for further hyperparameter tuning and validation based on F1-scores. After hyperparameter tuning, the best logistic regression model used the library for large linear classification solver and L2 regularization with strength of 0.01. The best gradient boosting model had a learning rate of 0.004, maximum considered features of 60, maximum depth of 3, and number of trees at 470.

Table 2. Average accuracy, precision, recall, and F1-score from 5-fold cross validation on STARR training dataset. Baseline prevalence was 9.2%.

<table>
<thead>
<tr>
<th>Model</th>
<th>F1-score</th>
<th>Accuracy</th>
<th>AUROC</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.17</td>
<td>0.90</td>
<td>0.75</td>
<td>0.35</td>
<td>0.12</td>
</tr>
<tr>
<td>Gradient Boosting</td>
<td>0.13</td>
<td>0.90</td>
<td>0.80</td>
<td>0.38</td>
<td>0.08</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>0.11</td>
<td>0.90</td>
<td>0.65</td>
<td>0.32</td>
<td>0.07</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.05</td>
<td>0.91</td>
<td>0.79</td>
<td>0.37</td>
<td>0.03</td>
</tr>
</tbody>
</table>

*AUROC = area under receiver operating curve.

Baseline STARR and MIMIC test set characteristics.

Overall, the baseline demographic characteristics are similar between the STARR and MIMIC test sets. Patients in both datasets are mostly men at approximately 60%, majority Caucasian, and on average in at least the 6th decade of life. Notable differences include a larger representation of Asian/Pacific Islander and Hispanic/Latino patients by 5 to 15 percentage points in the STARR dataset. On average, STARR patients have hospital stays that are 2-3x longer than that of MIMIC patients’ hospital stays, measured in this study as time from hospital admission to first ICU transfer out to lower acuity care in the hospital. See table 3 for details.
Table 3: Baseline demographic characteristics in STARR and MIMIC test set cohorts.

<table>
<thead>
<tr>
<th>FEATURE</th>
<th>STARR</th>
<th>MIMIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Readmit (n = 62)</td>
<td>Not readmitted (n = 560)</td>
</tr>
<tr>
<td>Sex, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>23 (37.1)</td>
<td>223 (39.8)</td>
</tr>
<tr>
<td>Male</td>
<td>39 (62.9)</td>
<td>337 (60.2)</td>
</tr>
<tr>
<td>Ethnicity, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>35 (56.5)</td>
<td>286 (51.1)</td>
</tr>
<tr>
<td>African American</td>
<td>7 (11.3)</td>
<td>42 (7.5)</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>5 (8.1)</td>
<td>80 (14.3)</td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>10 (16.1)</td>
<td>104 (18.6)</td>
</tr>
<tr>
<td>Native American/Alaska Native</td>
<td>0 (0.0)</td>
<td>2 (0.4)</td>
</tr>
<tr>
<td>Other</td>
<td>3 (4.8)</td>
<td>31 (5.5)</td>
</tr>
<tr>
<td>Unknown</td>
<td>2 (3.2)</td>
<td>15 (2.7)</td>
</tr>
<tr>
<td>Age (years), mean (std)</td>
<td>57.5 (15.4)</td>
<td>58.4 (17.2)</td>
</tr>
<tr>
<td>Direct admission from ED, n (%)</td>
<td>14 (22.6)</td>
<td>132 (23.6)</td>
</tr>
<tr>
<td>ICU length of stay (days), mean (std)</td>
<td>4.3 (3.9)</td>
<td>2.7 (3.1)</td>
</tr>
<tr>
<td>Hospital length of stay (days), mean (std)</td>
<td>26.1 (18.3)</td>
<td>10.2 (11.8)</td>
</tr>
</tbody>
</table>

* STARR = Stanford medicine Research data Repository; MIMIC = Medical Information Mart for Intensive Care; ED = emergency department; ICU = intensive care unit.

Internal validation with STARR
Baseline STARR test set readmission prevalence was 10%. Gradient boosting and logistic regression performed similarly across most metrics. Both classifiers had an accuracy of at least 0.90, AUPRC of 0.41, and NPV of at least 0.96. Gradient boosting had a higher AUROC of 0.85 and sensitivity of 0.84 compared to logistic regression with an AUROC of 0.81 and sensitivity of 0.68. However, logistic regression saw higher specificity (0.86) compared to gradient boosting (0.78). Predictions with the SWIFT criteria had the lowest accuracy (0.78), AUROC (0.67), AUPRC (0.17), F1-score (0.33), sensitivity (0.53), PPV (0.23), and NPV (0.94) though its specificity and NPV were comparable to the machine learning models. See table 4 for details.

External validation with MIMIC-IV
The baseline MIMIC test set readmission prevalence was 5.8%. On external validation, there was a drop in AUPRC, sensitivity, specificity, PPV, and F1-scores across all models. For logistic regression and SWIFT, all metrics decreased except for NPV. There was a slight increase in GBM accuracy on MIMIC-IV (0.93) compared to STARR (0.90) data. See table 4 for details. Figures 1-3 show ROC and PR curves that illustrate the decrease in performance from internal to external validation across all models.
Table 4: Classifier and SWIFT performance metrics on STARR and MIMIC test sets. F1-score, sensitivity, specificity, PPV, and NPV evaluated at optimal ROC threshold determined by Youden’s J statistic.

<table>
<thead>
<tr>
<th></th>
<th>STARR</th>
<th>MIMIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Current Standard</td>
<td>Select ML Models</td>
</tr>
<tr>
<td></td>
<td>SWIFT</td>
<td>LRM</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.78</td>
<td>0.91</td>
</tr>
<tr>
<td>AUROC</td>
<td>0.67</td>
<td>0.81</td>
</tr>
<tr>
<td>AUPRC</td>
<td>0.17</td>
<td>0.41</td>
</tr>
<tr>
<td>F1-score</td>
<td>0.33</td>
<td>0.45</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.53</td>
<td>0.68</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.81</td>
<td>0.86</td>
</tr>
<tr>
<td>PPV</td>
<td>0.23</td>
<td>0.34</td>
</tr>
<tr>
<td>NPV</td>
<td>0.94</td>
<td>0.96</td>
</tr>
</tbody>
</table>

*STARR = Stanford medicine Research data Repository; MIMIC = Medical Information Mart for Intensive Care; ML = machine learning; LRM = logistic regression model; GBM = gradient boosting machine; SWIFT = Stability and Workload Index for Transfer; AUROC = area under the receiver operating curve; AUPRC = area under the precision-recall curve; PPV = positive predictive value; NPV = negative predictive value.

A) Figure 1. Logistic regression performance on internal (STARR) and external (MIMIC) validation. A) ROC curve for STARR (AUC = 0.81) and MIMIC (AUC = 0.56) test data. B) PR curve for STARR (AP = 0.41) and MIMIC (AP = 0.07) test data.

B) Figure 2. Gradient boosting performance on internal (STARR) and external (MIMIC) validation. A) ROC curve for STARR (AUC = 0.85) and MIMIC (AUC = 0.60) test data. B) PR curve for STARR (AP = 0.41) and MIMIC (AP = 0.08) test data.
Figure 3. SWIFT performance on internal (STARR) and external (MIMIC) validation. A) ROC curve for STARR (AUC = 0.67 at threshold score of 15) and MIMIC (AUC = 0.51 at threshold score of 15) test data. B) PR curve for STARR (AP = 0.17 at threshold score of 15) and MIMIC (AP = 0.06 at threshold score of 15) test data.

Discussion

High morbidity and mortality rates in the intensive care setting are exacerbated by unplanned readmissions. The lack of objective uniformity in ICU discharge decisions leads to undesirable care quality variability, highlighting a need for generalizable clinical decision support and process improvement. To date, the only validated tool for estimated ICU readmission risk is the SWIFT score, which utilized logistic regression to determine that location prior to ICU admission, total length of ICU stay, last PaO\textsubscript{2}/FiO\textsubscript{2} ratio, last Glasgow Coma Scale (GCS) score, and last PaCO\textsubscript{2} were most predictive of ICU readmission with an AUROC of 0.75\textsuperscript{13}. However, SWIFT application was not found to improve patient or resource-related outcomes on multi-center implementation\textsuperscript{10–12}. For 7,175 patients, Kastrup et al. found an AUROC of 0.58 for unplanned ICU readmissions within 7 days of ICU discharge using the SWIFT score\textsuperscript{10}. SWIFT application to STARR (AUROC 0.67) and MIMIC-IV (AUROC 0.51) data in this study further supports this finding. There is a need for an improved clinical decision support tool.

Most previously published machine learning models for ICU readmission risk were trained and validated on MIMIC-III data (2001-2012). Logistic regression, random forests, support vector machines, gradient boosting, and neural networks yielded AUROC values from 0.71 to 0.88 with relatively low sensitivity (0.28 to 0.65) for ICU readmission risk spanning 24 hours to 30 days after discharge\textsuperscript{11–18}. Two studies reported SWIFT performance on the same dataset, and both showed improved performance with machine learning models compared to SWIFT. A boosted decision tree ensemble from Desautels et al. had an AUROC of 0.71 compared to 0.61 with SWIFT\textsuperscript{12}. Rojas et al. also saw better performance with a gradient boosting model on internal (gradient boosting AUROC 0.76, SWIFT AUROC 0.65) and external (gradient boosting AUROC 0.71, SWIFT AUROC 0.58) validation\textsuperscript{11}. Similarly, our gradient boosting model performed better than SWIFT on internal (gradient boosting AUROC 0.85, SWIFT 0.67) and external (gradient boosting 0.60, SWIFT 0.51) validation.

On internal validation, gradient boosting was the highest performing model for AUROC, sensitivity, and NPV (see table 4). Compared to prior machine learning (ML) work, our classifiers achieved one of the higher AUROC values on internal validation at 0.85. McWilliams et al. also saw an AUROC of 0.88 on interval validation with a logistic regression model and 0.89 on external validation (MIMIC-III) with a random forest model\textsuperscript{17}. However, model training was performed on a dataset that included patient records from both internal and external validation institutions. Rojas et al. reported an AUROC of 0.76 on data from University of Chicago Medical Center (internal validation) and 0.71 on data from MIMIC-III (external validation) with a gradient boosting model\textsuperscript{11}. Our AUPRC (0.41), PPV (0.29), and F1-scores (0.43) for gradient boosting were low on internal validation but is similar to past studies in which F1-scores have ranged from 0.32 to 0.53 and AUPRC is around 0.33\textsuperscript{13,14,17,25,26}. Moreover, our model is on the higher end of reported baseline sensitivity (0.55-0.88) and specificity (0.70-0.80) ranges with 0.84 sensitivity and 0.78 specificity\textsuperscript{12,14,17,25,26}. Our gradient boosting model demonstrated better or similar performance compared to most existing ML models.
While we cannot pinpoint the exact factors that contributed to our performance outcomes compared to those reported in literature, we hypothesize that the initial chosen feature set was a large contribution. Several studies have exclusively focused on only lab measurements or select organ systems such as the respiratory, nervous, and hematologic systems, likely based on the most common conditions that land patients in the ICU and most common causes of ICU mortality\textsuperscript{12,14}. However, literature shows that models with more comprehensive information across multiple organ systems, with different modalities (e.g., lab measurements, imaging), and at different stages of the ICU stay (e.g., initial admission and diagnosis, treatment, monitoring) tend to perform better\textsuperscript{11,15}. We aimed to provide a comprehensive view of a patient’s ICU course with a feature set that measures end organ dysfunction and failure for every organ system and documents the patient journey from diagnosis to treatment to maintenance and monitoring across different modalities such as lab measurements, imaging, intravenous medications, and procedural interventions. To obtain a fuller picture, we also represented the patient’s worst and typical health status based on extrema and counts of measurements outside of the normal physiologic range, respectively (see table 1). This feature set, curated by clinical experience, likely contributed to our model performance on internal validation.

Most prior work show the potential of machine learning-based clinical decision support in critical care but perceived benefits from implementation are unclear without additional validation in a variety of clinical institutions and settings. However, few studies performed external validation. Moreover, several included multiple ICU stays per patient, and many did not distinguish planned from unplanned readmissions, which may decrease performance at external sites and thereby reduce generalizability. Our model sought to improve generalizability and external validity by including only the first eligible ICU stay per patient and performing external validation on MIMIC-IV data. In an effort to target unplanned readmissions, we excluded patients who underwent a surgical procedure within 7 days of ICU discharge.

On external validation, gradient boosting performed better than logistic regression and SWIFT in all metrics except for specificity. Gradient boosting is more robust to overfitting compared to logistic regression and has consistently been shown to perform well on a wide range of problems and fields, especially bioinformatics\textsuperscript{27}. While gradient boosting is effective at predicting 7-day ICU readmission for STARR, several performance metrics decreased with external validation. Most notably, F1-scores decreased from 0.43 to 0.14 and AUROC decreased from 0.85 to 0.60 despite consistently high accuracy of 0.90 and 0.93 on internal and external validation, respectively. While Rojas et al. saw more consistent model performance between internal and external validation with AUROC of 0.76 and 0.71, respectively, they did not report F1-scores and had a smaller gap in readmission prevalence between internal and external validation datasets. As a drop in prevalence decreases PPV and F1-scores, we expected to see some decrease from internal to external validation since readmission prevalence was halved (10% STARR readmission to 5.8% MIMIC readmission prevalence). Though expected, this does not fully explain the low MIMIC performance metrics.

Additionally, the gap in F1-score may be attributed to other factors such as regional variations in clinical care, patient demographics, and health profiles. That the SWIFT model also performed more poorly on MIMIC-IV than Stanford suggests that differences ranging from clinical practice or patient populations to data documentation and representation may affect predictive models independent of intrinsic model properties. Hospital-level factors such as transfer timing and the number of subsequent transitions in primary team care (e.g., transfer to night team then primary receiving team) as well as the acuity level of the receiving ward (e.g., step down with more intensive monitoring and services than general hospital ward versus general ward) can influence readmission risk. More frequent care transitions can cause preventable adverse events due to information loss in transition of care communications. These factors illustrate how variability in institutional policies and staffing practices can greatly contribute to differences between cohorts and may not be fully captured in the data. Overall, our study and others suggest that while machine learning methods are an improvement over SWIFT and have potential to standardize ICU transfer decisions, there are still obstacles to overcome in generalizing any single model across multiple hospitals and clinical care settings.

Building a model for general use requires institutions and hospitals to commit to maintaining a comprehensive up to date electronic health record (EHR) database, ideally adhering to common data standards, and sharing this data for the academic community and others to build clinical decision support tools for improved care. Stanford and BIDMC are both quaternary care hospitals with robust academic critical care departments but differences in data storage and representation remained. For example, all records in MIMIC-IV were of ICU patients whereas ICU patients in STARR were distinguished based on admission orders, primary care team designations, and acuity levels. This difference in digital phenotype likely contributed to the decrease in select performance metrics on external validation. Additionally, text from notes and imaging reports were not included in the feature set, due to difficulties in consistently representing unstructured elements between STARR and MIMIC-IV. However, MIMIC has begun a transition to the Observational
Machine Learning Methods Can Improve ICU Discharge Triage: A Case Study Using MIMIC Data

Medical Outcomes Partnership (OMOP) Common Data Model (CDM) and STARR data is already accessible in OMOP format. Full MIMIC mapping to the OMOP CDM will improve interoperability and feature extraction in future work. In the meantime, it may ultimately be more prudent to take an approach of custom training each model for local use rather than aiming to create one model with variable accuracy depending on location of use. Last, we note that many machine learning models such as gradient boosting yield better performance at the expense of reduced interpretability. Our work and most current literature rely on these methods, and this tradeoff should be a design consideration at an implementation stage.

Additional implementation considerations should account for desired outcomes. While high specificity is desirable with regards to resource allocation, low sensitivity reduces clinical utility by reducing the number of patients who could be positively influenced by early identification. The sensitivity, specificity, PPV, NPV, and F1-score metrics reported in table 4 are based on an ROC threshold optimized by Youden’s J statistic. Practically however, clinicians can apply any threshold that is within their risk tolerance and choose to optimize whichever metric is deemed more important for the patient or the critical care unit in terms of bed allocation. NPV remained high on STARR and MIMIC datasets, regardless of specificity. Clinically, optimizing for NPV over specificity is preferred to see true non-readmissions when predicted as such by the ML model. Our gradient boosting model is one of the higher performing single-institution validated classifiers reported in literature and one of very few with external validation. Regardless, this study has shown that ML models, particularly gradient boosting, are an improvement upon the existing standard.

Conclusion
In conclusion, machine learning methods can improve current standards of ICU discharge triage based on 7-day ICU readmission risk prediction. The benefit of an automated ML model for clinical decision support lies in the objectivity, consistency, and data consolidation it can bring to clinical care though current application in ICU readmission risk prediction may require use of multiple locally trained models over one general model. Machine learning models have the potential to assist clinicians by reducing decision fatigue, providing consistent objective readmission risk probabilities, and standardizing the decision-making process.

Acknowledgements
VH was supported by the Stanford Intermountain Fellowship SPO137549. JHC was supported by the NIH/National Library of Medicine via Award R56LM013365; Gordon and Betty Moore Foundation through Grant GBMF8040; National Science Foundation SPO181514; Google, Inc SPO13604; Stanford Aging and Ethnogeriatrics (SAGE) Research Center under NIH/NIA grant P30AG059307; Stanford Clinical Excellence Research Center (CERC); and Stanford Departments of Medicine and Pathology.

This research used data or services provided by STARR, STAnford medicine Research data Repository,” a clinical data warehouse containing live Epic data from Stanford Health Care (SHC), the University Healthcare Alliance (UHA) and Packard Children’s Health Alliance (PCHA) clinics and other auxiliary data from Hospital applications such as radiology PACS. The STARR platform is developed and operated by the Stanford Medicine Research IT team and is made possible by the Stanford School of Medicine Research Office.

Author Contributions
KS contributed to the data collection and analysis system development, data collection, analyses, literature review, and manuscript writing. VH contributed to the initial data collection system development, literature review, and manuscript writing. JJS contributed to data analysis system development and manuscript writing. KB contributed to data analysis system development and manuscript writing. JHC was the primary advisor of this work.

References

454
An Informatics Analysis to Identify Sex Disparities and Healthcare Needs for Autism across the United States

Nate Tyler Stockham1, Kelley M. Paskov1, Kevin Tabatabaei1,2, Soren Sutaria1, Bennett Liu1, Jack Kent1, Dennis P. Wall1

1Stanford University, Stanford, California; 2McMaster University, Hamilton, Canada; ∗These authors contributed equally

Abstract

Autism is among the most common neurodevelopmental conditions. Timely diagnosis and access to therapeutic resources are essential for positive prognoses, yet long queues and unevenly dispersed resources leave many untreated. Without granular estimates of autism prevalence by geographic area, it is difficult to identify unmet needs and mechanisms to address them. Mining a dataset of 53M children using meaningful geographic regions, we computed autism prevalence across the country. We then performed comparative analysis against 50,000 resources to identify the type and extent of gaps in access to autism services. We find a steady increase in autism diagnoses from K-5, supporting delayed diagnosis of autism, and consistent under-diagnosis of females. We find a significant inverse relationship between prevalence and availability of resources (p < 0.001). While more work is needed to characterize additional trends including racial and ethnicity-based disparities, the identification of resource gaps can direct and prioritize new innovations.

Introduction

Autism spectrum disorder (ASD) is one of the most common developmental pediatric conditions, and is characterized by difficulties in social interaction and communication and restricted or repetitive interests and behaviors. For children with autism, early diagnosis and access to treatment results in improved outcomes1–3. However, the median age of diagnosis in the U.S. is 51 months4 with wait times for diagnostic services or ongoing care visits ranging from 3-14 months5. Over the past ten years, the prevalence of autism has grown from 1 in 110 to 1 in 544. This growth is likely to further stretch an already overburdened system of autism support services.

Current estimates of autism prevalence are produced by the Autism and Developmental Disabilities Monitoring (ADDM) Network which closely tracks the prevalence of autism among 8 year-old-children across 11 sentinel sites4. Autism prevalence varies widely from site to site from 1.31% (Colorado) to 3.14% (New Jersey), suggesting that autism prevalence is also likely to vary widely across the country. However, estimates of autism prevalence are only useful if they successfully guide the investment of healthcare resources. While ADDM does provide an estimate of autism prevalence for a few states, the actual availability and type of autism healthcare resources relative to need for a given geographic area is understudied. To estimate the relative disparities between autism prevalence and autism healthcare resources between different geographic areas and demographic groups we combined two national datasets; the Civil Rights Data Collection (CRDC) and GapMap.

The Civil Rights Data Collection is collected bi-annually by the U.S. Department of Education from every school in the U.S. and encompasses 53 million students, including the total student enrollment and the number of students in each school receiving special education services for autism under the Individuals with Disabilities Education Act (IDEA). Incorporating data from educational sources into autism prevalence estimates has been found to reduce racial/ethnic differences in prevalence rates6 and to mitigate healthcare access disparities7. The IDEA criteria used by schools to identify autism is quite effective, with over 90% of autistic students identified by schools also meeting the criteria for an autism diagnosis via the Autism Diagnostic Observation Schedule (ADOS), a diagnostic standard8. For sites where educational records were available, ADDM found that between 85.1% and 93.6% of children with autism are receiving special education services4. This makes the number of students receiving autism services in school an excellent proxy for the total number of autistic children in the school. The national comprehensiveness of the CRDC and its geographic detail make it an excellent data set for estimating the relative autism prevalence between geographic areas.

In order to understand how diagnostic and therapeutic resources for autism vary from state to state, we compare autism prevalence rates estimated from CRDC to the autism resources available in GapMap. GapMap is an autism
Figure 1: GapMap is an autism resource database containing the locations of diagnostic and therapeutic autism resources across the U.S.

resource database and interactive online map containing 51,000 geolocated autism resources. GapMap was designed to simultaneously track autism epidemiology and help families find autism services, and has been used to study ASD resource availability in the U.S., using national-scale prevalence estimates in order to identify geographic areas with insufficient resources. The average autistic individual is 20 miles (32 kilometers) away from the nearest autism diagnostic center and 83.86% of all US counties contain no autism diagnostic resources at all. By using estimates of autism prevalence drawn from schools, rather than national averages, we can more accurately analyze and identify geographical gaps in autism services.

Using CRDC data, we estimate the prevalence of male and female autism across the U.S. We show that autism prevalence varies by grade, and that a larger percentage of male autism is diagnosed during elementary school ages than female autism. We replicate previous findings that the number of children receiving special education services are highly correlated across IDEA categories. Finally, using the GapMap autism resource database, we show that the availability of autism resources per autistic student varies dramatically by state, with states with higher autism prevalence providing fewer resources per child. Our work demonstrates that CRDC school-based data can help understand the geographical distribution of autism as well as ASD resource gaps in the U.S.

Methods
Estimating autism prevalence

We use the CRDC to estimate autism prevalence at the sub-state level by aggregating school-level data into larger geographic units of Core-Based Statistical Areas (CBSA) as defined by U.S. Census Bureau. A CBSA is a geographic area chosen to delineate metropolitan or micropolitan boundaries associated commute regions, providing a natural way to group student enrollment across the US. We estimate overall prevalence rates using median prevalence across all CBSAs and empirically estimate 95% confidence intervals for prevalence across all CBSAs.

IDEA diagnosis correlated structure by multivariate ordinary least squares

We estimate the sex-specific correlation structure between autism and other IDEA categories using multivariate ordinary least squares; for each CBSA the total number of autistic students was regressed against total student enrollment
and the totals for the other twelve IDEA categories of deaf-blindness, developmental delay, emotional disturbance, hearing impairment, intellectual disability, multiple disabilities, orthopedic impairment, specific learning disability, speech or language impairment, traumatic brain injury, visual impairment, and other health impairment. However, students can only be assigned one of these IDEA categories; this restriction has implications for correlation structure as discussed in Results.

Estimating per-grade autism prevalence

The CRDC dataset provides the number of autistic students per school, but not per grade. Grade-level estimates of prevalence are useful in order to estimate age at first diagnosis for the population. Using per-grade total enrollment estimates from CCD, we estimate autism prevalence per grade for each sex using regularized regression. Let $X^e$ be a matrix where each entry $x^e_{ij}$ is the number of male students from race/ethnicity category $e$ in grade $j$ at school $i$. Our goal is to estimate $\beta^e$, a vector where entry $\beta^e_j$ is the prevalence of autism in males from race/ethnicity category $e$ for grade $j$. We will do this using $y^e$, a vector where each entry $y^e_i$ is the number of autistic male students from race/ethnicity category $e$ at school $i$. We model $y^e$ as the sum of binomials with different probabilities, since each grade will have a different autism prevalence. Le Cam’s theorem states that the sum of binomials with different probabilities approximately follows a poisson distribution with parameter $\lambda = X^e \beta^e$ in our case. We use maximum likelihood estimation to estimate $\beta^e$

$$\maximize_{\beta^e \forall e \in E} \sum_{e \in E} y^e \log(X^e \beta^e) - 1 X^e \beta^e - \lambda \| \beta^e - \beta \|_2 - \lambda \| \beta_1 - \beta_{-1} \|_2$$

subject to $0 \leq \beta^e \leq 1 \forall e \in E$

The first two terms in the sum represent the log likelihood of our model, where $1$ represents a vector of ones. The third term is a regularizer that encourages the autism prevalence estimates for each race/ethnicity category to be similar. The fourth term is another regularizer that encourages autism prevalence between consecutive grade levels to be similar. The regularization parameter $\lambda$ was set to $0.1n$ where $n$ is the number of schools included in the analysis. Finally, we constrain our autism prevalence estimates to be between 0 and 1. We fit our model in python using cvxpy for male and female autism prevalence separately.

We note that our regularization scheme encourages race/ethnicity categories to have similar prevalence estimates. This makes it difficult to interpret prevalence differences across race/ethnicity categories.

Next, we use our grade-level autism prevalence estimates to estimate the age that autistic children begin receiving services in school. We do this by assuming that the increase in our autism prevalence estimates between grades preK-5th are primarily due to the delayed diagnosis of autism throughout these age ranges. This is supported by findings from ADMM that the median age of autism diagnosis is 51 months. To estimate when autistic children begin receiving services in school, we calculate the difference in prevalence between neighboring grade levels, and divide this difference by the autism prevalence estimate for 5th grade (the grade with peak autism prevalence). This provides an estimate of the fraction of autistic children that begin receiving services in any given grade between preK-5th.

Generating K-12 ASD prevalence and resources maps and histograms

The state-level estimates of autism resources per thousand autistic student in Figure 6 was generated from GapMap and the CRDC. The data includes 49 states (excludes Iowa), Washington D.C. (denoted DC), and Puerto Rico (denoted PR), giving a total of 51 US state/territories.

Because states with higher populations contain more resources than less populous states, and diagnosis rates differ from place to place, it is necessary to normalize for ASD population to determine the relative abundance of resources per state. Fig. 4, Resources per 1000 ASD Diagnosed Per State, was produced by dividing the total number of resources in each state by the total number of ASD diagnosed students in the CRDC dataset, then multiplying by 1000. To determine the prevalence of ASD K-12 students per state (Fig. 4, Prevalence of ASD K-12 Students per
Results

Autism prevalence

Using CBSAs to aggregate school-level data provides consistent autism prevalence estimates across the U.S. as shown in Figure 2 yielding a prevalence estimate of autism among males to be 1.62% [0.86%, 2.58%], which is slightly lower than the current ADMM estimate of 2.94%. We estimate the female prevalence of autism to be 0.33% [0.14%, 0.59%], which is also lower than the current ADMM estimate of 0.69%. Finally, we estimate the male:female sex ratio for autism to be 5.3:1 [3.4:1, 9.6:1] which is in range of the current ADDM estimate of 4.3:1.

Our estimates of autism prevalence are lower than the most recent estimates from ADDM. It is possible that the 11 sentinel sites used by ADDM are not representative of the U.S. as a whole. However, it is more likely that children with autism are under-counted in the special education totals. Schools are required to choose a single IDEA category for each student, so autistic students that also qualify for other services may be included in a different IDEA category. In fact, ADDM found that for sentinel sites where special education information was fully available, an average of 32% of autistic students were reported under a different IDEA category.

Relationships between autism and other IDEA categories

To support our hypothesis that the number of autistic students are under-reported in the CDRC due to autistic students being reported in other IDEA categories, we examine the correlation between ASD and other IDEA categories in Figure 3. Three IDEA categories exhibit significant negative correlation with ASD counts: developmental delay, speech or language impairment, and multiple disabilities. This finding is supported by ADMM which found that among sites with access to school records for all students, children with autism are often included in other IDEA categories including developmental delay (0.7-23.2%), speech or language impairment (3.4-6.2%) and multiple disabilities (0.0-6.2%). Notably there is no significant correlation between the number of autistic students and the number of students with deaf-blindness, traumatic brain injury, hearing impairment, or visual impairment: this agrees with prior knowledge of the etiological distinction between these disabilities and autism.

Autism prevalence by grade

Using grade-level enrollment, we are able to estimate grade-level autism prevalence as shown in figure 4. Autism prevalence increases from preschool through 5th grade as students are identified as qualifying for autism services.
Figure 3: There are significant correlations between the number of ASD students in a CBSA and the number of students in other IDEA categories. Negative coefficients suggest the possibility that some students with ASD are being included in the other IDEA categories.

Figure 4: Autism prevalence varies by grade. Autism prevalence increases from preschool through 5th grade, then declines slowly. This pattern is consistent for both male and female students, and across every race/ethnicity category. The male-female ratio climbs from preschool through 11th grade ranging from 4x to 6x. Dotted black lines show most recent ADMM estimates. We exclude 12th grade due to data artifacts.
Female autistic students begin receiving services at younger ages than male autistic students. Female children are more likely to be identified as needing autism services before school begins than male students.

Interestingly, autism prevalence declines slightly from 5th-11th grade. There are several explanations for this decline. Changing diagnostic criteria for ASD between DSM-IV and DSM-V may have affected primary school students differently than secondary school students. Another possibility is that for some students, school-based autism services are only necessary in the younger grades.

Using estimates for autism prevalence in each grade, and assuming that autism prevalence is constant across each grade level, we estimate the fraction of autistic students that begin receiving autism services in each grade, as shown in Figure 5. Approximately 60% of female autistic students are identified as requiring autism services before starting kindergarten as compared to 48% of male autistic students. This is supported by ADMM findings that female children with autism are more likely to be evaluated before 3 years old than male children\(^4\). By first grade, 97% of female autistic students will be identified, as compared to only 87% of male autistic students.

There are several explanations for this finding. Female autism may be more severe than male autism, resulting in younger age of diagnosis. Alternatively, schools may be more effective at identifying male autism than female autism, resulting in missed autism diagnoses for females during elementary-school.

Comparing autism prevalence to resources across the United States using GapMap

Next, we compared autism prevalence to resources using GapMap, as shown in figure 6. There is a significant inverse relationship \(p < 0.001\) between autism prevalence among K-12 students and available autism healthcare resources.
There were a median of 24 ABA resources, 36 therapy resources and 27 diagnostic resources per thousand autistic students across the 50 states. South Dakota and North Dakota have the largest number of resources per autistic student across all categories while Puerto Rico, Nevada, and California consistently have the fewest across all categories.

available as indexed by GapMap; a percentage point increase in autism prevalence translates to approximately 90 fewer resources per thousand autistic students. The notable outliers are Montana, North Dakota, and South Dakota, with autism prevalence of 0.589%, 1.003%, 0.903% and resources per thousand autistic students of 260, 313, 243, respectively. In contrast, New York state has similar autism prevalence of 0.685% but roughly half (163) the autism healthcare resources per autistic student.

It is interesting to note that New York has similar characteristics to states in the Northeast or West – California, Massachusetts, or Washington – since their overall population is dominated by high density population areas. Thus, it would make sense if they had similar ASD provider infrastructures and followed similar data trends of high ASD prevalence but low number of resources available. However, this was not the case. In fact, New York was the opposite because it has low ASD prevalence (0.685%) and a relatively high number of resources available (163.1 resources per 1000 diagnosed). In future research, it would be imperative to explore why New York differs from similarly populated states. This would enable us to better understand how states with varying population density can effectively provide accessible ASD resources across their state.

Classification of resources into service categories

To analyze resource frequency by state in context, we need to cross reference the frequency of types of resources (7) with prevalence data (6) and population growth trends. Resource frequency shortcomings in Western states are likely due to recent population growth, which entailed an increase in total ASD K-12 diagnosed cases. Western states have undergone significant population growth from 2008-2018. All states have grown significantly more than the 50-state median of 0.63% growth (California 0.78%, Oregon 1.07%, Arizona 1.34%, Nevada 1.35%, Idaho 1.35%, Utah 1.73%)\(^{12}\). With this quick population growth, many of these states are now under-equipped in resources. Nevada has a 1.31% diagnosed over total enrolled which is above the 51 state/territory average of 1.11%. It also has the lowest resources per 1000 diagnosed across ABA, therapy, and diagnostic resources among all US states (7). Thus, with a rapidly growing population (2.09% from 2017-2018, highest in the US) and high childhood ASD prevalence, Nevada has shortcomings in autism resources. It is recommended that more resources of all categories be opened across the state.
Similarly, Utah’s resource shortcomings and low prevalence rate are due to the recent population spike. Utah has the 4th lowest resources per 1000 for ABA and diagnostic resources and the 7th lowest resources per 1000 for therapy resources among the 51 US state/territories (7). While Utah’s childhood ASD prevalence is low at 8.95%, it had the largest growing population in the US from 2008-2018 (12). Utah’s childhood ASD undiagnosed cases in the coming years has likely grown from the population spike. Therefore, it is recommended that Utah opens more resources, especially diagnostic resources, across the state. California and Arizona also consistently fall in the lower bins across all categories of resources (7). Despite lower population growth from 2008 to 2018 than NV and UT, these states contain a wide variety of population density by geographical location. California and Arizona should devote attention to increasing total resources available to better serve their growing population.

Montana (74.37), New York (59.19), Louisiana (29.08), and Colorado (32.22) all have high diagnostic resource frequency per 1000 diagnosed, and low K-12 ASD prevalence rates (MT 0.589%, NY 0.685%, LA 0.721%, CO 0.733%). North Dakota and South Dakota are outliers above upper thresholds for resources per 1000 diagnosed across all categories (7). Even so, both states have mild prevalence rates of ASD K-12 cases (1.00% and 0.903% for ND and SD respectively) that are slightly lower than the nationwide average of 1.11%. Both states have low population density, so it may be necessary for these states to have more resources per 1000 autistic students in order for resources to be geographically accessible to every student.

Discussion

The CRDC is a valuable resource for estimating autism prevalence in the U.S. It encompasses 53M students and provides a fine-grained geographical distribution of students receiving autism services. However, autism prevalence rates estimated from this data are lower than other studies. Our estimates are likely underestimating the prevalence of autism in the U.S. This discrepancy is due in part to the CRDC requirement that students be included in only one IDEA category. Our analysis shows that IDEA categories are significantly correlated across CBSAs, suggesting that some students with autism are recorded in other IDEA categories including developmental delay, emotional disturbance, intellectual disability, orthopedic impairment, or specific learning disability. Another explanation for lower prevalence rates in school data is parents choosing not to report their child’s autism diagnosis to the school. For ADMM sites where educational records were available for all students, 6.4-14.9% of autistic students did not have special education records. More work is needed to understand the concordance of medical records and school records for autism. Finally, autism may be under-diagnosed in some regions of the U.S., which could result in lower overall prevalence estimates.

Despite this limitation, we were able to estimate changes in autism prevalence by grade level and use these estimates to understand the age at which autistic children begin receiving services. Male autistic students are identified throughout the primary grades, while female autistic students are most often identified before entering school. This suggests that female autism may be easier to identify at younger ages as compared to male autism, or alternatively that female autism is under-diagnosed in grades K-5.

Since states are so large and include a diverse array of urban/suburban/rural areas, in the future, we plan a more granular analysis at the CBSA, county, or school level. Doing so would enable us to recognize gaps in resources within states. Furthermore, by using GapMap’s resource data of precise location (longitude and latitude), we could calculate the average driving time to resources from the closest educational provider. Availability is likely constrained by excessive wait times and/or difficulty in reaching these resources (driving 20 miles in a rural area would take far less time than driving 20 miles in Los Angeles or New York). Along with other factors like traffic congestion and public transit availability, it may be possible to define an accessibility index that would give us a better grasp on where resources should be opened such that they can consistently serve families before or after school.

Another source of further analysis is GapMap’s future capability of recording capacity and wait times for all resources in the database. The analysis executed using GapMap counts the number of resources without any ability to gauge capacity of that resource. This could explain why northern mid-west states seem saturated in resources as compared to coastal states. If a therapy resource in California is a two story clinical building with 10+ therapists while a therapy resource in North Dakota is an office with one therapist, GapMap is currently unable to account for the difference in throughput for children with autism. With capacity data, we could better evaluate where resource gaps are more or less prevalent. Moreover, GapMap does not assess the current capability or wait time of a resource. In certain urban areas,
diagnostic service centers may not have availability for many months, rendering them unavailing for families who are seeking a timely diagnosis for their child. Timely diagnosis is vital to early intervention which in turn significantly ameliorates symptoms of ASD. Therefore, with the knowledge that the average wait times for diagnostic resources in a location is high, we would be able to better pinpoint where alternative diagnosis capabilities are needed to reduce the congestion of resource demand in an area.

Finally, in the northern parts of the United States (namely North Dakota, Montana, and South Dakota), there was low prevalence of ASD, but a relatively high number of resources available per autistic student. This may indicate that these Northern states have better autism resource coverage than other states. However, these particular states are geographically large with areas of low population density. It may be that many resources are required in order to be geographically accessible to students. Further research looking into precise geographical locations of clinics, clinic size and capacity, and average wait times at clinics would be needed to better understand the landscape of autism resources in relation to where autism rates are lower. Our map analysis suggests that there may be an increasing nearly nationwide shortage of resources for the growing ASD K-12 population. More research is needed to better understand where and why there are scarcities of resources in relation to ASD prevalence.

Our work shows that CRDC data is a valuable resource for analyzing the geographical distribution of autism. In concert with autism resource databases like GapMap, it can help us understand where more resources are needed to support autistic students in the U.S. CRDC is an incredibly comprehensive resource, containing data on 53M students. We encourage other researchers to consider working with this resource.

Acknowledgements

The work was supported in part by funds to DPW from the National Institutes of Health (1R01EB025025-01, 1R21HD091500-01, 1R01LM013083), the National Science Foundation (Award 2014232), The Hartwell Foundation, Bill and Melinda Gates Foundation, Coulter Foundation, Lucile Packard Foundation, the Weston Havens Foundation, and program grants from Stanford’s Human Centered Artificial Intelligence Program, Precision Health and Integrated Diagnostics Center (PHIND), Beckman Center, Bio-X Center, Predictive and Diagnostics Accelerator (SPADA) Spectrum, Spark Program in Translational Research, MediaX, and from the Wu Tsai Neurosciences Institute’s Neuroscience:Translate Program.

References


DaT3M: A Data Tracker for Multi-faceted Management of Multi-site Clinical Research Data Submission, Curation, Master Inventorying, and Sharing

Shiqiang Tao, PhD1,3, Licong Cui, PhD2,3, Wei-Chun Chou, MS1,3, Samden Lhatoo, MD1,3, Guo-Qiang Zhang, PhD1,2,3

1 Department of Neurology, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX, 77030
2 School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, 77030
3 Texas Institute for Restorative Neurotechnologies, The University of Texas Health Science Center at Houston, Houston, TX, 77030

Abstract

Managing research data is an important and challenging aspect of clinical studies, especially for multi-site collaboratives. To address this challenge, we designed, developed and deployed a multi-faceted, multi-level interactive data tracker (DaT3M) for multi-site clinical research data submission, curation, master inventorying, and sharing. Components of DaT3M include data overview, data portal, data status panel, data query engine, and data downloader. DaT3M managed clinical research data for the Center for SUDEP Research (CSR). The CSR instance of DaT3M includes 2,743 subjects from seven data contributing institutions, 7 data modalities and 10,678 data components: 3,398 Epilepsy Monitoring Unit reports, 3,440 electroencephalography recordings, 629 MRI imaging datasets, 177 bio-chemistry datasets, 722 DNA datasets, 2,289 follow-up forms, and 30 SUDEP forms. Preliminary, structured, one-on-one usability evaluations were performed with 7 researchers from four institutions. System Usability Score reached 85.3, showing that DaT3M has achieved high levels of user satisfaction based on our pilot evaluation.

1 Introduction

Managing clinical research data following FAIR data principles1,2 is an important and demanding task for clinical studies. The challenges involved become pronounced for large-scale multi-site collaboratives. While multi-site collaboratives bring people and resources together to provide better opportunities to solve complex research problems3–5, they also bring new challenges for data management. Common challenges include the lack of data consistency and common data format, semantic interoperability, continuous data integration and availability, and those highlighted by FAIR6. Therefore, multi-site collaboratives require a modernized informatics infrastructure to manage clinical data that supports real-time data access by clinical investigators while achieving essential elements of FAIR.

In this paper, we present the design, development, and user evaluation of a Data Tracker for Multi-faceted Management of Multi-site clinical research data submission, curation, master inventorying, and sharing called DaT3M. To elevate the FAIRness of research data, we provide five main features in DaT3M: (1) a data overview interface; (2) individual center data portal; (3) a data status panel for real-time status checking of multi-modal data; (4) a data query engine to build patient cohort; and (5) a data slice downloader. The data overview interface provides a quick summary of the entire data collection in the central repository. Information such as the total number of complete datasets, the number of datasets from each individual institution, available data types in each dataset, and the total number of datasets of each data type are presented in the data overview interface. This is the most frequently used function to coordinate data integration from all institutions.

The individual center data portal provides a dedicated work-space for each participating institution to manage their own contributed data. The data status panel is provided in institutional portal using data visualization techniques to track the statuses of all data types for each subject. For large research data repositories, a query engine is necessary to build patient cohorts to identify the subset of patients with specific characteristics. It is critical for researchers to be able to find subgroups of study subjects for specific hypotheses7,8. In DaT3M, we reuse MEDCIS9 to build patient cohorts. The data slice downloader delivers user-specified data subsets from the central repository for secondary data analysis.

Many tools are available for research data management including Microsoft Excel, REDCap10, MACRO11, Oracle Clinical12, OpenClinica13 and OpenCDMS14. Although these software tools are widely used for research data management, we do not find a single tool that can address all the data FAIRness challenges faced by large-scale multi-site collaboratives. Therefore, we develop DaT3M as an integrated web application to provide customized functions to support data submission, curation, master inventorying, and sharing of large-scale multi-site collaboratives.
DaT3M has been piloted to support the data management for the Center for SUDEP Research (CSR), a National Institute for Neurological Disorders and Stroke (NINDS) funded Center Without Walls for Collaborative Research in the Epilepsies. DaT3M has integrated 10,678 data components of 7 different modalities for 2,743 patients from CSR’s seven data contributing institutions. The features of DaT3M make it an ideal tool for supporting large-scale multi-site collaboratives.

2 Background

2.1 Center for SUDEP Research

CSR is a multi-site cross-disciplinary collaboration composed of researchers from 16 institutions across the United States and Europe to understand Sudden Unexpected Death in Epilepsy Patients (SUDEP). The precise mechanism of SUDEP is still unknown. And due to the low incidence rates of SUDEP (0.2 per 1,000 persons/year in children and 1.0 per 1,000 persons/year in adults), multi-site studies are the key to collect sufficient datasets for statistical analysis. This investment by NINDS over nearly five years promises to catalyze research on SUDEP and dramatically enhance our understanding of this devastating phenomenon. The participating institutions of CSR includes Baylor College of Medicine, University Hospitals Cleveland Medical Center (UH), Lurie Children’s Hospital of Chicago, Columbia University, Harvard University, NYU School of Medicine (NYU), Northwestern University (NW), Texas Children’s Hospital, Thomas Jefferson University (TJU), UCLA, UCSF, University College London (UCL), University of Iowa (UIowa), University of Kentucky (UKY), University of Michigan, and The University of Texas Health Science Center at Houston. CSR collects seven types of clinical data for analysis including patient reports from epilepsy monitoring units (EMUs), electroencephalography (EEG) signal data, imaging data, bio-chemistry data, DNA data, follow-up forms and SUDEP forms. UH, NYU, NW, UCLA, UCL, TJU, and UIowa are the seven CSR institutions that contribute clinical data to the central data repository. Figure 1 shows the clinical data flow from distributed institutions into the CSR central data repository using SFTP. Multi-modality clinical data are captured and uploaded from individual institutions and processed at the central data repository. Data processing tasks such as data integration, data curation, and data conversion happen at the same time handled by different working groups of CSR.

![Figure 1. Clinical research data flow in CSR.](image)

2.2 FAIR, Rigor, and Reproducibility for multi-site Clinical Studies

Two cornerstones of science advancement, rigor in designing and performing scientific research and the ability to reproduce research findings, can be facilitated by FAIR data principles\(^1\). FAIR requires research data to be Findable, Accessible, Interoperable, and Reusable. Research data that meet FAIR principles have better machine-actionability and readiness for secondary analysis. Reproducibility of research findings is especially important in view of the inherent biological variations and expected technical and methodological varieties. Rapid sharing of verified raw data and associated data analytics has played a critical role for data reproducibility\(^17,18\). It is important to establish stringent data quality control and quality assurance processes to ensure that the data generated will be broadly referenced and
used by the research community. However, data management tools facilitating and enabling FAIR data principles are lacking in general, especially for complex data generated from multi-site collaboratives.

2.3 Gaps and Challenges

Gaps and challenges exist in collecting, managing, and sharing of research data in multi-site settings. One is that the traditional way of research data management heavily relies on general data management software such as spreadsheets. Such data management software is not capable enough to handle complex data management tasks such as facilitating data submission and sharing, and tracking data flow, data provenance, and data use statistics. For data sharing, the traditional way requires manual effort to build patient cohorts, prepare datasets, and distribute them to researchers when they request. The whole process is labor-intensive and error-prone. To overcome such challenges, we create an interactive research data management platform called DaT3M to automate complex tasks such as tracking data flow and provenance, building patient cohorts, and preparing datasets for sharing.

3 Methods

3.1 System Architecture

Figure 2 illustrates the system architecture of DaT3M. With MySQL database and Ruby on Rails application as the foundational backbone support, users can interact with DaT3M’s different function modules using web browsers. After a target patient cohort is built, users can proceed to download the data using our data downloader, which is a command line tool.

DaT3M is implemented using Ruby on Rails, an agile web application development framework with MySQL as the backend database. Its iterative agile development features allow us to rapidly respond to new requirements for data management tasks. We design our function modules following Web Interface-driven Development (WIDD), which is an effective software development method for clinical applications by involving domain experts in the process of interface design.

3.2 Data Modeling and Visualization

Three core data models are designed in DaT3M: Patient, Data Type, and Data Status. As depicted in Figure 3, one patient can have many data types while each data type has one data status (i.e., available or not). In the current instance of DaT3M for CSR, there are seven data types: EMU reports (short name as P), EEG signal data (E), MRI imaging data (M), biochemistry data (B), DNA data (D), follow-up forms (F), and SUDEP forms (S). Different types of data are linked by the patient’s unique CSR study ID.

To provide an intuitive and concise view of patient data and their statuses, we use a colored squared box to represent each data type as well as its status. The short name of the data type is enclosed in the squared box. We use green color to indicate that the data status is available and red to convey that the data status is missing or not available. For instance, given a specific patient, a squared box with letter “P” in green indicates that the patient’s EMU reports are available. Figure 4 shows a panel of seven colored square boxes, representing the overall status of one patient dataset in CSR. The data types are evenly placed horizontally. The color of each data type indicates the availability status of the corresponding patient data. Therefore, the available data for the patient shown in Figure 4 include EMU reports,
 EEG signal data, MRI imaging data, follow-up forms, and DNA data, while biochemistry data and SUDEP form are not available.

Figure 4. Patient data statuses of seven data types: P - EMU reports, E - EEG signal data, M - MRI imaging data, B - biochemistry data, D - DNA data, F - follow-up forms, S - SUDEP forms.

3.3 Data Status Tracking

To address the clinical research data management needs of cohort identification and data curation assistance, we created an efficient and powerful mechanism supporting querying patients by their data statuses. DaT3M provides a query widget (see Figure 5) with the similar appearance as the patient status panel consisting of 7 data status boxes. Importantly, the boxes in the query widget are interactive buttons that users can click to change colors to green, red, or white. White color in this query widget means the underlying data type is not used in the current query. In the query widget, the default status is optional with status box button in white; a mouse click will change the white color to green; a further mouse click will change the green color to red; and at last a click will change the red color back to white. That is, the order of color changing forms a loop: White $\rightarrow$ Green $\rightarrow$ Red $\rightarrow$ White.

Figure 5. Filter or find patients with P (EMU reports) and E (EEG signal data) available.

Therefore, constructing a query in DaT3M is straightforward. Users can simply click on the status buttons in the query widget to configure them in desired colors. For example, one query task in CSR is to find all patients with both EMU reports and EEG data available in the central data repository. We can efficiently construct such a query as displayed in Figure 5 with two clicks by setting buttons “P” and “E” as green (i.e., available), leaving all the other buttons as white (i.e., optional). The interactive query widget is powerful since it supports various queries with respect to data statuses – a total of $2^{7} = 128$ different combinations, where each combination is an independent query.

3.4 Robust Data Slice Downloader

DaT3M integrates with MEDCIS\textsuperscript{9} query engine which employs concepts from Epilepsy and Seizure Ontology (EpSO)\textsuperscript{27} to build patient cohorts. Once a patient cohort is created, DaT3M allows end users to download the cohort data using a scalable download framework as shown in Figure 6. A patient cohort often contains thousands of data files with size in terabyte level. It is not practical to use the typical browser-based individual file downloading method to retrieve a

469
patient cohort. To support scalable and smooth data download, we create a ruby gem called CSR Data Inventory which is a software package supporting batch-downloading of CSR patient cohort. This gem supports and has been tested on multiple operating systems (Windows, Linux, and OS X).

There are five steps involved in the downloading workflow (see Figure 6). First, the user gets a data download token from the built patient cohort. Then the user can run DaT3M data downloader in the local machine and feed in the data download token. In step 3, a file list will be downloaded containing the catalog of all files to be downloaded. Data downloader then traverses the file list and sends request to retrieve data files using the same data download token. In the last step, files are downloaded and stored in the user’s local device. The whole workflow is highly automated. The user only needs to start the program and feed in the data download token. It is called data slice downloader because the user has a choice to select what data type to download during the step of data download token generation. Typically, researchers only need specific types of data for their study, such as EEG data for signal processing and analysis or MRI data for imaging-related investigation. Slicing data can effectively reduce the workload of data download and researchers’ local data management. Besides, the data downloader is robust since it supports resumable download. The first-downloaded file list contains meta-data about every file to be downloaded such as file name, path, and size. With this information, we can check if a data file is already downloaded completely. Once interrupted, the data downloader will skip those downloaded files and continue with the remaining data files.

![Figure 6. Data downloading framework of DaT3M.](image)

3.5 Evaluation Method

To evaluate the usability of DaT3M, we invite participants who work as a researcher in the health informatics or healthcare workers that might get involved in data downloading and analysis using CSR. We design a structured one-on-one interview that consists of two sections. In the first section, each participant is instructed to complete five tasks about the system’s key features (see Table 1). The participant is asked to share screen and encouraged to think aloud when performing tasks. The interviewer observes and records the interview, and calculates the time and steps that the participant spend on each task.

<table>
<thead>
<tr>
<th>Task</th>
<th>Task Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>Pick the UH data source and filter cases with MRI imaging</td>
</tr>
<tr>
<td>Task 2</td>
<td>Select one subject and explore the data details</td>
</tr>
<tr>
<td>Task 3</td>
<td>Build a query for subjects with generalized tonic-clonic seizure only during admission</td>
</tr>
<tr>
<td>Task 4</td>
<td>Generate a data download token for the query built in Task 3</td>
</tr>
<tr>
<td>Task 5</td>
<td>Download the data of patients found in Task 3</td>
</tr>
</tbody>
</table>

In the second section, each participant is provided with a 10-question survey based on the participant’s experience in the first section. Survey questions are as follows:
1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

We use the System Usability Scale (SUS) score that has a 5-point Likert scale ranging from strongly disagree (1) to strongly agree (5). The SUS score is a widely used usability evaluation questionnaire that allows us to evaluate our system with a small sample size of participants.

4 Results

4.1 Patient-level Data Availability Tracked in DaT3M

DaT3M is currently deployed at The University of Texas Health Science Center at Houston. Table 2 shows that as of August of 2021, DaT3M had tracked 10,678 data components for 2,741 patients for CSR. These patients are from seven collaborating institutions of CSR, including 1,086 from UH, 297 from NYU, 235 from UCLA, 450 from NW, 210 from TJU, 293 from UCL, and 170 from IUowa. It can be seen that different institutions show disparate patterns of data availability. For example, only UH, UCLA, and UCL have contributed imaging data; and bio-chemistry data only come from UH and TJU.

<table>
<thead>
<tr>
<th>Center</th>
<th># of Patients</th>
<th>EMU Reports</th>
<th>EEG Recordings</th>
<th>MRI Imaging</th>
<th>Bio-chemistry Data</th>
<th>DNA Data</th>
<th>Follow-up Forms</th>
<th>SUDEP Forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>UH</td>
<td>1,086</td>
<td>1,644</td>
<td>1,676</td>
<td>126</td>
<td>137</td>
<td>456</td>
<td>981</td>
<td>22</td>
</tr>
<tr>
<td>NW</td>
<td>450</td>
<td>504</td>
<td>505</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>296</td>
<td>1</td>
</tr>
<tr>
<td>NYU</td>
<td>297</td>
<td>288</td>
<td>308</td>
<td>0</td>
<td>0</td>
<td>124</td>
<td>283</td>
<td>1</td>
</tr>
<tr>
<td>UCLA</td>
<td>237</td>
<td>215</td>
<td>235</td>
<td>207</td>
<td>0</td>
<td>0</td>
<td>143</td>
<td>0</td>
</tr>
<tr>
<td>TJU</td>
<td>210</td>
<td>231</td>
<td>251</td>
<td>0</td>
<td>40</td>
<td>135</td>
<td>161</td>
<td>2</td>
</tr>
<tr>
<td>UCL</td>
<td>293</td>
<td>345</td>
<td>294</td>
<td>296</td>
<td>0</td>
<td>0</td>
<td>288</td>
<td>3</td>
</tr>
<tr>
<td>IUOA</td>
<td>170</td>
<td>171</td>
<td>171</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>137</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>2,741</td>
<td>3,398</td>
<td>3,440</td>
<td>629</td>
<td>177</td>
<td>722</td>
<td>2,289</td>
<td>30</td>
</tr>
</tbody>
</table>

4.2 Data Overview Dashboard

DaT3M provides a dashboard to overview all the data availability, statistics, and distributions of CSR as shown in Figure 7. Six different interactive charts present different aspects of CSR Data. Bar chart A visualizes the same information of Table 2. Panel B lists the total number of datasets for each data type. Charts C and D depict the data distribution in terms of gender and age respectively. Pie chart E shows the number patients from each center. And pie chart F shows the total number of SUDEP cases and its percentage in the whole CSR data.

4.3 Center Data Portal

DaT3M provides a center data portal for each CSR center to track their data submission and data completeness statuses. Figure 8 shows a screenshot of the data portal of CSR center UH, consisting of four areas highlighted and labeled from 1 to 4. Area 1 shows the total number of patients from this center. Area 2 displays the number of data components of each data type. DaT3M’s data status tracking widget is labeled as area 3. Users can click on each status box to change its color and combine various data statuses as a filter to query the center’s data records shown in area 4.
5 Evaluation

Seven researchers from four different health institutes participated one-on-one interviews for the usability evaluation of DaT3M. For each task in the first section of interviews, we calculated the average time in seconds taken for the task, average steps to perform the task, number of participants who were able to complete the task with minimal required steps, and number of participants who failed in completing the task (see Table 3). In task 1, the participants were asked to use the filter function in data tracker to find patients with MRI imaging data. All the participants were able to complete this task, and three of them completed the task with the minimal required steps. In task 2, the participants were asked to explore a patient’s data records and no major issue was found. And this task has no minimum required steps to complete since it is a exploratory task. Task 3 asked the participants to use query builder to find patients with a specific type of seizure. Participants with more domain knowledge of neurology were able to complete the task faster.
than those with less, and 4 participants completed this task with the minimum steps. Task 4 asked the participants to generate a unique token for data download. This is a relatively simple task, and all the participants completed it without any issue. Task 5 is the most complicated task, asking participants to use the command line tool to download research data. Two participants failed this task, since their institutional IT policies did not allow installation of our data download tool, while those who succeeded in the task favored the user-friendliness of this tool.

Table 3. Average time and steps taken for performing usability evaluation tasks.

<table>
<thead>
<tr>
<th>Task</th>
<th>Average time (in seconds)</th>
<th>Average steps (minimum steps)</th>
<th>Participants achieved minimum steps (out of 7)</th>
<th>Participants failed on task (out of 7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Pick the UH data source and filter cases with MRI imaging.</td>
<td>59</td>
<td>4.0 (3)</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>2. Select one patient and explore the data details.</td>
<td>76</td>
<td>9.5 (-)</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>3. Build a query to find patients with generalized tonic-clonic seizure only during admission.</td>
<td>57</td>
<td>7.7 (7)</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>4. Generate a data download token for the query built in Task 3</td>
<td>37</td>
<td>4.2 (4)</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5. Download the data of subjects found in Task 3.</td>
<td>62</td>
<td>6.6 (5)</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 9 summarizes the participants’ responses to the 10-question survey in the second section of interviews. Most people thought that the system was easy to use and they would like to use this system frequently. A final SUS score of 85.3 was achieved. Usually, a SUS score above 68 is considered acceptable, and above 85 is considered as excellent.

6 Discussion

The primary motivation and objective of DaT3M is to promote the FAIRness of research data. DaT3M allows data to be more findable not only at the data set level, but also at data modality and subject levels. Its integrated query engine MEDCIS allows users to conveniently create cohort-level data subsets. DaT3M makes data more accessible through a streamlined downloading process. DaT3M supports data interoperability by adopting common data formats (e.g. European Data Format-EDF) and domain ontologies (e.g. EpSO). DaT3M integrates data submitted from different centers using a general common data model: patient, data type, and data status. Its data scope can be expanded to include data from additional sites. For example, consented patients in EpiToMe, our bespoke EHR for epilepsy care deployed at Memorial Hermann Hospitals, are being added to DaT3M. DaT3M provides support for data quality control and quality assurance processes through the interactive data status panel that tracks data curation workflow and data provenance. High-quality curated data is the key to achieve data reusability.

Patient cohort building and quick data sharing. DaT3M’s data sharing mechanism overcomes the traditional way of manual data preparation and distribution per request. It integrates patient cohort building, data preparation, and data downloading in a highly automated pipeline. With this pipeline, investigators with proper privileges can locate their target datasets and start to download them within a few minutes.

Evaluation feedback. It can be seen from Figure 9 that most participants share similar attitudes toward the survey questions. Most of them (6 out of 7) agreed that the system was easy to use and learn. Most of them (6 out of 7) felt confident while using the system, and they also agreed that other people could learn how to use it very quickly. From additional feedback of the survey, one of the participants thought that the system could save a lot of his time in...
Figure 9. Responses of the survey questionnaire for DaT3M’s usability evaluation.

locate the patient data he needed. Other feedback includes functional suggestions such as providing a different way of downloading data in case that the command line tool does not work, which will be addressed in our future work.

**General applicability.** Although DaT3M has been developed for the CSR study, its data modeling is generally applicable for other multi-site studies. Its mechanism of query by data statuses, patient cohort building, and data slice downloader can be also adapted for other studies.

**Limitation.** The evaluation of DaT3M are preliminary due to the small sample size of participants, although the excellent usability conclusion is valid according to the SUS method. A full-scale evaluation may need to involve crowdsourcing methods such as Amazon Mechanical Turk. Another limitation of DaT3M is the requirement of installation of the data downloader, which was prevented by one of the four evaluation institutions. Alternate data distribution mechanisms such as SFTP are needed to address such institutional policy challenges.

7 Conclusion

In this paper, we introduced a novel research data management tool – DaT3M for large-scale multi-site clinical studies. It models patient data with data modalities and statuses, designs intuitive visualizations for patient data, and automates a pipeline to integrate patient cohort building to data sharing. DaT3M has served as the data management platform for the center for SUDEP research. It facilitates data submission and sharing, tracks data flow and data provenance, and thus improves the FAIRness of CSR data. The pilot evaluation with researchers from multiple institutions indicated that DaT3M achieved highly satisfying usability.

**Acknowledgement**

This research was supported by the National Institutes of Health (NIH) through grants U01NS090408, U01NS090405, and R01NS116287. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

**References**

Physician Experience Design (PXD):
More Usable Machine Learning Prediction for Clinical Decision Making

Lu Wang, PhD1, Mark Chignell, PhD1, Yilun Zhang, MS1, Andrew Pinto, MD3,4, Fahad Razak, MD6,7,8, Kathleen Sheehan, MD2,5, and Amol Verma, MD6,7,8

1Dept. of Mechanical & Industrial Engineering, University of Toronto (UofT), Toronto, ON M5S 2E4, Canada (CA).
2GEMINI, Unity Health, Toronto, ON M5B 1W8, CA.
3Dalla Lana School of Public Health, UofT, Toronto, ON M5S 3G8, CA.
4Dept. of Family and Community Medicine, St. Michael’s Hospital, Unity Health, Toronto, ON M5B 1W8, CA.
5Dalla Lana School of Public Health, UofT, Toronto, ON M5T 2S8, CA.
6Dept. of Psychiatry, UofT, Toronto, ON M5B 1W8, CA.
7Dept. of Medicine, UofT, Toronto, ON M5S 1A8, CA.
8Inst. of Health Policy, Management and Evaluation, UofT, Toronto, ON M5T 3M6, CA.

Abstract
Delirium is an acute neurocognitive disorder, which is difficult to identify and predict. Using GEMINI1, Canada’s largest hospital data and analytics study, we had a labeled sample of around 4,000 cases with approximately 25% of cases being labeled as having delirium. Based on this labeled data, we developed machine learning (ML) models and interacted with physicians to interpret the ML models and their predictions. We developed a preliminary Explainable Artificial Intelligence (XAI) framework for physician experience design (PXD) to improve the uptake of ML models by improving the transparency of model results, thereby increasing physician trust in models as well as the uptake of model results for clinical decision making. We developed our PXD approach first with Conceptual Investigation to collect and extract physicians’ feedback on ML models and their evaluation requirements. We carried out a case study, working closely with the physicians in a participatory design process to develop a dashboard2 that presents ML delirium identification results interactively based on physician selections and inputs. In this approach a physician-preferred ML model for clinical decision making is selected through PXD evaluation.

1 Introduction
Our discussion of physician experience design (PXD) will be framed around a delirium prediction case study. Delirium is described as “acute brain failure” and is considered both a “medical emergency” and “quiet epidemic” [1, 2]. It is the most common neuropsychiatric condition in medically ill and hospitalized patients [3] and recognized as a quality of care indicator in Canada and the United States (U.S.) [4, 5]. Symptoms of delirium can be severe and distressing for both patients and caregivers and result from a complex interaction between predisposing and precipitating factors [1, 2]. Affecting up to 50% of older hospital patients, those with delirium are more than twice as likely to die in hospital or require nursing home placement [6, 7]. The long-term effects of delirium are serious, as it is associated with worsening cognitive impairment and incident dementia [7, 8]. Patients with delirium have longer hospitalizations, increased readmission rates, and more than double the healthcare costs. More recent estimates suggest that it accounts for $183 billion dollars of annual healthcare expenditures in the U.S. [8]. With the growing availability of electronic clinical data repositories such as the one used in this study, methods such as data mining and machine learning can support clinical decision making. However, in order to be useful machine learning (ML) predictions need to be incorporated into clinical workflow. In practice this means that physicians should find predictions, and the manner in which they are presented, to be useful and usable. While there has been considerable discussion of how to improve the Human-AI interaction [9] using methods such as explainable AI (XAI), there has been little if any discussion of how to improve the user experience for physicians when they interact with ML models.

There is, however, relevant prior research concerning the physician user experience when retrieving and viewing

1https://www.geminimedicine.ca/
2https://pxd-dashboard.herokuapp.com/
clinically relevant information on a tablet or handheld device. For instance, the authors in [10] found that physicians preferred the use of white space, alternating shading of rows, and tabular formats for information presentation. The authors in [11] found that the preferred form factor for viewing relevant clinical evidence depended on physician roles. For instance, family physicians preferred tablets, while residents preferred a more mobile form factor that could fit easily into a pocket.

Additional motivation for PXD is provided by examples from human factors engineering. For instance, there were many crashes in World War 2 pilot training due to pilots misreading the height displayed by the altimeter. In the same way that altimeters need to be designed so that they are interpretable by the pilots who use them, ML prediction results need to be presented and framed in a way that will make them understandable and usable for physicians. Another key lesson from human factors in aviation is that short of full automation where the pilot is no longer needed, the pilot as supervisory controller needs to remain in “in the loop” so that the human operator can take over when the automation fails, or when manual input is needed [9]. This problem of designing appropriate user interfaces for highly, but not fully, automated systems is also one that threatens the safety of highly automated vehicles [12].

Clinical decision making is more like a supervisory control task than a continuous vehicular control task. Clinical decision making also has a feature where there is a division of roles between nurses who continuously monitor the patient and provide planned inputs, and physicians who consult with the patient and receive reports from time to time, ordering tests or interventions as needed. If ML prediction is to be more than an academic exercise, then prediction needs to be effectively integrated into clinical decision making and physician workflow. Since ML prediction is not now, and probably never will be, 100% correct, physicians need information about the quality and uncertainty of prediction being provided to them.

Unfortunately, automated machine learning (aML) methods such as deep learning often have a hidden decision making process and lack transparency. Inspired by the benefits of putting the human in-the-loop, XAI was proposed to help human users understand ML models cognition, trust the results and effectively manage the decision making process [13].

As we worked with the physicians in our research study we found that their informational needs went well beyond traditional evaluation metrics such as accuracy, sensitivity and specificity, F1, or Area under the ROC curve. In this paper we focus on the questions and concerns raised by the physicians while reviewing the output of ML models for identifying delirium. Inspired by the user experience design (UXD) approach we derive a preliminary model of PXD, within XAI, that is built around the key information that physicians need in their clinical decision making.

This work was carried out as part of a project to efficiently identify delirium cases during hospitalization using all data available from admission to discharge and in our work we initially focused on the methodological goal of demonstrating the value of incorporating physician expertise in evaluating the delirium identification performance of ML models.

Figure 1: Scheme of making machine learning models more usable integrating physician expertise with physician experience design (PXD).

Figure 1 summarizes our workflow incorporating physicians in ML models evaluation with two major parts, i.e., Conceptual Investigation (CI) for studying the physicians’ needs on evaluating ML models and Technical Investigation (TI) for implementing a platform (a dashboard in our case) to present the physician-selected results based on extracting information from CI. Instead of presenting physicians with the results that ML experts prefer, we show the delirium identification results based on physician preferences. The PXD approach enables physicians to evaluate model performance in an interactive way and thus makes ML models, with corresponding results, interpretable and usable. In the remainder of this paper we report on the lessons learned during this work with respect to the needs that physicians have for an appropriate PXD when incorporating ML models for delirium identification into their clinical decision
making. The main contributions of this paper are as follows:

- A physician experience design (PXD) approach is proposed within the XAI framework for making ML prediction model interpretable and usable in the context of clinical decision making.
- Methods are proposed for collecting and extracting data to satisfy physicians’ requirements for model evaluation.
- Methods are proposed for reporting model evaluation metrics within the PXD framework.
- Methods are proposed for reporting the calibration and bias of models.
- Methods are proposed for assessing the stability of models, and evaluation metrics, over time.
- A PXD dashboard is developed to facilitate review of key aspects of model prediction during clinical decision making (using radar, calibration, and temporal stability charts) – in dropdown menu format.

2 Materials & Methods

2.1 GEMINI Study

GEMINI is a unique big data collaborative supporting cutting-edge quality improvement and research projects [14, 15, 16, 17]. It includes 370,000+ patient admissions from 207,000+ unique patients and 20+ hospitals across Ontario, Canada and it contains over a billion data points. A rigorous internal validation process has demonstrated 98-100% accuracy across key data types [18]. In this study we focus on GEMINI data from six large hospitals (St. Michael’s Hospital, Toronto General Hospital, Toronto Western Hospital, Trillium Credit Valley Hospital, Trillium Mississauga Hospital and Sunnybrook Hospital) containing data on all hospitalizations to general internal medicine (N= 240,000), in those six hospitals, from 2010-2017.

![Figure 2: Data Contained in GEMINI project.](image-url)

### 2.1.1 Research Ethical Review Board (REB) Approval

The Research Ethics Review Board (REB) at the Toronto Academic Health Science Network approved the GEMINI study on 08/31/2019 with REB reference number 15-087. The extension of the REB approval was issued on 08/17/2020 by the Unity Health Toronto REB under the same reference number 15-087.

Our paper is also part of the GEMINI sub-study, named "Using artificial intelligence to identify and predict delirium among hospitalized medical patients", which was approved by University of Toronto (UofT) REB on 10/15/2019 with RIS Protocol Number 38377. The UofT REB approved the renewal of this sub-study on 09/01/2020 under the same reference number 38377.

2.1.2 GEMINI Data Set

In GEMINI, administrative health data are linked with clinical data extracted from hospital information systems at the individual patient level (Figure 2).

- Administrative Data: Patient-level characteristics are collected from hospitals as reported to the Canadian Institute for Health Information Discharge Abstract Database (CIHI-DAD) and the National Ambulatory Care Reporting System (NACRS). Diagnosis data and interventions are coded using the enhanced Canadian International Statistical Classification of Diseases and Related Health Problems (ICD-10-CA) and the Canadian Classification of Health Interventions.
• Clinical Data: Data from the electronic information systems in GEMINI include: laboratory test results (biochemistry, hematology, microbiology), blood transfusions, in-hospital medications, vital signs, room transfers, and routine clinical monitoring. The quality of key elements of these data has been ensured through statistical quality control processes and direct data validation [18]. GEMINI data extraction methods allow access to a wealth of data ideal for text processing methods, including radiologist reports of diagnostic imaging.

The delirium cases were identified through manual medical record review by trained medical professionals using a validated method [19] in GEMINI. Developed by [20], this method has good sensitivity (74%) and specificity (83%) compared to clinical assessment and is considered a suitable gold-standard for identification of delirium for research and quality improvement [20]. Inter-rater reliability was assessed by having 5% of the charts blindly reviewed by a second abstractor, achieving 90% inter-rater reliability.

We used 11 data files from a subset of the GEMINI data set, which contained 3,862 hospital admissions that were labelled with delirium status (positive or negative). The data files include clinical and administrative data as described below.

2.2 Machine Learning (ML) Models Construction and Training for Delirium Status Identification

The scheme shown in Figure 1 was run with 12 supervised classification algorithms with the task of predicting delirium status. The 12 machine learning (ML) algorithms used, covering most categories of ML models used, were:

• Ensemble ML models: Gradient Boosting Classifier (GBC); AdaBoost Classifier (ABC); Random Forest (RF); Voting Classifier Soft (VC-S).
• Non-parametric ML models: k Nearest Neighbors (kNN); Decision Tree (DT).
• Linear-parametric ML models: Logistic Regression (LR); Linear Support Vector Machine (LSVM); Linear Discriminant Analysis (LDA).
• Non-linear parametric ML models: Quadratic Discriminant Analysis (QDA); Neural Network (NNW): Multi-layer Perceptron classifier in deep learning.
• Bayesian-based ML models: Gaussian Naïve Bayes (GNB)

For the modeling, we split our integrated complete data into two parts, a training set and a testing set. As shown in Figure 3, the data extended over a five year period, from 04/01/2010 to 03/31/2015. We divided this period into 10 six-month segments. We treated the first 9 segments, i.e., 04/01/2010 to 09/30/2014, as the training set. The last-6-month period, i.e., 10/01/2014 to 03/01/2015, was used as hold-out data (i.e., the testing set) to test our models’ performance in making a prospective prediction. Cycling through different holdout sets over time allowed us to assess whether there was any non-stationarity in the data, which would affect our ability to predict delirium in the future based on models developed on currently available data. In the training set, we used 5-fold cross validation to tune the model parameters for each of 12 machine learning algorithms. We then used the tuned parameters from the 5-fold cross validation to identify delirium status of each admission in the testing/hold-out set.

2.3 Commonly Used ML Model Evaluation Metrics

We then tested the model performance on the hold-out testing set and calculated six evaluation metrics, i.e., accuracy, precision, recall/sensitivity, specificity, Area Under the Receiver Operating Characteristic Curve (ROC-AUC), F1 score.

Accuracy was calculated as the proportion of predicted labels that matched the corresponding ground truth labels. 
\[
\text{Accuracy} = \frac{TP}{TP + FP},
\]
where \( TP \) denotes the number of true positives. 
\[
\text{Precision} = \frac{TP}{TP + FP},
\]
where \( TP \) and \( FP \) denote the number of true positives and the number of false positives, respectively. 
\[
\text{Recall} = \frac{TP}{TP + FN},
\]
where \( TP \) and \( FN \) denote the number of false negatives, respectively. 

ROC curve was plotted using the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. ROC-AUC was obtained via the probability that our binary classifier ranked a randomly chosen positive instance higher than a randomly chosen negative one (assuming ‘positive’ ranks
Holdout set is varied from TS1 to TS10, i.e., TS$_i$, $i \in \{1, 2, 3, \ldots, 9, 10\}$, for testing.

All TS but TS$_i$ for training.

Figure 3: Data splits for models training and testing on a rolling basis.

higher than ‘negative’). F1 score is a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. $F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$.

2.4 Physician Experience Design (PXD) Evaluation

2.4.1 Physician Team

We interacted with medical delirium experts who address delirium in both their research and clinical work. Our physician teams had four members (also co-authors on this paper): Dr. Fahad Razak is an internist and epidemiologist, with specialization in observational methods, ‘big data’ projects, and global health at St. Michael Hospital, Toronto, CA. Dr. Kathleen Sheehan is a staff psychiatrist with the medical psychiatry program at the Centre for Mental Health-University Health Network. Dr. Amol Verma is staff physician in General Internal Medicine at St. Michael’s Hospital and a scientist in the Li Ka Shing Knowledge Institute. Dr. Andrew Pinto is a Public Health and Preventive Medicine specialist and family physician at St. Michael’s Hospital, and an Assistant Professor at the University of Toronto.

2.4.2 Conceptual Investigation (CI)

Working with the physicians we developed a questionnaire over several iterations. Questions of interest related to the types of evaluation metric that they found most useful, and how they would prefer different evaluation metrics, or combinations of evaluation metric, to be visualized. Sample questions are shown in Figure 4.

We collected physician responses and summarize their recommendations as follows: 1). Use all six evaluation metrics but also allow a way for the physician to focus on an evaluation metric of particular interest; 2). False alarm and miss rates should also be available for viewing (e.g., in a confusion matrix); 3). Measures of temporal stability of prediction should be shown both as plots of prediction performance over time and also as plots or tables of standard deviations of performance measures; 4). Calibration is an important issue and a visualization of calibration results should be available; 5). Radar charts (i.e., plots showing the value of all evaluation metrics in a single plot, can help physicians and other stakeholders see all the values together and compare them visually); 5). The F1 score (i.e., a balanced score of precision and recall/sensitivity) is the most useful single measure, but recall and sensitivity are also of particular interest.
2.4.3 Technical Investigation (TI)

**PXD Evaluations** Based on the CI, we added two metrics into the results table: 1) False positive rate (FPR)\(=\frac{FP}{FP+TN}\), also called false alarm, which answers the question like of all the people who are without delirium, how many of those we incorrectly predict? 2) False negative rate (FNR)\(=\frac{FN}{FN+TP}\), also called miss rate, which answers the question of all the people who are with delirium, how many of those we incorrectly predict?

We also developed a stability measure integrating the aforementioned eight evaluation metrics, i.e., accuracy, precision, recall/sensitivity, specificity, ROC-AUC, F1 score, FPR and FNR. We calculated the standard deviation (SD) and range over 10 time segments on each of the eight metrics, using their variation to show the stability of performance. We used the harmonic mean to show average performance across the matrices, i.e., \(\frac{1}{m_1 + \frac{1}{m_2} + \frac{1}{m_3} + \cdots + \frac{1}{m_n}}\), where \(n\) is the total number of metrics and \(m_i\) is the \(i^{th}\) metric value.

We showed how well calibrated predictions were by using reliability diagrams (also called calibration curves) with one example being shown in Figure 7.(b). The dashed line is the perfectly calibrated line and the line around the perfectly calibrated curve is the calibration performance of the ML model (e.g., the blue curve in Figure 7.(b) is the calibration curve of the Gradient Boosting Classifier, abbreviated as GBC). When the curve falls below the diagonal, the ML model has over-forecast (estimating the delirium probability to be higher than it actually is, while the curve above the diagonal, indicates under-forecasting. Thus the calibration plot indicates the amount of bias towards positive or negative labelling at different levels of probability for the positive label.

The Radar chart is another visualization that we added in the TI stage, where multiple evaluation metrics are compared in the form of a 2D plot with values plotted on a set of axes radiating from a common origin. An example radar plot is shown in Figure 7.(c), displaying several evaluation metrics for the GBC model in a single plot.

**PXD Dashboard** For better interaction with physicians, we developed a PXD dashboard to present the ML models results. Based on physicians’ feedback from CI and TI, we implemented three types of dashboard built in one overall interface, which can be accessed at https://pxd-dashboard.herokuapp.com/.

Figure 5 shows partial views of our PXD dashboard, where three dropdown menus allow users to customize their view of the results. The first dropdown allows a selection from different types of dashboard. The first type is PXD selected machine learning (ML) models. Only four ML models’ results were included in this type, based on the physicists’ feedback from CI. The second type of dashboard is a comparison of 12 ML Models. The third type shows results for all 12 models in a single view. Note that the performance ranking (PR) is created based on the F1 score consistent with the preferences of our physicians. The next dropdown chooses the model. Note that users can choose a particular ML model from the list of 12 models and can then choose the form of evaluation feedback from
the following options: Confusion matrix, Calibration plot, Original Result Table, Radar chart, Temporal Performance and Temporal Stability Measure. Aside from the last option for temporal evaluation, which reviews 10 time segments (TS), the other evaluations only present results from the last TS/last-6-month (prospective hold-out) period. The Radar Chart, Calibration Plot and Result Table are presented with 12 results reflecting the 12 models used in our case study.

3 Results

3.1 Experimental Setup

We built 12 machine learning (ML) models for identifying delirium status. The sample size of input data in each path is presented in Table 1. For more details of features description, please refer to [14, 15, 16, 17].

<table>
<thead>
<tr>
<th>Number of admissions/rows</th>
<th>Number of features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. iML-Delirium scheme</td>
<td>3862</td>
</tr>
<tr>
<td></td>
<td>324</td>
</tr>
</tbody>
</table>

The 12 ML models, along with hyperparameter tuning and cross validation, were implemented in the Python package Scikit-learn [21]. Hyperparameter tuning was conducted using RandomizedSearchCV and GridSearchCV functions. Cross validation was employed via cross_val_score, cross_val_score and cross_val_score functions. Gradient boosting classifier was trained via GradientBoostingClassifier function. AdaBoost classifier used AdaBoostClassifier function. Neural network classifier was implemented by MLPClassifier function. Decision tree classifier was employed via DecisionTreeClassifier function. k nearest neighborhood (kNN) was trained via KNeighborsClassifier function. Logistic regression classifier was conducted using LogisticRegression function. Random forest classifier was implemented via RandomForestClassifier function. Linear Support vector machine (SVM) was employed via svm function with kernel = 'linear'. Gaussian Naive Bayes was implemented via GaussianNB function. Linear discriminant analysis classifier was trained via LinearDiscriminantAnalysis function. Quadratic discriminant analysis classifier was employed via QuadraticDiscriminantAnalysis function. Voting classifiers with soft setting was implemented by VotingClassifier function.

3.2 Experimental Results

We trained these models with hyperparameter tuning and 5-fold cross validation across 9 time segments. We then tested the tuned model on a tenth time segment on a rolling basis so that all of 10 time segments were used as the holdout set on one occasion. The 12 ML models’ results can be accessed at https://pxd-dashboard.herokuapp.com/ in the third type of dashboard. Due to page limitation in this paper we only compare the performance of the 12 models in Figure 7 and show the best-performing one (i.e., GBC), selected by physicians through
4 Discussion

Our proposed physician experience design approach, within an XAI framework, uses a dashboard that enables effective interaction with physicians. The dashboard presents information to physicians efficiently, so that the ML models and their corresponding results can be interpreted in a way that physicians understand. The dashboard presented here only includes PXD selected ML models based on that CI stage, so that the demonstration of ML model performance can be quickly displayed to physicians. We implemented a screening process so that physicians wouldn’t be overwhelmed with information. The diverse visualization of results should give physicians a more intuitive sense of ML model performance. To give consideration to various users’ needs, we also developed two other types of dashboard comparing performance across 12 models.

In contrast to ML model selection based only on the values of evaluation metrics, physicians selected models based on more complex evaluation criteria such as the interpretability/explainability of the algorithm, good calibration, stable performance over time without failure in any time period, and higher sensitivity. As a result, GBC was selected by physicians because it’s a tree based model with good interpretability [22, 23], is well calibrated and its performance was relatively stable over time and with generally higher sensitivity than the other models had.

Currently, communication between developers team and physicians is via emails, meeting and questionnaire, which exhibits time latency and limits efficiency. To address this limitation, we plan to develop a compatible physician-centered interactive ML system including both ML models development and evaluation to boost the efficiency of the PXD process.

5 Conclusion

Delirium is a highly prevalent, preventable and treatable neurocognitive disorder, which is associated with very poor outcomes when untreated. It is characterized by acute onset of fluctuating mental status, psychomotor disturbance...
and hallucinations and thus is difficult to spot, creating an opportunity for higher quality care through automated identification of delirium, or of delirium risk. In the research reported in this paper, we have presented a PXD approach within an XAI framework using a dashboard that physicians can interact with. Allowing physicians to explore models and their predictions through the dashboard should improve trust in ML models and their corresponding results as well as an understanding of likely defects in model prediction so that the physicians can use ML more effectively in clinical decision making for delirium. Developing PXD for physicians is a delicate balance between providing a rich set of information (able to dive into details) and providing a concise overview (just what the physician needs to know). The physicians working with us in this study were comparatively knowledgeable about data science and ML prediction. Further research is needed to determine how much training different types of physician will need to have in order to be able to use ML prediction results effectively in their clinical decision making, and to determine the extent to which a PXD approach can reduce those training requirements.

Acknowledgment

The authors would like to thank the Canadian Institutes of Health Research Foundation and the National Science and Engineering Research Council for funding this work with a Collaborative Health Research Projects Grant (Application Number 415033). This research was also supported by an NSERC Discovery Grant to the second author (RGPIN-2018-06591).

References


Prior Knowledge Enhances Radiology Report Generation

Song Wang\(^1\), Liyan Tang\(^1\), Mingquan Lin, Ph.D.\(^2\), George Shih, M.D.\(^3\), Ying Ding, Ph.D.\(^1\), Yifan Peng, Ph.D.\(^2\)

\(^1\) The University of Texas at Austin, Austin, TX, USA; \(^2\) Population Health Sciences, Weill Cornell Medicine, New York, NY, USA; \(^3\) Department of Radiology, Weill Cornell Medicine, New York, NY, USA

Abstract

Radiology report generation aims to produce computer-aided diagnoses to alleviate the workload of radiologists and has drawn increasing attention recently. However, previous deep learning methods tend to neglect the mutual influences between medical findings, which can be the bottleneck that limits the quality of generated reports. In this work, we propose to mine and represent the associations among medical findings in an informative knowledge graph and incorporate this prior knowledge with radiology report generation to help improve the quality of generated reports. Experiment results demonstrate the superior performance of our proposed method on the IU X-ray dataset with a ROUGE-L of 0.384 ± 0.007 and CIDEr of 0.340 ± 0.011. Compared with previous works, our model achieves an average of 1.6% improvement (2.0% and 1.5% improvements in CIDEr and ROUGE-L, respectively). The experiments suggest that prior knowledge can bring performance gains to accurate radiology report generation. We will make the code publicly available at https://github.com/bionlplab/report_generation_amia2022.

1 Introduction

A radiology report provides a translation of radiographs into text, presenting a synopsis of the process of detailed findings and thoughtful impressions\(^1\). It gives descriptive information about a patient’s history, symptoms, and interpretations of relevant radiology images\(^2\). Therefore, radiology report writing has long been a time-consuming and labor-some process that requires domain expertise. To alleviate the burden of radiologists, there is an unmet need to develop automatic report generation systems.

Radiology report generation takes chest X-ray images as input and generates a descriptive report to support better diagnostic conclusion inferences beyond disease labels. This task is close to visual captioning\(^3\)–\(^6\) but posts new challenges. First, radiology report generation outputs a sequence of sentences, while visual captioning usually produces only one sentence. Second, radiology report generation requires extensive domain knowledge to produce clinical-coherent texts. For example, it must follow critical protocols, including the correct use of medical terms to describe normal and abnormal medical observations\(^2\).

Radiology report generation has attracted more attention recently\(^7\)–\(^12\). Xue et al. fused the visual features and the semantic features of the last sentence through an attention mechanism, and used this fusion to generate the next sentence in a recurrent generative manner\(^11\). Wang et al. presented a text-image embedding network to jointly learn the text and image information, and integrated an end-to-end trainable CNN-LSTM architecture with multi-level attention models for chest X-ray reporting\(^8\). Jing et al. employed a co-attention model.
over visual features and textual embeddings, and proposed a multi-task learning framework that jointly predicts disease tags and generates reports\(^7\). Chen et al. recorded vital information during the generation process by presenting a memory-driven Transformer to further assist report generation\(^10\). Jing et al. used reinforcement learning to exploit the structure information between and within report sections for generating high-quality reports\(^9\). Liu et al. combined self-critical sequence training with reinforcement learning to optimize the correct mentions of disease keywords in reporting\(^12\).

Though achieving good results, few works considered incorporating prior knowledge, which can provide supplementary information for accurate reporting. For example, medical observations presented in a chest X-ray image are usually not isolated from each other, where underlying mutual influences may exist. Compared to experienced radiologists aware of such relationships, deep learning methods tend to suffer from the lack of knowledge if not explicitly taught, which limits the generation accuracy. Modeling the associations among medical observations in the form of a knowledge graph enables us to further utilize prior knowledge to produce high-quality reports. To this end, Zhang et al.\(^13\) and Li et al.\(^14\) combined graph-based knowledge inference with an encoder-decoder pipeline for radiology report generation. However, their prior knowledge is manually pre-defined, thus, requires domain experts to be closely involved in the design and implementation of the system. Owing to the rigid graph, their approaches usually achieve a high precision but may miss some important findings. While it is feasible to manually identify and implement a high-quality knowledge graph to achieve good precision, it is often impractical to exhaustively encode all the nodes and relations in this manner. Our work enables the text-mined prior knowledge as a universal knowledge graph to mitigate some of these concerns.

This paper presents an innovative framework of knowledge-based report generation, which seamlessly integrates prior domain and linguistic knowledge at different levels. First, we study a data-driven approach to automatically capture the intrinsic associations among the concepts in the RadLex radiology ontology\(^15\). This prior knowledge serves as a natural extension to the human-designed one\(^13\). Disease findings are defined as nodes in the graph, and correlated findings are connected to influence each other during graph propagation. Second, we build a graph convolutional neural network to model the prior knowledge on chest findings\(^16\). Frontal-view and lateral-view images of chest X-ray are fed into a convolutional neural network extractor for image feature extraction, and the features together with the built graph are passed to a three-layer graph convolution network through an attention mechanism to learn dedicated features for each graph node. Later these node features are passed to two branches, one linear classifier for disease classification, and one two-level decoder for report generation. Different from previous studies, extra text-mined concepts are included in the model as auxiliary nodes so that the model enriches its expression power by training on existing datasets with image-level diseases annotated. We hypothesize that these text-mined labels may reflect known features identified in chest X-rays and add granularity to the association strength of those features in the generated reports.

To train the model, we adopt an existing two-step procedure\(^13\): multi-label classification followed by report generation. Such a training strategy simulates the reading routine of radiologists by first observing multiple findings when they read medical images and then compiling radiological reports. Specifically, we first train a multi-label classifier where each class label corresponds to one medical finding and hence one node in the knowledge graph. After training the multi-label classifier, we keep the classifier frozen, and train a two-level decoder that consists of one topic-level Long Short-Term Memory (LSTM\(^17\)) and one word-level LSTM. The two-level decoder encourages each generated sentence to focus on one different topic. Fig 1 shows the generation pipeline and an example of the generated report.

Our contributions are outlined as follows. (1) We text-mine and model the prior knowledge in a graph; (2) We incorporate the prior knowledge with graph-based knowledge inference to enhance report generation; (3) Extensive experiments on the IU X-ray dataset\(^18\) show that our proposed model outperforms state-of-the-art methods. (4) We make codes, models, and pre-processed data publicly available.

The rest of the paper is organized as follows. We describe the multi-task learning in Section 2, followed by our experimental setup, results, and discussion in Section 3. We conclude with future work in the last section.
2 Methods

2.1 Framework

Simulating the reading routine of radiologists by first observing multiple findings when they read medical images and then compiling radiological reports, our proposed method generates radiology report $S$ from the frontal-view $I_f$ and lateral-view $I_l$ of chest X-ray images following several steps (Fig 2). We first constructed the prior knowledge graph (i.e., relationships between medical findings) in a data-driven manner (Section 2.2). The frontal-view and lateral-view images are fed to an image encoder to extract visual features (Section 2.3), which are then passed to a graph convolution network (GCN) based on the knowledge graph (Section 2.4). The network then propagates the semantic correlations among radiology concepts based on the prior knowledge graph, therefore, injecting domain knowledge into concept representation learning. We then concatenate the node features of both views and train a report decoder (Section 2.5).

![Image encoder](DenseNet) → Report

Figure 2: The proposed framework.

2.2 Prior Knowledge Graph Construction

In our study, the nodes in the knowledge graph are radiology concepts (e.g., diseases or body parts) and edges are the semantic correlations among the concepts. Our knowledge graph consists of two parts. The first part was manually defined by domain experts. The second part consists of supplementary concepts and their correlations text-mined from the radiology reports in a data-driven manner. Specifically, we first build a rule-based tool to greedily match concepts in RadLex on sequences of the lemmatized tokens in the reports (i.e., longer matches are returned where possible). We then select the concepts with top-$q$ appearing frequencies if they have not been included in the graph in Zhang et al. Here, we only focus on three categories of interest: Anatomical entity, Clinical finding, and Imaging observation. We then examine the document-level co-occurrences of concepts to build a correlation matrix and binarize the matrix to prevent overfitting.

2.3 Image Encoder

We employ DenseNet-121 (pre-trained on CheXpert) as our image encoder backbone. One frontal-view and one lateral-view chest X-ray image are fed into DenseNet-121 to extract visual features, which are then used to initialize graph node features in two steps. The first step initializes the feature of the global node by average pooling the visual features of the frontal-view and lateral-view images. The second step initializes the features of the remaining finding nodes in the graph through a spatial attention mechanism. We use a convolution layer with kernel size one followed by a softmax to compute the spatial attention weights, where the number of channels is equal to the number of finding nodes in the graph. Finally, we concatenate the global node feature computed from step one, with the weighted sum of visual features, where the weights come from step two, to initialize the graph node features.

2.4 Graph Convolution Network

We employ a GCN to model inner correlations among radiology concepts. The graph structure is constructed based on the graph detailed above. The GCN updates its node representations by message passing. The graph convolution is...
expressed as:

\[
\hat{H}^l = ReLU(BN(Conv1d(H^l)))
\]

\[
m = ReLU(D^{-1/2} \hat{A}D^{-1/2}H^lW^l)
\]

\[
H^{l+1} = ReLU(BN(Conv1d(concat(\hat{H}^l, m))))
\]

where \(H^l\) is the states in the \(l\)-th layer, with \(H^0\) initialized using the output of image encoder. \(\hat{A} = A + IN\) is the adjacency matrix with added self-connections, where \(A\) is the graph adjacency matrix, \(IN\) is the \(N\)-dimension identity matrix, \(D = \text{diag} \sum_j A_{ij}\) is the diagonal node degree matrix, \(BN\) is the batch normalization, and \(W^l\) is a trainable layer-specific weight matrix.

2.5 Report Generation Decoder

Radiology reports usually contain several sentences where each sentence focuses on one topic. Therefore, we adopt a two-level LSTM structure\(^13\). We input the graph node features to an attention module to obtain the context vector, which attends graph node features to different topics. The vector is then fed to a topic-LSTM to generate topics, and the output topic vectors are passed to a word-LSTM to generate sentences in a word-by-word fashion.

2.6 Training Procedure and Loss Functions

Our framework combines two loss functions. Suppose \(p(S_t)\) is the probability of observing the correct word \(S_t\) at time \(t\). The first loss is the cross-entropy loss of the predictions on the whole report,

\[
- \sum_{t=1}^{N} \log p(S_t | I_t, I_f; \theta)
\]

Additionally, we take the features of each node in the graph and do average pooling. We then fit a linear classifier to predict the diseases present in the images. Clinically, it simulates the reading routine of radiologists by first observing multiple findings when they read medical images and then compiling radiology reports.

To use the ground truth labels, we divide the nodes in the knowledge graph into two types: primary and auxiliary. The primary nodes are the chest observations with ground truth labels in the dataset. The auxiliary nodes are the supplementary concepts mined from the reports. We then use weighted binary cross-entropy loss defined on the primary nodes\(^23\) for training.

Our model is trained using the same two-step training procedure as in Zheng et al.\(^13\): the multi-label classifier is trained first, and then we fixed the image encoder and GCN modules when training the report decoder.

3 Results

3.1 Datasets

We use the publicly available IU X-ray dataset\(^18\). IU X-ray dataset contains 3,955 de-identified radiology reports, with each report associated with one frontal-view, and one lateral-view chest X-ray image. Several sections (e.g., findings and impressions) are covered in each radiology report, where findings describe the medical findings and impressions summarize the overall diagnoses.

We only consider the cases with complete findings and impression sections, and with both frontal-view and lateral-view images present, which results in 2,912 reports and 5,824 images. We concatenate the findings and impression sections in each report as the ground truth. All the words in the ground truth are tokenized, converted to lower case. Infrequent words with a frequency of less than three are dropped, which results in 1,103...
unique tokens. The ground truth labels are obtained by detecting the corresponding labels in the Mesh part of the reports, where findings are listed in a formatted manner. After pre-processing, we have a total of 2,912 reports and 5,824 images (Table 1), where each report is associated with one frontal-view and one lateral-view of the chest X-ray image.

3.2 Experimental Settings

We adopt DenseNet-121 pre-trained on CheXpert as our backbone CNN model. Images are randomly cropped to 512×512 with padding if needed, and the feature map from DenseNet-121 block four is of size 1024×16×16. We replace the last fully connected layer of DenseNet-121 with a multi-label classification layer appended with attention and three graph convolution layers with 256 hidden units.

Our model is trained for 150 epochs with a batch size of 8. Adam is used for optimization with a learning rate of 1e-6 and weight decay of 1e-5. We use 200-dimension GloVe word embeddings in the decoder.

3.3 Evaluation metrics

To evaluate the generation results, we compute BLEU scores, ROUGE-L, and CIDEr, which reflect different aspects of the performance, where BLEUs reflect the precision, ROUGE-L is closer to recall, and CIDEr measures consensus.

In this study, we used the bootstrap to assess the statistical significance of the results. For the training dataset, we sampled 1,747 instances with replacement to train the models. We then evaluated the model on the held-out test dataset. By repeating this sampling, training, and evaluation 15 times, we obtained a distribution of the performance metrics and reported the 95% confidence intervals (CI).

3.4 Results and Discussion

3.4.1 Report generation

We compare our method with four previous works on radiology report generation, including SAT framework in Xu et al., multi-level LSTM framework in Yuan et al., two-level LSTM augmented with knowledge graph in Zhang et al., and CoAtt in Jing et al. Table 2 shows the performance comparison. Our method achieves 0.4%-2.0% improvements over previous works, especially 2.0% improvement in CIDEr, 1.5% in ROUGE-L, and 1.0% in BLEU-2.

Table 2: Comparisons on the test set of IU X-ray dataset. The 95% CI is reported for our model.

<table>
<thead>
<tr>
<th></th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>ROUGE-L</th>
<th>CIDEr</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xu et al.</td>
<td>0.433</td>
<td>0.281</td>
<td>0.194</td>
<td>0.138</td>
<td>0.361</td>
<td>0.320</td>
<td>0.288</td>
</tr>
<tr>
<td>Yuan et al.</td>
<td>0.445</td>
<td>0.289</td>
<td>0.200</td>
<td>0.143</td>
<td>0.359</td>
<td>0.268</td>
<td>0.284</td>
</tr>
<tr>
<td>Jing et al.</td>
<td>0.455</td>
<td>0.288</td>
<td>0.205</td>
<td>0.154</td>
<td>0.369</td>
<td>0.277</td>
<td>0.291</td>
</tr>
<tr>
<td>Zhang et al.</td>
<td>0.441</td>
<td>0.291</td>
<td>0.203</td>
<td>0.147</td>
<td>0.367</td>
<td>0.304</td>
<td>0.292</td>
</tr>
<tr>
<td>Our model</td>
<td>0.450±0.005</td>
<td>0.301±0.003</td>
<td>0.213±0.004</td>
<td>0.158±0.005</td>
<td>0.384±0.007</td>
<td>0.340±0.011</td>
<td>0.308</td>
</tr>
</tbody>
</table>

3.4.2 Ablation studies

We conducted ablation studies to verify the effectiveness of different modules (Table 3). According to (a), adopting the pre-trained word embeddings (GloVe) improves all metrics, where the average score increases by 1.4%. When the proposed GCN is added (ablation study (b)), the BLEU scores increase by up to 0.6% and ROUGE-L increases by 0.6%. Note that CIDEr drops probably because the generated reports contained repeated sentences due to a lack of contextual coherence; while compared to other metrics, CIDEr intends to weigh more on the details of sentences.

We also compare different sizes of nodes in GCN. In addition to the nodes in Zhang et al., we text-mined ten auxiliary
Table 3: Ablation studies on the test set of IU X-ray dataset.

<table>
<thead>
<tr>
<th></th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>ROUGE-L</th>
<th>CIDEr</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our model</td>
<td>0.450</td>
<td>0.301</td>
<td>0.213</td>
<td>0.158</td>
<td>0.384</td>
<td>0.308</td>
<td></td>
</tr>
<tr>
<td>(a) Random embs</td>
<td>0.421</td>
<td>0.281</td>
<td>0.202</td>
<td>0.151</td>
<td>0.367</td>
<td>0.340</td>
<td>0.294</td>
</tr>
<tr>
<td>(b) GCN in Zhang et al.</td>
<td>0.446</td>
<td>0.297</td>
<td>0.209</td>
<td>0.152</td>
<td>0.378</td>
<td>0.357</td>
<td>0.307</td>
</tr>
<tr>
<td>(c) 20 Nodes</td>
<td>0.460</td>
<td>0.307</td>
<td>0.215</td>
<td>0.155</td>
<td>0.380</td>
<td>0.357</td>
<td>0.307</td>
</tr>
<tr>
<td>(d) 40 Nodes</td>
<td>0.460</td>
<td>0.311</td>
<td>0.223</td>
<td>0.164</td>
<td>0.387</td>
<td>0.310</td>
<td>0.309</td>
</tr>
<tr>
<td>(e) 60 Nodes</td>
<td>0.446</td>
<td>0.296</td>
<td>0.209</td>
<td>0.152</td>
<td>0.372</td>
<td>0.318</td>
<td>0.299</td>
</tr>
</tbody>
</table>

nodes from the IU X-ray dataset, including Chest pain, Shortness of breath, Granuloma, Lymphadenopathy, Deformity, Granulomatous disease, Congestion, Tuberculosis, Infection, and Hypertension. Ablation studies (c) - (e) show that the performance drops significantly when switching to a larger graph with 60 nodes. This observation indicates that a larger graph may not always lead to more accurate report generation.

3.4.3 Error analysis

We further studied the correlations between text complexity and evaluation metrics. Fig 3 shows the average number of sentences in the report for different BLEU scores. We find that reports with more sentences are prone to be less accurate and tend to have lower BLEU scores.

3.4.4 Multi-label classification

We monitored the node features in the GCN and measured how suitable they are for the disease classification. We apply global average pooling after the graph convolution layers to obtain the graph-level feature, and further, append a fully connected layer to predict probabilities for each finding as a multi-label classification task. We use Binary Cross Entropy loss during training. While we are using 30 knowledge graph nodes, we only classify on the first 20 finding nodes, leaving the rest 10 finding nodes as auxiliary nodes and not including them when calculating the loss. The numbers of reports labeled with different diseases are as shown in Table 4.

Table 4 also shows that our method achieves comparable results in AUC with the method described in Zhang et al.\textsuperscript{13}. For diseases that appear in more than 10% of the reports (Scoliosis, Hypoinflatin, Opacity, and Cardiomedgaly), the correlation between AUC and BLEU is moderate (0.46, 0.44, 0.45, 0.44, respectively). For other diseases, there are no dependencies between disease classifications and BLEU scores. For example, pleural effusion has a better classification accuracy than Emphysema, but this does not suggest that pleural effusion will have a better report generation accuracy. However, since the number of reports per disease is imbalanced in the IU dataset, more datasets are needed to further investigate their correlations.

3.4.5 Error case analysis

Fig 4 shows three examples where our model respectively generates a low-scored report and two high-scored reports. The first example is labeled with Calcinosis and Thickening. However, our generated report omits the fact that there is a calcified granuloma. The BLEU-1 score is 0.1389. Example 2 and Example 3 are respectively labeled with Normal
Table 4: Disease report distributions, classification accuracies and the quality of generated reports.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Reports</th>
<th>AUC %</th>
<th>Zhang et al.</th>
<th>Proposed</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>1,491</td>
<td>38</td>
<td>0.81</td>
<td>0.81</td>
<td>0.47</td>
<td>0.33</td>
<td>0.24</td>
<td>0.16</td>
</tr>
<tr>
<td>Airspace disease</td>
<td>125</td>
<td>3</td>
<td>0.86</td>
<td>0.81</td>
<td>0.31</td>
<td>0.18</td>
<td>0.10</td>
<td>0.05</td>
</tr>
<tr>
<td>Atelectasis</td>
<td>332</td>
<td>8</td>
<td>0.67</td>
<td>0.70</td>
<td>0.41</td>
<td>0.27</td>
<td>0.19</td>
<td>0.13</td>
</tr>
<tr>
<td>Calciosi</td>
<td>305</td>
<td>8</td>
<td>0.91</td>
<td>0.91</td>
<td>0.30</td>
<td>0.18</td>
<td>0.11</td>
<td>0.06</td>
</tr>
<tr>
<td>Cardiomegaly</td>
<td>375</td>
<td>10</td>
<td>0.73</td>
<td>0.79</td>
<td>0.32</td>
<td>0.19</td>
<td>0.11</td>
<td>0.05</td>
</tr>
<tr>
<td>Cicatrix</td>
<td>196</td>
<td>5</td>
<td>0.89</td>
<td>0.94</td>
<td>0.25</td>
<td>0.15</td>
<td>0.09</td>
<td>0.04</td>
</tr>
<tr>
<td>Edema</td>
<td>46</td>
<td>1</td>
<td>0.67</td>
<td>0.76</td>
<td>0.36</td>
<td>0.18</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>Effusion</td>
<td>161</td>
<td>4</td>
<td>0.88</td>
<td>0.69</td>
<td>0.32</td>
<td>0.18</td>
<td>0.10</td>
<td>0.06</td>
</tr>
<tr>
<td>Emphysema</td>
<td>106</td>
<td>3</td>
<td>0.78</td>
<td>0.79</td>
<td>0.37</td>
<td>0.24</td>
<td>0.15</td>
<td>0.09</td>
</tr>
<tr>
<td>Fracture bone</td>
<td>84</td>
<td>2</td>
<td>0.94</td>
<td>0.97</td>
<td>0.34</td>
<td>0.22</td>
<td>0.15</td>
<td>0.10</td>
</tr>
<tr>
<td>Hernia</td>
<td>48</td>
<td>1</td>
<td>0.81</td>
<td>0.80</td>
<td>0.29</td>
<td>0.18</td>
<td>0.10</td>
<td>0.07</td>
</tr>
<tr>
<td>Hypoinflation</td>
<td>507</td>
<td>13</td>
<td>0.64</td>
<td>0.60</td>
<td>0.27</td>
<td>0.16</td>
<td>0.09</td>
<td>0.05</td>
</tr>
<tr>
<td>Lesion</td>
<td>126</td>
<td>3</td>
<td>0.80</td>
<td>0.82</td>
<td>0.35</td>
<td>0.22</td>
<td>0.14</td>
<td>0.09</td>
</tr>
<tr>
<td>Medical device</td>
<td>362</td>
<td>9</td>
<td>0.86</td>
<td>0.83</td>
<td>0.37</td>
<td>0.25</td>
<td>0.18</td>
<td>0.12</td>
</tr>
<tr>
<td>Opacity</td>
<td>455</td>
<td>12</td>
<td>0.84</td>
<td>0.77</td>
<td>0.30</td>
<td>0.19</td>
<td>0.11</td>
<td>0.07</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>120</td>
<td>3</td>
<td>0.83</td>
<td>0.83</td>
<td>0.28</td>
<td>0.18</td>
<td>0.12</td>
<td>0.08</td>
</tr>
<tr>
<td>Pneumothorax</td>
<td>27</td>
<td>1</td>
<td>0.93</td>
<td>0.87</td>
<td>0.29</td>
<td>0.17</td>
<td>0.09</td>
<td>0.05</td>
</tr>
<tr>
<td>Scoliosis</td>
<td>559</td>
<td>14</td>
<td>0.66</td>
<td>0.64</td>
<td>0.37</td>
<td>0.24</td>
<td>0.15</td>
<td>0.08</td>
</tr>
<tr>
<td>Thickening</td>
<td>56</td>
<td>1</td>
<td>0.73</td>
<td>0.77</td>
<td>0.34</td>
<td>0.23</td>
<td>0.15</td>
<td>0.10</td>
</tr>
<tr>
<td>Others</td>
<td>411</td>
<td>10</td>
<td>0.60</td>
<td>0.61</td>
<td>0.44</td>
<td>0.30</td>
<td>0.21</td>
<td>0.14</td>
</tr>
</tbody>
</table>

and Thickening. We can see that the generated reports are very close to the ground truth reports in terms of concept mentions and descriptions. They report BLEU-1 scores 0.9546 and 0.6757, respectively.

<table>
<thead>
<tr>
<th>Images</th>
<th>Reference</th>
<th>Generated report</th>
<th>Ground-truth labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>negative for acute cardiopulmonary abnormality . no focal consolidation effusion or pneumothorax . normal heart size . stable left subclavian chest . xxxx catheter tip tip in mid svc . lungs are well expanded . calcified granuloma posteriorly in the medial right lower lobe . postsurgical changes of anterior lower cervical spine . thoracic spine demonstrates mild degenerative changes with some mild xxxx loss of &lt;unk&gt; and t7 vertebral bodies .</td>
<td>no acute cardiopulmonary disease . &lt;unk&gt; silhouette is within normal limits . no focal airspace consolidation . no focal airspace consolidation . no pleural effusion . no pneumothorax . the mediastinum is unremarkable .</td>
<td>Calciosi, thickening</td>
</tr>
<tr>
<td>Example 2</td>
<td>no acute disease . the heart is normal in size . the mediastinum is unremarkable . the lungs are clear .</td>
<td>no acute cardiopulmonary disease . the heart is normal in size . the mediastinum is unremarkable . the lungs are clear .</td>
<td>Normal</td>
</tr>
<tr>
<td>Example 3</td>
<td>no acute cardiopulmonary abnormality . the &lt;unk&gt; silhouette and pulmonary vasculature are within normal limits . there is no pneumothorax or pleural effusion . there are no focal areas of consolidation . cholecystectomy clips are present .</td>
<td>no acute cardiopulmonary findings . the &lt;unk&gt; silhouette and pulmonary vasculature are within normal limits . there is no focal airspace consolidation . no pleural effusion or pneumothorax . there are no acute bony abnormality .</td>
<td>Thickening</td>
</tr>
</tbody>
</table>

Figure 4: Three examples where our model generates low-scored reports and high-scored reports.
3.4.6 Limitations and Discussions

Our framework employs DenseNet-121 pre-trained on CheXpert as our backbone, which has the underlying limitation that CheXNet\textsuperscript{31} was fine-tuned on one single-institution dataset. Chest X-ray image quality can also affect the training performance. To understand the impact of image acquisition quality, we experimented adding random Gaussian noises to the training images, and the average multi-label classification AUC drops from 0.786 to 0.683. With the six report generation metrics dropping to 0.384, 0.271, 0.197, 0.145, 0.394, 0.301 respectively, the average accuracy of the generated reports drops from 0.308 to 0.282.

4 Conclusions

In this paper, we propose to incorporate text-mined prior knowledge with radiology reporting by employing a graph convolution module for knowledge inference, followed by multi-label disease classification and report generation. Our model achieves better performances than previous works on the IU X-ray dataset. Furthermore, we verified the effectiveness of different modules through ablation studies.

In the future, we plan to adopt Transformer to improve the contextual coherence and explore other domain knowledge that can be utilized in report generation. We plan to train and evaluate our model on different datasets. While our work only scratches the surface, we hope it will shed light on the future directions for radiology reporting.

Acknowledgment

This work is supported by Amazon Machine Learning Research Award 2020. It was also supported by the National Library of Medicine under Award No. 4R00LM013001.

References


Comparison of Alexa Voice and Audio Interfaces for Home-Based Physical Telerehabilitation

Chenhao Wei, MS 1, Joseph Finkelstein, MD, PhD, FAMIA 1
1Icahn School of Medicine at Mount Sinai, New York, NY

Abstract

The goal of this pilot study was to compare Alexa voice and video interfaces for home-based telerehabilitation dialog by conducting cognitive walkthrough testing. All task performance scores were higher in video interface as compared to the audio interface. The overall task score was significantly higher for video interface (42.4±4.6) as compared to the audio score (41.3±5.9). Comparative usability survey demonstrated higher preference of the video interface as compared to the audio interface. Based in the comparative survey, 85.7% stated they definitely prefer video interface, 85.7% felt that video introduction was simpler to understand, 71.4% felt that exercise instructions were simpler to understand with the video interface, and 78.6% felt that overall navigation was easier with the video interface. The overall time to accomplish all three tasks was significantly shorter (p<0.05) for the video interface (170.5±12.2 seconds) as compared to the audio interface (194.2±10.3 seconds). This is the first study systematically comparing two major Alexa interfaces in a telerehabilitation system. These results are instrumental for future development of Alexa-based telerehabilitation systems.

Introduction

Physical rehabilitation has been shown to slow down disease progression and improve quality of life in a wide range of chronic health conditions [1-3]. Access to life-long rehabilitation programs may be limited due a spectrum of barriers including insurance coverage and mobility limitations [2]. Telemedicine approaches have been shown effective in addressing these barriers [3]. Making telerehabilitation services more acceptable for people with limited health and computer literacy may potentially affect ability of older adults with chronic health conditions to effectively use home-based telemedicine systems. Recent projects explored use of Amazon Alexa as means for delivering various skillsets to individuals with minimal computer literacy. Alexa assistant can potentially help older adults who don’t feel comfortable using computers or computer-like devices to overcome computer literacy barriers by allowing natural voice–driven interactions with the telerehabilitation system. However, use of Amazon Alexa for delivery of rehabilitation services has not been systematically explored. With regard to interface, Alexa devices can be divided in two groups: voice operated units without video capability and voice-operated units with integrated video capability [4]. It is not clear whether addition of video affects usability of these devices for telerehabilitation services. The goal of this pilot study was to compare Alexa voice and video interfaces for home-based telerehabilitation dialog by conducting cognitive walkthrough testing.

Methods

Study Design

The system for online rehabilitation exercise with personalized schedules has been designed based on the voice interaction with Amazon's voice assistant Alexa for aging patients to exercise at home. Patients can use their voice to finish the whole rehabilitation exercise in the system via the Amazon Echo devices. The system design of the Alexa based Remote Rehabilitation System is depicted in Figure 1. All the voice data is collected by the Echo devices and then sent to the Alexa voice-processing server. After the server translates the voice data to string commands, our program will analyze the command's intent and fetch the exercise schedule or media resources to generate the response to send back to the echo device. The whole rehabilitation system contains four stages: exercise schedule check, watch the exercise introduction video, exercise with video, post-exercise survey.
Study Design

All the participants were given a list of instructions and questionnaires to carry out the system's cognitive walkthrough. The participants underwent two tests: one test required interaction with an Alexa device supported only voice communication; and another test required interaction with an Alexa device with a video screen providing visual cues and short videos in parallel to voice communication. The order of these two tests was randomly selected for each of the participant in order to prevent the training bias. Before the test, participants filled out a pre-survey to collect the socio-demographic information and were familiarized with the Alexa commands. All the participants were using the same Echo Show and Echo Dot devices to interact with the system. All the participants were timed for each task. If the participant needed additional help to complete the task, the research assistant noted it in the test report. At the completion of each cognitive walkthrough task, each participant was asked to grade the task from 1 (very difficult) to 5 (very easy) using a survey that contained the following questions: (1) How difficult or easy was it to complete this task? (2) How satisfied are you using this application/system to complete this task? (3) How would you rate the amount of time it took to complete this task? After participants finished all the tasks, they were asked System Usability Scale (SUS) and System Usability Comparison.

Results

The user interface and conversation tree is depicted in Figure 2. Fourteen cognitive walkthrough experiments have been completed. The profiles of the 14 study participants are presented in Table 2. The usability analysis is presented in Tables 2-6. From Table 4, all task performance scores were higher in video interface as compared to the audio interface. Specifically, the total Task 2 score for the video interface (14.3±1.5) were significantly higher than the total audio score (13.7±1.9), the total Task 3 video score (13.6±2.9) was significantly higher than audio score (13.4±3.0), and overall task score was significantly higher for video interface (42.4±4.6) as compared to the audio score (41.3±5.9). Comparative usability survey in Table 6 demonstrated higher preference of the video interface as compared to the audio interface. Based in the comparative survey (table 5), 85.7% stated they definitely prefer video interface, 85.7% felt that video introduction was simpler to understand, 71.4% felt that exercise instructions were simpler to understand with the video interface, and 78.6% felt that overall navigation was easier with the video interface. The overall time to accomplish all three tasks was significantly shorter (p<0.05) for the video interface (170.5±12.2 seconds) as compared to the audio interface (194.2±10.3 seconds).
Table 1. Tasks performed by study participants during cognitive walkthrough (The sentences with * mark are specifically for with only audio experiment).

<table>
<thead>
<tr>
<th>Task 1 Review your exercise plan</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Sit in front of the Alexa Echo Show / Alexa Echo Dot*</td>
</tr>
<tr>
<td>2. Speak out “Alexa, open my exercise.”</td>
</tr>
<tr>
<td>3. Wait for the Alexa response.</td>
</tr>
<tr>
<td>4. Speak out “Alexa, show me the exercise.”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task 2 Review Instruction for “Stand-to-Sit” Exercise</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Choose “Stand to sit” exercise by saying “Alexa, stand to sit.”</td>
</tr>
<tr>
<td>2. Wait for the Alexa to ask you whether you want the introduction or not.</td>
</tr>
<tr>
<td>3. Speak out “Yes” to start watching the introduction video / to start listening the introduction audio.*</td>
</tr>
<tr>
<td>4. Finish the video/audio*.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task 3 Perform the Exercise “Stand-to-Sit”</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Speak out “Alexa, start exercise.”</td>
</tr>
</tbody>
</table>

Figure 2. User Interface and Conversation Tree
2. Follow the video to do the “stand to sit” exercise for 3 times.

3. Click the “Exercise Finished” button / Speak out “Finish*.

4. Answer the post-exercise survey questions.

Table 2. Participant profile (N=14).

<table>
<thead>
<tr>
<th></th>
<th>Mean(SD)</th>
<th>Mean(SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age(years)</td>
<td>37.2 (12.5)</td>
<td></td>
</tr>
<tr>
<td>How many days Exercise</td>
<td>1.9</td>
<td></td>
</tr>
<tr>
<td>%(%N)</td>
<td></td>
<td>%(%N)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>42.9(6)</td>
<td>28.6(4)</td>
</tr>
<tr>
<td>Male</td>
<td>57.1(8)</td>
<td>71.4(10)</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>57.1(8)</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>35.7(5)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>7.2(1)</td>
<td>100(14)</td>
</tr>
<tr>
<td>ATM Use</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Once a day</td>
<td>0</td>
<td>100(14)</td>
</tr>
<tr>
<td>Once a week</td>
<td>21.4(3)</td>
<td></td>
</tr>
<tr>
<td>Once a month or less</td>
<td>71.4(10)</td>
<td>Excellent</td>
</tr>
<tr>
<td>Never</td>
<td>7.2(1)</td>
<td>Good</td>
</tr>
<tr>
<td>Internet Proficiency</td>
<td></td>
<td>None</td>
</tr>
<tr>
<td>Excellent</td>
<td>78.6(11)</td>
<td>64.2(9)</td>
</tr>
<tr>
<td>Good</td>
<td>21.4(3)</td>
<td>7.2(1)</td>
</tr>
<tr>
<td>Very limited</td>
<td>0</td>
<td>Several times</td>
</tr>
<tr>
<td>Native English Speaker</td>
<td></td>
<td>Frequent</td>
</tr>
<tr>
<td>Yes</td>
<td>28.6(4)</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>71.4(10)</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Post-task survey

<table>
<thead>
<tr>
<th>Questions asked after each task</th>
<th>Score Range</th>
<th>Sub Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>How difficult or easy was it to complete this task?</td>
<td>1, “Very Difficult,” to 5, “Very Easy.”</td>
<td>Task X.1</td>
</tr>
</tbody>
</table>
How satisfied are you with using this application/system to complete this task?  

Task X.2

How would you rate the amount of time it took to complete this task?  
1, “Too Much Time,” to 5, “Very Little Time.”

Task X.3

<table>
<thead>
<tr>
<th>Table 4. Results of patient testing of the Alexa-based home automated tele-management system with screen or with only audio</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Screen</strong></td>
</tr>
<tr>
<td><strong>Mean</strong></td>
</tr>
<tr>
<td>Task 1. Review Your Exercise Plan</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Task 2. Review Instruction Video for Stand-to-Sit Exercise</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
### Task 3. Perform Exercise Stand-to-sit

| 1. How difficult or easy was it to complete this task? | 4.857 | 0.36 | 4.786 | 0.43 | 0.214 | 0.082 |
| 2. How satisfied are you with using this application/system to complete this task? | 4.714 | 0.61 | 4.643 | 0.63 | 0.286 | 0.040 |
| 3. How would you rate the amount of time it took to complete this task? | 4.357 | 1.15 | 4.643 | 0.93 | -0.286 | 0.040 |

**Total Task 3 Score**

| 14.286 | 1.49 | 13.714 | 1.90 | 0.571 | 0.041 |

**Total Task Score (Task1-Task3)**

| 42.429 | 4.75 | 41.286 | 5.93 | 1.143 | 0.036 |

### Exit Survey

| The Alexa exercise program is appealing | 4.643 | 0.84 | 4.439 | 0.94 | 0.214 | 0.082 |
| The Alexa exercise program is easy to navigate: | 4.571 | 0.94 | 4.286 | 1.07 | 0.286 | 0.302 |

**Total Exit Survey Score**

| 9.214 | 1.72 | 8.714 | 1.94 | 0.500 | 0.151 |
Table 5. System Usability Comparison Survey

<table>
<thead>
<tr>
<th>Questions asked after each task</th>
<th>Score Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1. I prefer the Alexa with video screen more than the pure voice</td>
<td>1. “Strongly Disagree,” to 5, “Strongly</td>
</tr>
<tr>
<td>screen more than the pure voice system</td>
<td>Agree”</td>
</tr>
<tr>
<td>Q2. The Video introduction guides me more clearly than the Audio</td>
<td>1. “Strongly Disagree,” to 5, “Strongly</td>
</tr>
<tr>
<td>introduction</td>
<td>Agree”</td>
</tr>
<tr>
<td>Q3. The exercise video instructions are better than the exercise</td>
<td>1. “Strongly Disagree,” to 5, “Strongly</td>
</tr>
<tr>
<td>audio explanation</td>
<td>Agree”</td>
</tr>
<tr>
<td>Q4. The video screen system is easier to follow than the pure voice</td>
<td>1. “Strongly Disagree,” to 5, “Strongly</td>
</tr>
<tr>
<td>system.</td>
<td>Agree”</td>
</tr>
</tbody>
</table>

Table 6. System Usability Comparison and System Usability Scale of with Screen System†

<table>
<thead>
<tr>
<th>Items</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>4.7</td>
<td>0.7</td>
</tr>
<tr>
<td>Q2</td>
<td>4.3</td>
<td>1.0</td>
</tr>
<tr>
<td>Q3</td>
<td>4.1</td>
<td>1.0</td>
</tr>
<tr>
<td>Q4</td>
<td>4.4</td>
<td>1.0</td>
</tr>
</tbody>
</table>

† 1: strongly disagree – 5: strongly agree

Discussion

The Alexa-based remote rehabilitation system's usability inspection with 14 participants demonstrated its high acceptance. Comparison of voice and video interfaces demonstrated statistically significant better performance of the video interface. Our results are congruent with previous reports demonstrating the significant potential of patient-centered digital health [5] tailored to patient preferences. Previous studies demonstrated potential utility of Alexa-based voice assistants as voice Interface technology in patients with heart failure [6] and in geriatric care [7]. Our findings establish an evidence-based usability framework for further expansion of Alexa-based devices into telerehabilitation domain by demonstrating utility of combined voice and video interface. Home-based telerehabilitation systems with user-friendly interfaces implemented with direct user input have been shown well accepted in patients with COPD [8], post-acute hip fracture recovery [9], geriatric syndromes [10], multiple sclerosis [11] and cancer [12]. Future steps would include developing Alexa-based user-friendly interfaces for these patients which include interactive video components.

Conclusion

This is the first study systematically comparing two major Alexa interfaces in a telerehabilitation system. These results are instrumental for future development of Alexa-based telerehabilitation systems.

References


Automatic Cohort Determination from Twitter for HIV Prevention amongst Black and Hispanic Men

Davy Weissenbacher, PhD\textsuperscript{1}, J. Ivan Flores, B.S.\textsuperscript{1}, Yunwen Wang, M.A., M.S.\textsuperscript{2}, Karen O’Connor, M.S.\textsuperscript{1}, Siddharth Rawal, B.S.\textsuperscript{1}, Robin Stevens, PhD, MPH\textsuperscript{2}, Graciela Gonzalez-Hernandez, PhD\textsuperscript{1}

\textsuperscript{1}University of Pennsylvania, Philadelphia, Pennsylvania, USA; \textsuperscript{2}University of Southern California, Los Angeles, California, USA

Abstract

Recruiting people from diverse backgrounds to participate in health research requires intentional and culture-driven strategic efforts. In this study, we utilize publicly available Twitter posts to identify targeted populations to recruit for our HIV prevention study. Natural language processing and machine learning classification methods were used to find self-declarations of ethnicity, gender, age group, and sexually-explicit language. Using the official Twitter API we collected 47.4 million tweets posted over 8 months from two areas geo-centered around Los Angeles. Using available tools (Demographer and M3), we identified the age and race of 5,392 users as likely young Black or Hispanic men living in Los Angeles. We then collected and analyzed their timelines to automatically find sex-related tweets, yielding 2,166 users. Despite a limited precision, our results suggest that it is possible to automatically identify users based on their demographic attributes and Twitter language characteristics for enrollment into epidemiological studies.

1 Introduction

Recruiting people from diverse backgrounds to participate in health research requires intentional and culture-driven strategic efforts. New media and technology provide novel opportunities to reach diverse populations. In recent decades, online networking sites have gained increased popularity as research recruitment tools, e.g. using Facebook to recruit young adults for inspecting their alcohol and drug use\textsuperscript{1} or young smokers for a cessation trial\textsuperscript{2} (see a review\textsuperscript{3}). While social media can in theory increase the reach of recruitment efforts and health promotion messaging, it does not ensure the equity of reach. Social media recruitment and interventions can reproduce long-standing health disparities if those efforts fail to target diverse groups based on their varying media use patterns and digital access.

Geolocation data has been used for public health surveillance and has been used in conjunction with temporal, textual, and network data to monitor e-cigarette use\textsuperscript{4}, opioid abuse\textsuperscript{5,6}, influenza\textsuperscript{7}, and HIV-related social media discussions\textsuperscript{8\textsuperscript{-}10}.

However, many of these applications remain largely descriptive of the social media content itself, rather than predictive or proactively applied to public health practice.

In consort with these geolocation approaches, we can utilize additional social media user data to identify target groups that meet specific demographic and behavioral characteristics for recruitment. A targeted social-media-based recruitment approach can identify communities in greatest need and can overcome the challenges of traditional recruitment approaches. Integrating the geographical location of social media users with content-based attributes, such as what we propose, provides a valuable tool for initial triage that can unobtrusively and inexpensively help identify potential study participants based on these detectable attributes, within the validated performance limitations of the automatic methods.

In this study, we utilize publicly available social media posts to identify predetermined populations to recruit for our HIV prevention study, targeting specific ethnic, gender, and age groups. The open user-generated content on Twitter was used to identify distinct subgroups of users who possibly meet the study's selection criteria and can be purposefully contacted. Compared to the undifferentiated social media postings that are likely confined by personal networks, our approach aims to better unleash the potential of large-scale social media data, to more efficiently find a larger number of the target population that is otherwise missed through traditional approaches. Specifically, this study aims to identify the likely members of ethnic (Black/Hispanic), gender (male), and age (18-24) groups in a particular geographic location (Los Angeles, California), whom we will recruit as the participants for an HIV prevention intervention. In this paper, we describe how we used geolocation techniques to bound the sampling frame to a digital sphere of people.
living in a specific urban space. We then used novel Natural Language Processing (NLP) approaches to identify users’ demographic characteristics and behaviors based on the content of their profiles and tweets.

The main contributions of our work are the collection of two corpora, (1) a corpus of 2,577 Twitter users profiles annotated with the demographic information available in their profiles, (2) a corpus of 3,500 tweets annotated with the use of sexually-explicit language and likely posted by young Black and Hispanic men living in Los Angeles, and (3) the release of a classifier trained to identify such language. Our annotation guidelines and the code of our classifier are available at https://bit.ly/33nF2p3.

2 Related Work

Health professionals are increasingly integrating social media in their daily activities, not only to disseminate new knowledge and extend their professional network but also to communicate with patients and collect health-related data. The possibility to identify and recruit participants for research studies using social media is one of its most appealing promises for health research. As trials continue to be canceled or discontinued due to poor quality recruitment through traditional methods, recruitment through social media appears to be a good complement, or a better alternative to reach underserved populations.

Although targeted advertising on Facebook has been used for recruiting participants, we opted for Twitter because it offers several advantages for our study. Twitter has a large base of active users, a free application programming interface (API) allowing for simple, real-time access to the data, and, due to its nature, it represents a default media for the users to publicly share and for researchers to have open access to their expressed ideas, activities, and beliefs; enabling selection via what they post and from where they post rather than only passively waiting for them to click on an ad. These makes Twitter a valuable recruitment tool that can complement other social media platforms. For example, Twitter helped to recruit cancer survivors, young cigarillo smokers, or users of direct-to-consumer genetic testing services.

Still, the most common method to recruit participants with Twitter is to use it as a broadcast platform. For example, researchers create a web page to describe their study and through their personal Twitter accounts, or an account dedicated to the study, advertise the study by frequently posting links to the web page. They may also ask colleagues, relevant organizations, or influential actors in their network to tweet/retweet the links to the study to reach more users. This approach relies on existing social networks and can lead to the exclusion of users in demographic groups outside of those research networks. As done in traditional recruitment through TV or local journals, researchers may also buy ads to promote their study among potentially eligible users as identified by the marketing tools of the platform. While this may be effective for some studies, this approach is passive: researchers advertise their studies and wait for interested individuals to contact them.

Few researchers have used the data posted by the users themselves to identify eligible users and directly contact them to join their study. In a seminal work, Krueger et al. identified transgender people communicating on Twitter by collecting tweets mentioning 13 transgender-related hashtags in order to examine their health and social needs. Hashtags allow for the collection of tweets of interest but they have strong limitations. Amongst these, we note: (a) 58% of the tweets collected in their study were irrelevant to the transgender community and had to be manually excluded, (b) the hashtags used by a community change quickly over time, and (c) all users of interest not using the hashtags were missed. The protocol published by Reuter et al. proposed a similar approach, but complemented the hashtags with keywords to detect patients with cancer on Twitter. The results of their study are, at the time of writing, under peer-review but they indicate the feasibility of this approach using a combination of manual curation and automated processes. In this study, we will follow an approach that builds on our prior work where we successfully combined keywords and machine learning to detect women publicly announcing their pregnancy.

3 Methods

We aimed to find on Twitter potential users to consent for inclusion in our study: young men (aged 18 to 24) who are Black and/or Hispanic and live in the metropolitan area of Los Angeles, California, USA and who publicly post sex-related content. We summarize our process in Figure 1. The following sections detail each step.
3.1 Geographical Locations and Demography Detection

To detect our target population, we first collect tweets posted in the Los Angeles (LA) area, using the official Twitter Search RESTful API. We defined two circular areas covering the city of LA. The center of the first circle was positioned at latitude 33.987816 and longitude -118.313296, and the center of the second circle at latitude 33.877130 and longitude -118.21004. Both circles had a radius of 16km. These circles roughly cover the areas of Compton and South-Central LA which are neighborhoods with predominantly Black and Hispanic populations. The free API facilitates collecting tweets matching the given geographic criteria posted up to seven days before the query. We collected samples of 200,000 tweets posted in these areas every 6 days between July 06, 2020, until February 26, 2021. After removing duplicate tweets, our initial collection included 47.4 million tweets.

We then applied two filters sequentially, based on results from the existing classifiers Demographer and M3, on the 47.4 million tweets initially collected, shown in Figure 1. Among competing tools, we chose Demographer and M3 based on their availability and performance. The first filter, which consists of rules on the output by Demographer, helped us remove tweets not posted by Black or Hispanic male users. Given a single tweet, Demographer searches for the given name of the user in the profile or the username. It distinguishes individuals from organizations by relying on neural networks to learn morphological features from the name and higher-level features from the user’s profile. When a user is identified as an individual, the gender and race or ethnicity of the person are inferred by the software. We removed 34.9 million tweets (73.5%) posted by users that were flagged as either institutional accounts or not meeting the gender and race/ethnicity inclusion criteria (Black or Hispanic men) from our initial collection, leaving a total of 12.6 million tweets to be processed by the second filter.

The second filter uses rules on the output of M3 to remove tweets posted by users younger or older than the target age bracket of 19-29 years old. Although our original study calls for users aged 18-24, 19-29 is the closest age bracket computed by the tool. M3 relies on a deep neural network to jointly estimate the age, gender, and individual/organization-status of the user from their profile images, usernames, screen names, and the short descriptions in their profiles. We removed 34.9 million tweets (73.5%) posted by users that were flagged as either institutional accounts or not meeting the gender and race/ethnicity inclusion criteria (Black or Hispanic men) from our initial collection, leaving a total of 12.6 million tweets to be processed by the second filter.

The second filter rules on the output of M3 to remove tweets posted by users younger or older than the target age bracket of 19-29 years old. Although our original study calls for users aged 18-24, 19-29 is the closest age bracket computed by the tool. M3 relies on a deep neural network to jointly estimate the age, gender, and individual/organization-status of the user from their profile images, usernames, screen names, and the short descriptions in their profiles. We removed 34.9 million tweets (73.5%) posted by users that were flagged as either institutional accounts or not meeting the gender and race/ethnicity inclusion criteria (Black or Hispanic men) from our initial collection, leaving a total of 12.6 million tweets to be processed by the second filter.

In order to validate the two initial filters (Demographer for gender and race/ethnicity prediction and M3 for age prediction), we randomly selected and annotated 2,577 users’ profiles (19%) from the 13,515 users. In agreement with best practices for deriving race/ethnicity information from Twitter datasets, three annotators (one graduate student and two undergraduates) that match the target demographic (young Black and/or Hispanic) were trained using the guidelines developed for the task. We provided annotators the profile information of users directly obtained through the Twitter API including the username, screen name, profile description, and listed location. We also allowed the
annotators to search the user’s Twitter profile directly, including other social media platforms or websites that the user provided links to in their Twitter profile, to help establish the information needed to validate the automatic prediction of user demographics.

For each of the 13,515 users identified by our pipeline, we collected up to 3,000 of their latest tweets, which is the maximum allowed through the Twitter API. This yielded 21 million tweets (4.35G) to be analyzed for our last aspect of the selection criteria: identify users who post tweets with sex-related language. The next section details this last classifier, which we developed specifically for this study.

3.2 Sex-related Language Detection

We developed an ensemble of classifiers to identify users posting tweets with sex-related language in our collection. Sex-related content included tweets about sexual intercourse, sexual desire, and sexual behavior. We excluded tweets focused solely on genitalia. We trained our ensemble with supervision, the most effective training approach, to learn a binary function: for a given tweet, our ensemble returns 1 if the content of the tweet is sex-related, 0 otherwise.

We used an existing corpus of 11,175 annotated tweets to pre-train our classifiers, as pre-training has often been found to improve the training phase. The corpus is described in a recent study. The authors randomly collected tweets from the 1% of publicly available tweets posted between January 1, 2016, and December 31, 2016. They isolated 6,949 tweets posted by young men living in the US by using existing tools and a list of keywords related to sex and HIV, such as HIV or condoms, among many. They manually annotated a subset of 3,376 tweets as pornographic (759, 22.5%), sex-related (1,780, 52.7%), and not sex-related tweets (837, 24.8%). In their study, they classified pornographic tweets based on a tweet format that included extensive use of hashtags and links to other websites. These hashtags were often of several sexually explicit words. They estimated the inter-annotator agreement to have an intraclass correlation of 91.2%. They later extended their corpus by selecting 7,799 additional tweets by following the same process and annotated the tweets as either “not pornographic but sex-related” or “not sex-related nor pornographic”. For our study, we used all of the annotated tweets, that is, 11,175 tweets (3,376 + 7,799 tweets).

We then fine-tuned and evaluated our classifier on a second, new, corpus sampled from the 21 million tweets posted by our population of interest, the 13,515 users identified by our pipeline. We selected a subset of candidates likely to be sex-related tweets by searching in the 21 million tweets mentions of the 53 keywords or emojis that have a sexual connotation from the list used in the previous study. This search returned 69,969 tweets posted by 10,491 users. We randomly sampled 3,500 (about 5%) of the tweets for annotation. Of these, 1,164 tweets were tagged as having sex-related language (but not pornographic) and 2,336 were tagged as not sex-related or pornographic. The inter-annotator agreement was substantial with 0.652 Cohen Kappa score. We found very few pornographic tweets among the 69,969 tweets because of the selection of our filters which identify individual users living in LA and filtered out organization and bot accounts, likely to be the main sources of pornographic tweets.

Our ensemble is composed of recurrent neural networks and transformers classifiers since these models are state-of-the-art and have been shown to encode efficiently the semantic content of sentences. We chose well-known architectures and available embeddings: 1) a Bidirectional Long Short-Term Memory (biLSTM) network which inputs are static word embeddings pre-trained on 2 billion tweets with GloVe; 2) a biLSTM which inputs are contextualized BERT embeddings (Bidirectional Encoder Representations from Transformers); 3) a single feedforward network which inputs are BERT embeddings; 4) another single feedforward network which inputs are RoBERTa embeddings (Robustly optimized BERT approach); 5) a biLSTM which inputs are RoBERTa embeddings.

Given the relatively small size of our corpora, we evaluated our ensemble of classifiers with 5-fold cross-validation, per accepted evaluation standards. For each fold, we pre-trained each classifier on the existing corpus of 11,175 tweets, we randomly split our corpus of 3,500 tweets into three sets, keeping 70% of the data for training, 10% for validation, and 20% for evaluation. During training, we fine-tuned the pre-trained transformer models with an AdamW optimizer, setting the learning rate to 4e-5 and a batch size of 8. For the biLSTM classifiers, we used default parameters of the Flair NLP library: a Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.1 and a batch size of 32. We trained the models during 5 epochs and evaluated them at each epoch on the validation set. We selected the models at the epochs where they achieved the highest accuracy on the validation set for testing. We used weighted
average voting to ensemble the decisions of the classifiers on the test set, with the weight of each classifier being their respective performance on the validation set. We evaluated our classifiers using the standard metrics of precision and recall for classification, as well as their harmonic mean, the F1-score.

3.3 Accounting for Prediction Errors in our Cohort Collection

Although we applied state-of-the-art methods to automatically predict the demographic features of Twitter users and the linguistic content of their posts, our predictions remain noisy. We estimated the real number of users detected by our pipeline by correcting our numbers using the precision of each component in our pipeline, calculated using our manual annotation of the 3,500 tweets sampled from our corpus. With these performance metrics (precision), we estimated how many users eligible for our study were actually amongst the 5,560 resulting after running our complete pipeline (see the last steps in Figure 1 and the Discussion section for details).

4 Results
4.1 Geographical Locations and Demography

We first assessed agreement on gender validation. Prior to calculating agreement, any user whose account was removed at the time of annotation for any one of the annotators was removed from the corpus. In total, 2,577 unique user profiles were validated by at least one annotator. All agreement measures were calculated using the IRR package in R (version 4.0.0). Agreement was calculated using Cohen’s kappa for two raters and Fleiss kappa for three raters. Table 1 shows the results of the agreement calculations.

<table>
<thead>
<tr>
<th>Annotator group</th>
<th>A+B</th>
<th>B+C</th>
<th>A+C</th>
<th>A+B+C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of users</td>
<td>720</td>
<td>382</td>
<td>382</td>
<td>382</td>
</tr>
<tr>
<td>Percent agreement</td>
<td>77.4</td>
<td>78.3</td>
<td>90.3</td>
<td>74.1</td>
</tr>
<tr>
<td>Cohen’s $\kappa$</td>
<td>0.568</td>
<td>0.575</td>
<td>0.776</td>
<td>—</td>
</tr>
<tr>
<td>Fleiss $\kappa$</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.618</td>
</tr>
</tbody>
</table>

Table 1: Inter-annotator agreement for gender determination.

Disagreements were resolved by taking the majority class for users that were validated by three annotators. In the event that there was no majority, and for disagreements on users validated by only two annotators, the annotation was adjudicated by the developer of the guidelines. Table 2 shows the results of validation after adjudication.

<table>
<thead>
<tr>
<th>Annotation set</th>
<th>Number of users</th>
<th>Account removed (%)</th>
<th>Female (%)</th>
<th>Not a personal account (%)</th>
<th>Unable to determine</th>
<th>Male (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triple</td>
<td>400</td>
<td>20 (5.0)</td>
<td>66 (16.5)</td>
<td>14 (3.5)</td>
<td>22 (5.5)</td>
<td>278 (69.5)</td>
</tr>
<tr>
<td>Double</td>
<td>347</td>
<td>22 (6.3)</td>
<td>73 (21.0)</td>
<td>11 (3.2)</td>
<td>26 (7.5)</td>
<td>215 (62.0)</td>
</tr>
<tr>
<td>Single</td>
<td>1830</td>
<td>44 (2.4)</td>
<td>349 (19.1)</td>
<td>14 (0.7)</td>
<td>117 (6.4)</td>
<td>1306 (71.4)</td>
</tr>
<tr>
<td>Total</td>
<td>2577</td>
<td>86 (3.3)</td>
<td>488 (18.9)</td>
<td>39 (1.5)</td>
<td>165 (6.4)</td>
<td>1799 (69.8)</td>
</tr>
</tbody>
</table>

Table 2: Results of gender validation.

Agreement for race/ethnicity annotations was then calculated for the 1,799 users validated as male. Before calculating agreement, all annotations were normalized to three groups, ‘y’ if the annotator identified the user as Black and/or Hispanic, ‘n’ for any other race/ethnicity identified, and ‘u’ for those the annotator was unsure. Table 3 shows the agreement measures for race/ethnicity.

<table>
<thead>
<tr>
<th>Annotator group</th>
<th>A+B</th>
<th>B+C</th>
<th>A+C</th>
<th>A+B+C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of users*</td>
<td>441</td>
<td>241</td>
<td>274</td>
<td>240</td>
</tr>
<tr>
<td>Percent agreement</td>
<td>76.9</td>
<td>71.8</td>
<td>71.5</td>
<td>58.8</td>
</tr>
<tr>
<td>Cohen’s $\kappa$</td>
<td>0.543</td>
<td>0.459</td>
<td>0.545</td>
<td>—</td>
</tr>
<tr>
<td>Fleiss $\kappa$</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.449</td>
</tr>
</tbody>
</table>

Table 3: Inter-annotator agreement for race/ethnicity determination.

Note. * The number of users differs for each group due to missing annotations, based on the guidelines if an annotator determined a user to be non-male, they did not annotate other demographic information.

Adjudication for race/ethnicity was performed in the same way as was done for gender. Table 4 shows the results after
adjudication. Note that the accounts removed were those not found when adjudicating the disagreements and those accounts were removed from the corpus.

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Number of male users</th>
<th>Black and/or Hispanic (%)</th>
<th>Not Black and/or Hispanic (%)</th>
<th>Unable to determine (%)</th>
<th>Account removed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triple</td>
<td>278</td>
<td>166 (59.7)</td>
<td>55 (19.8)</td>
<td>56 (20.1)</td>
<td>1</td>
</tr>
<tr>
<td>Double</td>
<td>215</td>
<td>133 (61.7)</td>
<td>61 (28.4)</td>
<td>18 (8.4)</td>
<td>3</td>
</tr>
<tr>
<td>Single</td>
<td>1306</td>
<td>694 (53.1)</td>
<td>391 (29.9)</td>
<td>221 (16.9)</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>1799</td>
<td>993 (55.2)</td>
<td>507 (28.2)</td>
<td>295 (16.4)</td>
<td>4</td>
</tr>
</tbody>
</table>

**Table 4**: Results of race/ethnicity validation.

The annotators derived location exclusively from the location field provided in the profile, therefore the agreement was very high (95.8%) for location annotations. Of the 993 users, 982 (98.9%) had information in the location field that the annotators used for validation. After adjudication and normalization of the names of neighborhoods and cities in Los Angeles County, there were 26 locations annotated as outside LA county, 8 were due to the user entering a non-place name in the location field, 17 were locations outside the target area, and 1 was a missed annotation.

Our annotators only used the descriptions in the user profiles or information provided on other linked sites to validate their age; however, age information was sparse. Of the 993 users validated as male and Black or Hispanic, only 67 (6.7%) had an age annotated by one of the annotators; 62 (92.5%) of the 67 users were in the 19-29 year-old age range. We are currently validating the age by crossing multiple dimensions of the users’ profiles: searching for self-declaration of ages in the timelines, looking at the photos posted by the users, and searching for the ages of other members of the networks the users frequently interact with.

### 4.2 Detecting sex-related tweets

We report the performance of the ensemble as well as the performance of each classifier composing the ensemble in Table 5. With an average of 0.7814 F1-score during the 5-fold cross-validation ($SD = 0.019$), the ensemble achieved the best performance. We checked if the disagreements between the predictions of the ensemble and those of the best performing individual classifier, Roberta + biLSTM, were statistically significant. For our computation, we used the predictions of the fold where their F1-scores were the closest. With a McNemar test, we found the proportions of disagreements between the two classifiers too close to rejecting the null hypothesis (with $\alpha=0.1$). We also evaluated the performance of our ensemble when trained and evaluated only on the 3500 tweets of our corpus, that is, when none of the classifiers composing the ensemble were pre-trained on the existing corpus of 11,175 annotated tweets. With an average of 0.7596 F1-score during the 5-fold cross-validation, the ensemble without pre-training performed slightly worse than the ensemble trained on all available tweets. We compared the predictions of both ensembles during the fold where their F1-scores were the closest with a McNemar test. Whereas the ensemble trained with all examples achieved better performance than the other ensemble during all folds, we found the proportions of disagreements between the two ensembles too close to rejecting the null hypothesis (with $\alpha=0.05$), showing a marginal improvement of the pre-training for our task.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>$P$</th>
<th>$R$</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT + FF</td>
<td>0.7138</td>
<td>0.6984</td>
<td>0.7044</td>
</tr>
<tr>
<td>RoBERTa + FF</td>
<td>0.7221</td>
<td>0.8003</td>
<td>0.7586</td>
</tr>
<tr>
<td>GloVe + biLSTM</td>
<td>0.6661</td>
<td>0.6940</td>
<td>0.6772</td>
</tr>
<tr>
<td>BERT + biLSTM</td>
<td>0.7111</td>
<td><strong>0.8170</strong></td>
<td>0.7601</td>
</tr>
<tr>
<td>RoBERTa + biLSTM</td>
<td>0.7283</td>
<td>0.8012</td>
<td>0.7619</td>
</tr>
<tr>
<td>Ensemble (pre-trained)</td>
<td>0.7810</td>
<td>0.7825</td>
<td><strong>0.7814</strong></td>
</tr>
<tr>
<td>Ensemble (no pre-training)</td>
<td><strong>0.7908</strong></td>
<td>0.7322</td>
<td>0.7596</td>
</tr>
</tbody>
</table>

**Table 5**: Performance of classification of non-pornographic sex-related tweets.

We analyzed the false positive predictions of our ensemble during the k-fold where it achieved its best precision score of 0.8206, and 0.7821 Recall, 0.8009 F1-score on our test set. The classifier made 40 false positives (FP) predictions. Most FPs, 25% (10/40), were tweets not talking about sex but the keywords or phrases occurring in the tweets could have sexual meanings in other contexts. Closely followed tweets with insults or swearing 22.5% (9/40) or jokes
15% (6/40) using sexual references. These tweets have explicit phrases but their main messages were not sexually related. Also, 12.5% (5/40) were tweets with sexual concepts used as elements of comparison in metaphors; 7.5% (3/40) were tweets posted by users speaking about fictional characters, masturbation, or describing their genitals, which our guidelines excluded. We found that only 7.5% (3/40) tweets were related to pornography and incorrectly labeled by our ensemble, including two tweets posted by sex workers self-promoting their content, and a third tweet as an advertisement written in a descriptive style very close to the style used by users commenting on pornographic materials. In the end, 7.5% (3/40) were deemed annotation errors by the final adjudicator. The decisions computed by neural networks are difficult to explain, as a result, we were unable to identify the reason for the misclassification of the last remaining tweet by our model.

We also analyzed the 51 false negatives (FN) predictions of the classifier. Most FNs, 35.2% (18/51) were tweets where sexual statements occurred in long sentences discussing a topic not sexually related; 23.5% (12/51) tweets were sexual meanings suggested often as jokes; 9.8% (5/51) tweets were tweets where a sexual keyword occurred but the tweets were very short - less than five words - providing limited context for the ensemble to determine their meaning. Finally, 7.8% (4/51) were tweets written with phonetic spellings or not written in English. We were unable to identify the reasons for the misclassification of the remaining 23.5% (12/51) tweets.

### 4.3 Detecting users posting sex-related tweets

Taking all 21 million tweets by the 13,515 users identified by our pipeline as Black or Hispanic men that are 19-29 years old and live in the Los Angeles area, and matching our list of keywords related to sex in their timeline and classifying the tweets detected with our ensemble, there were 5,560 users at the end of our entire pipeline. We observed that users are likely to post several sex-related tweets, with 58% (3211/5560) posting more than 3 such tweets. Whereas our ensemble did not detect perfectly single sex-related tweets, the fact that users post sex-related tweets more than once allows us greater precision when estimating the likelihood of a user having posted such Tweets. We grouped the 5,560 users according to the total number of sex-related tweets posted as predicted by the ensemble. We modeled the detection of users posting sex-related tweets as a binomial experiment, where a success is the ensemble correctly identifying a sex-related tweet. Since the precision of our ensemble was estimated to be 0.781, out of the 453 users having posted only 1 tweet predicted as sex-related we can assume that the ensemble correctly detected 353 users (453-(453*0.219)). Out of the 503 users having posted exactly 2 tweets predicted as sex-related, we estimated that the ensemble correctly detected 478 users (503-(503*0.048)) since the probability of our ensemble to be wrong on both tweets predicted as sex-related is 0.048 (with B(n=2, p=0.781) and k=0, where n is the number of predictions of the ensemble, p its precision and k the number of time the ensemble was correct). Following similar reasoning, we estimate that we have correctly detected 448 out of 453 users who posted exactly 3 tweets. We assume that the remaining 4,151 users who posted 4 or more tweets predicted as sex-related by our ensemble were correctly detected since the probability of k=0 was lower than 0.0025. In total, we estimate that among the 5,560 users, our ensemble was able to correctly identify 5,430 users posting sex-related tweets, indicating a high score of 0.977 precision.

### Discussion

From a sample of 2,577 users’ timelines that we manually inspected for validation, we were able to identify 993 Black or Hispanic men most likely to live in Los Angeles; the validation of the age of the users is still ongoing. With a limited score of 39.9% precision (993/(2577 − 90)), automatic tools for inferring Twitter users’ demographic information are still in their infancy and require improvement. Yet, despite the imprecision and sparsity in the demographic attributes of Twitter users, we were able, using these tools, to automatically identify around 5,392 (13,515*0.399) users who are really from the population of interest. Given we will reach out to the 13,515 users for consent and enrollment, any errors in the automatically derived demographic information will be evident and the users excluded if needed when verifying their eligibility for our HIV prevention study. In future work, we will improve the detection of errors when determining the gender and ethnicity attributes. In this study, we only relied on the Demographer to compute these attributes, whereas both Demographer and M3 distinguish individuals from organizations and, for individuals, deduce their gender. We will integrate both tools in our pipeline and submit for manual verification the users where the tools disagree. We will also search for self-declarations of these attributes in the timelines with regular expressions, a dimension that has not been exploited by the previous tools.
Our ensemble achieved a moderate performance when detecting sex-related tweets with 0.781 F1-score, however, since users are most likely posting multiple sex-related tweets, we found its precision to be high when detecting users posting multiple sex-related tweets with a score of 0.977. Among the 13,515 users of interest our ensemble selected 5,430 users posting sex-related tweets. Accounting for the errors in the predictions of the demographic of the users, we assume that among these 5,430 users, 2,166 (5430*0.399) are actually Black or Hispanic men in the Los Angeles Area, given only 39.9% of the users were identified as such in the set of 2,577 user profiles manually inspected.

In this study, we designed our pipeline to be precise because we have limited resources to contact users and validate their demography. A few changes can be made in our approach to increase the size of our cohort (semi-)automatically. First, we could replace the Twitter’s streaming API with the Twitter’s firehose to collect tweets posted during a given period. Second, we could extend semi-automatically the list of sexually connoted keywords by looking for and including in our list all sex-related phrases frequently occurring in the tweets predicted by our ensemble. Last, we did not store a large number of tweets from our initial collection of 47.4 million tweets based on the assumption that the users were not Black or Hispanic users. However, we found in our results that the ethnicity was imprecisely inferred automatically with a low 55.3% precision score. In future research, we may remove the ethnicity filter from the pipeline and solely rely on the geolocations of the tweets, given they originate within two neighborhoods with predominantly Black and Hispanic populations, which may largely increase the number of users detected.

We acknowledge the limited applicability of the approach to populations with non-binary gender identities. The concept of “gender” encompasses biological, psychological, social, and cultural factors. Our approach, as primarily estimating what appears to be the Twitter user’s biological sex (female/male), presents limitations in capturing the nuanced social construction and self-identification of gender. During manual verification of gender prediction, we encountered users that are non-binary or present themselves as women despite the classifier’s prediction of being men. In future research, we will explore ways to better include sexual and gender minority populations.

5 Conclusion

In this study, we aimed to automatically identify young Black or Hispanic men living in Los Angeles who use sex-related language publicly on Twitter. It was important to narrow the target cohort as much as possible as we intend to recruit them into the HIV prevention study that is the focus of our grant (NIDAR21DA049572). We used the official Twitter API and available tools, Demographer and M3, to extract the age, gender, and ethnicity of users tweeting from Los Angeles. In addition, we developed an ensemble of neural networks to detect sex-related language. Our ensemble achieved a 0.7814 F1-score on our test set when identifying individual tweets. However, because users are more likely to post multiple sex-related tweets, the performance is much higher in practice. After accounting for the prediction errors, we identified 2,166 users who post sex-related content on Twitter, and are likely Black or Hispanic men living in Los Angeles from an initial set of 47.4 million tweets collected over 8 months. Despite the imprecision of the current automatic detection methods, our results suggest that it is possible to find men in the target age bracket living in the right location (Los Angeles) and characterize their language on Twitter to a degree that makes it feasible to then contact them for enrollment into epidemiological studies without reaching out to too many users that do not fit the demographic profile. Contacting them will allow us to validate the automatic methods, and complete the missing demographic information, if any.

The classifier has the potential to identify significant numbers of Black and Hispanic men at high HIV risk inclusive of all sexual identities. We are able to systematically analyze large quantities of social media data from men, a volume of data that would be infeasible to manually review. This approach is a significant departure from current online recruitment procedures for HIV prevention and care efforts, which are often limited by the under recruitment of men of color and men who have sex with women.

Funding

This work was supported by National Institute on Drug Abuse (NIDA) grant number R21DA049572-01 to SR. The content is solely the responsibility of the authors and does not necessarily represent the official view of NIDA.
References


Predicting Avoidable Emergency Department Visits Using the NHAMCS Dataset

Yuyang Yang, BS¹, Jingzhi Yu, BS², Songzi Liu, BA, MS³, Hanyin Wang, BS, MS⁴, Scott Dresden, MD², Yuan Luo, PhD⁶
¹-⁶Northwestern University, Chicago, IL, USA

Abstract

Despite the important role avoidable emergency department (ED) visits play in healthcare costs and quality of care, there has been little work in development of predictive models to identify patients likely to present with an avoidable ED visit. We use a conservative definition of ‘avoidable’ ED visits defined as visits that do not require diagnostic or screening services, procedures, or medications, and were discharged home to classify visits as avoidable. Models trained using data from emergency departments across the US yielded a training AUC of 0.723 and a testing AUC of 0.703. Models trained using the full dataset were tested on demographic groups (race, gender, insurance status), finding comparable performance between white/black patients and male/female with reductions in performance in Hispanic populations and patients with Medicaid. Predictors strongly associated with non-avoidable ED visits included increased age, increasing number of total chronic diseases, and general as well as digestive symptoms. Reasons for visit stated as injuries and psychiatric symptoms influenced the model to predict an avoidable visit.

Introduction

Avoidable emergency department (ED) visits put significant strain on the healthcare system by increasing overall cost and leading to ED overcrowding. This problem has received more attention in the US in recent years due to significant increases in annual ED visits and reductions in ED admission rates among the elderly, the Medicare-reimbursed, and patients with multiple comorbidities. Previous studies suggest that diverting ED visits for nonurgent conditions that are treatable at retail clinics or urgent care facilities may lead to a projected saving of $4.4 billion annually. Avoidable ED visits are also believed to compromise quality of care by contributing to excessive testing and treatment as well as compromising the longitudinal relationship between the patients and the primary care physicians. Yet despite the heightened interest in reducing potentially avoidable ED visits, only a handful of studies had explored the factors associated with avoidable ED visits such as reasons for visit, demographics features, and encounter characteristics. In one national study of ED visits and utilization, non-urgent ED visits were shown to be more prevalent in the older, non-Hispanic white, and Medicare-insured patients. Another review showed that, based on limited evidence, younger age, convenience of the ED compared to alternatives, and referral to the ED by a physician all contributed to driving up non-urgent ED use.

To our knowledge, none of the studies had developed advanced models to classify and predict avoidable ED visits, the insights from which could be used to inform more efficient and cost-effective care management. Thus, the goal of this study is to use several machine learning algorithms to predict avoidable ED visits based on patient and visit characteristics known before the point of triage using a conservative definition of avoidable ED visit.

Methods

Study Population

The emergency department visits data used for this study comes from the Center for Disease Control (CDC) National Hospital Ambulatory Medical Care Survey (NHAMCS), which is an annual survey tool to collect the utilization of ambulatory services in hospital emergency and outpatient departments. NHAMCS surveys a sample of non-institutional and general hospitals across the United States. For this study, data from the NHAMCS survey collected between the years 2014 to 2018 were included in our sample.

Variables of Interest

From the NHAMCS dataset, we included the following variables of interest: age, gender, race and ethnicity, arrival time, whether the patient was seen at the ED with the past 72 hours, alcohol abuse, substance abuse, comorbidities, including Alzheimer’s, Asthma, Cancer, Chronic Kidney Disease, Chronic Obstructive Pulmonary Disease, Chronic
Heart Failure, Coronary Arterial Disease, Depression, Diabetes, End Stage Renal Disease, HIV, Hyperlipidemia, Hypertension, Obstructive Sleep Apnea, Osteoporosis, and Obesity, the total number of chronic conditions, time of day, day in the week of visit, month of visit, and reasons for visit codes. The NHAMCS dataset also contains patient stated reasons for visit, broken down into broad categories of chief complaints by organ system. Reasons for visit, in the form of organ system categories were used as variables in this study. In addition, patient time of arrival, which is encoded as a 24-hour time was recoded into arrival at daytime (between 7 AM and 5PM), night (5 PM and midnight), and overnight (midnight to 7AM).

**Data Preprocessing and Avoidable ED Definition**

Variables with more than 30% missing values (n=7 variables) and observations with missing data values (18.55%) were dropped. Training and testing datasets were created via a 7:3 split. Subsets of the dataset were created by stratifying by race, gender, and insurance type.

The avoidable ED visit definition of this study is based on a prior avoidable ED visit study by Hsia et al that uses NHAMCS data from earlier years. An avoidable ED visit is defined as discharged ED visits not requiring any diagnostic tests, procedures or medications. Patients admitted for observation, hospitalized, transferred, died in the ED or were dead on arrival were considered to be non-avoidable ED visits and excluded (Figure 1).

![Figure 1. Flowchart for defining avoidable ED visit and removing data with missingness or errors](image)

**Modeling Methods & Statistical Analysis**

To determine the best performing method for the classification task, multiple machine learning methods were experimented, including logistic regression, random forest, gradient boosted tree (XGboost), and multi-layer perceptron (MLP). Random forest and XGboost models were tuned to improve accuracy. After training 4 models with each machine learning method, we compared the area under curve (AUC) of each model. Modeling was performed using python 3.8.

To understand the performance of our model in various key demographic groups, we tested our full model on the demographic-specific testing sets. In addition, we created demographic-specific models, by retraining on demographic-specific training sets, and tested those models on the corresponding demographic-specific testing sets.

In addition, we examined the importance of variables used in the XGboost model. We obtained the Shapley Additive Explanations (SHAP) values from each model using the shap python package. Individual SHAP plots for the main model is shown in Figure 2 and SHAP plots for demographic subgroups are included as supplementary figures.

**Results**
There are a total of 77,714 individuals in our sample population. There are more females (55%) than males in the data and the mean age is 37.06. 57% of the population is white, 23% is black, and 16% is Hispanic. Private insurance makes up 18% of the population, with 27% of patients using Medicare and 35% using Medicaid. In comparison to the white and black population, the Hispanic population was significantly younger, had decreased usage of Medicare and increased usage of Medicaid (Table 1).

### Table 1. Baseline characteristics of the sample population

<table>
<thead>
<tr>
<th>Data</th>
<th>Visits</th>
<th>Non-ED%</th>
<th>Age (mean)</th>
<th># Comorbidities (mean)</th>
<th>Race (White %)</th>
<th>Gender (Female %)</th>
<th>Insurance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>77714</td>
<td>4.44</td>
<td>37.06</td>
<td>1.05</td>
<td>57%</td>
<td>55%</td>
<td>34.6 (Medicaid) 27.1 (Medicare) 17.8 (Private Insurance)</td>
</tr>
<tr>
<td>White</td>
<td>44402</td>
<td>3.75</td>
<td>41.26</td>
<td>1.18</td>
<td>-</td>
<td>55.13%</td>
<td>26.34 (Medicaid) 23.24 (Medicare) 32 (Private Insurance)</td>
</tr>
<tr>
<td>Black</td>
<td>17688</td>
<td>4.85</td>
<td>32.52</td>
<td>1.01</td>
<td>-</td>
<td>55.7%</td>
<td>43.8 (Medicaid) 12.2 (Medicare) 20 (Private Insurance)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>12689</td>
<td>4.89</td>
<td>29.00</td>
<td>0.69</td>
<td>-</td>
<td>54.5%</td>
<td>50.6 (Medicaid) 7.67 (Medicare) 19.0 (Private Insurance)</td>
</tr>
</tbody>
</table>

Table 2. Accuracy metrics for best machine learning model experimented in this study

<table>
<thead>
<tr>
<th>Model</th>
<th>Training AUC</th>
<th>Testing AUC</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGBoost</td>
<td>0.723</td>
<td>0.703</td>
<td>0.975</td>
<td>0.635</td>
<td>0.769</td>
</tr>
</tbody>
</table>

We tested four types of models XGBoost, Random Forest, Logistic Regression, and MLP. Out of these models XGBoost was found to have the best testing AUC (Table 2). The training AUC of our full model is 0.723 and the testing AUC is 0.703. The AUC of the full model tested on demographic-specific groups is shown in Table 3. The AUC of the full model is comparable between the female-only (AUC = 0.706) and male-only (AUC = 0.700) testing sets, as well as between white (AUC = .6984) and black patients (AUC = .7006) testing sets. There is a decrease in model performance among the Hispanic group (AUC = .6724). Within the insurance subgroups, the performance of...
the model is highest in the Medicare-only group (AUC = 0.7289) and lower in the private insurance-only group
(AUC = 0.685) and Medicaid-only group (AUC = 0.677).

Table 3a. Race Models

<table>
<thead>
<tr>
<th>Race, test set N</th>
<th>Training Set, N</th>
<th>Training AUC</th>
<th>Testing AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>White, 13345</td>
<td>Full</td>
<td>0.7234</td>
<td>0.6984</td>
</tr>
<tr>
<td>White-only, 31057</td>
<td></td>
<td>0.7282</td>
<td>0.7006</td>
</tr>
<tr>
<td>Black, 5295</td>
<td>Full</td>
<td>0.7234</td>
<td>0.7050</td>
</tr>
<tr>
<td>Black-only, 12393</td>
<td></td>
<td>0.7390</td>
<td>0.6899</td>
</tr>
<tr>
<td>Hispanic, 3777</td>
<td>Full</td>
<td>0.7234</td>
<td>0.6724</td>
</tr>
<tr>
<td>Hispanic-only, 8912</td>
<td></td>
<td>0.7634</td>
<td>0.6733</td>
</tr>
</tbody>
</table>

Table 3b. Gender Models

<table>
<thead>
<tr>
<th>Gender, test set N</th>
<th>Training Set, N</th>
<th>Training AUC</th>
<th>Testing AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female, 13119</td>
<td>Full</td>
<td>0.7234</td>
<td>0.7064</td>
</tr>
<tr>
<td>Female-only, 30535</td>
<td></td>
<td>0.7304</td>
<td>0.7064</td>
</tr>
<tr>
<td>Male, 10651</td>
<td>Full</td>
<td>0.7234</td>
<td>0.6995</td>
</tr>
<tr>
<td>Male-only, 24926</td>
<td></td>
<td>0.7193</td>
<td>0.6928</td>
</tr>
</tbody>
</table>

Table 3c. Insurance Models

<table>
<thead>
<tr>
<th>Insurance, test set N</th>
<th>Training Set, N</th>
<th>Training AUC</th>
<th>Testing AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private, 6509</td>
<td>Full</td>
<td>0.7234</td>
<td>0.6848</td>
</tr>
<tr>
<td>Private, 15109</td>
<td></td>
<td>0.7613</td>
<td>0.6742</td>
</tr>
<tr>
<td>Medicare, 4242</td>
<td>Full</td>
<td>0.7234</td>
<td>0.7289</td>
</tr>
<tr>
<td>Medicare, 9858</td>
<td></td>
<td>0.8163</td>
<td>0.7220</td>
</tr>
<tr>
<td>Medicaid, 8225</td>
<td>Full</td>
<td>0.7234</td>
<td>0.6770</td>
</tr>
<tr>
<td>Medicaid, 19082</td>
<td></td>
<td>0.7235</td>
<td>0.6617</td>
</tr>
</tbody>
</table>

We identified several variables to have strong influence on model prediction of non-avoidable or avoidable ED
visits. (Figure 2). Age, increasing number of total chronic diseases, and general as well as digestive symptoms all
strongly influenced the model to label a visit as avoidable. Reasons for visit stated as injuries, psychiatric symptoms,
and younger age strongly influenced the model to predict an avoidable visit. The time of day and day of week of the
visit were not considered to be particularly important by our model. We also noted several differences in the feature
importance of models trained on specific demographic subgroups. Genitourinary symptoms were strong predictors

517
of non-avoidable ED visits in women (Supplementary Figure 1a). The presence of Psychiatric symptoms was seen as important in the Medicaid population, but not in either Medicare or Private insurance (Supplementary Figure 3).

Discussion

In this study, we developed a classification algorithm using machine learning methods to predict avoidable ED visits prior to the patients arriving at the ED. Our results suggest a moderate ability to predict avoidable ED visits using variables known prior to the point of triage, with a max testing AUC just over 0.70. There was little difference in model performance between genders, and between black and white populations, suggesting that our model is not sensitive to racial differences between the two groups. Model performance was worse in Hispanic, as well as patients on Medicaid, suggesting that our model may be biased against these populations. Compared to the Black and White populations, the Hispanic population was significantly younger, used Medicaid more often, and had less comorbidities. Differences in the Hispanic population within the data in comparison to Black and White groups as well as confounding factors not represented in the data such as language or cultural barriers, which have been noted in machine learning work involving Hispanic groups, may explain reduced model performance in this group.

Some of the variables identified in figure 2 as strong predictors of avoidable/non-avoidable determination are consistent with existing knowledge surrounding avoidable ED visits. It is noted that the emergency department setting is not well equipped to treat mental health disorders and substance abuse, which may explain why psychiatric symptoms were associated with avoidable ED visits. Our finding that pediatric age is associated with increased likelihood of avoidable ED visits is consistent with previous findings, potentially due to a mismatch in perceived severity of a child’s symptoms between parents and healthcare providers. The presence of Medicare insurance was not considered an impactful metric, although PAYTYPE 1 (Private Insurance) was associated with increased avoidability and 5 (Self-Pay) and -8 (Unknown Payment) were associated with decreased avoidability. The time-of-day variable was not considered to be important in the model, suggesting that there is no difference in avoidable visits between overnight and daytime admissions. The day of week variable was also not considered to be important. It is known that patient access to primary care is an important factor in ED utilization. It has been shown that patients seek ED care during windows when primary care physicians are unavailable, such as overnight arrivals and weekend visits in single site studies. As our data comes from the national level, the impact of primary care availability that may be visible at a local level may be obscured.

Some features were more important in the race, gender, and insurance specific models compared to the base model. Genitourinary symptoms were important in females but not males, possibly due to an increased abundance of urinary tract infections and other gynecological symptoms that present more commonly in females. Arrival by ambulance was important in females but not males and digestive symptoms were more important in females than males. Women are
known to have higher rates of digestive primary complaints compared to men (and lower admissions, but same when adjusted for age and comorbidities)\(^{14}\). Pregnant women have been noted to use emergency department services non-urgently\(^{15}\).

Our work represents the first attempt to use machine learning in the classification of avoidable emergency visits. We identified several factors associated with avoidable and non-avoidable ED visits that may be useful for clinicians and patients in managing their care options. For example, in a primary care or community clinic setting, our model could be deployed to provide patient guidance on whether to schedule a meeting with their physician or go to the emergency department.

The addition of longitudinal care data from healthcare providers would likely improve the accuracy of our model in this setting. Future work should examine the effect of additional variables known prior to the point of triage to improve predictive performance, such as information regarding prior ED utilization.

Supplementary Figures

![Supplementary figure 1a. Female only model.](image-url)
Supplementary figure 1b. Male only model.

Supplementary figure 2a. Black only model.
Supplementary figure 2b. Hispanic only model.

Supplementary figure 2c. White only model.
Supplementary figure 3a. Private-Insurance only model.

Supplementary figure 3b. Medicare only model.

Supplementary figure 3c. Medicaid only model.
References

A Computational Framework for Identifying Age Risks in Drug-Adverse Event Pairs

Zhizhen Zhao1, Ruoqi Liu2,3, Lei Wang, PhD3, Lang Li, PhD3, Chi Song, PhD4, Ping Zhang, PhD, FAMIA2,3,*

1Department of Statistics, The Ohio State University, Columbus, Ohio, USA
2Computer Science and Engineering, The Ohio State University, Columbus, Ohio, USA
3Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio, USA
4Division of Biostatistics, The Ohio State University, Columbus, Ohio, USA
*Corresponding authors: zhang.10631@osu.edu

Abstract

The identification of associations between drugs and adverse drug events (ADEs) is crucial for drug safety surveillance. An increasing number of studies have revealed that children and seniors are susceptible to ADEs at the population level. However, the comprehensive explorations of age risks in drug-ADE pairs are still limited. The FDA Adverse Event Reporting System (FAERS) provides individual case reports, which can be used for quantifying different age risks. In this study, we developed a statistical computational framework to detect age group of patients who are susceptible to some ADEs after taking specific drugs. We adopted different Chi-squared tests and conducted disproportionality analysis to detect drug-ADE pairs with age differences. We analyzed 4,580,113 drug-ADE pairs in FAERS (2004 to 2018Q3) and identified 2,523 pairs with the highest age risk. Furthermore, we conducted a case study on statin-induced ADE in children and youth. The code and results are available at https://github.com/Zhizhen-Zhao/Age-Risk-Identification

Introduction

Adverse Drug Events (ADEs) are considered any noxious, unintended or undesired effect of a drug that occurs at a dose normally used in humans for prophylaxis, diagnosis or treatment1. ADEs are the 4th leading cause of death in the United States, which significantly increases the economic burden, stay length and death risk for hospitalized patients2. It is estimated that hospitalized patients have more than 2,216,000 ADEs in U.S., which cause more than 106,000 deaths annually3.

Age is an important risk factor to experience ADEs. Previous studies illustrated significant ADE differences among various age groups4. The 20-29 years old group has the lowest rate of adverse events, whereas the 0-9 years old group has the highest rate5,6. Young children have greater potential to report adverse events than adults, especially in the anti-infective, respiratory, dermatological and nervous system, because the detoxification mechanisms of children are immature7. Among children, a clinical trial showed that treating asthma through inhaled corticosteroids, younger children developed more cough and perioral dermatitis, while older children report hoarseness more frequently5. Furthermore, multiple studies have shown that the risk of an ADE increases with age5,8. For instance, with patent foramen ovale, older cryptogenic stroke patients have greater risks of having adverse events than younger patients5. Among patients with heart failure, risks of ADEs after taking digoxin enhance significantly with age, from 1.7% for patients below 50 to 5.4% for patients older than 80-year-old8. Although previous studies revealed certain ADEs with age risks, there still lacks comprehensive exploration, especially for ADEs associated with particular drugs.

To detect an association among interested drug-ADE pairs, data mining methods have been developed on measures of disproportionality, such as Reporting Odds Ratio (ROR) and the Proportional Reporting Ratio (PRR)9. However, many of them ignored heterogeneous characteristics of patients (e.g. age, gender and primary diseases) and gave the same weight to the information from the whole population when calculating the expected number of reports for a specific drug-ADE pair, which might mask the true signals or flag false associations as potential signals12.

The US Food and Drug Administration (FDA) provides post-marketing drug surveillance data Adverse Event Reporting System (FAERS)13, which collects data for suspected adverse drug events for further analysis. It covers a wide range of products aimed at diverse medical indications and are used across a broad range of patient populations. We aim to analyze the FAERS database and identify the age of patients who are more likely to experience certain ADEs.
after taking a specific drug. In precision medicine, a few studies adopted data-driven statistical subgroup methods
to estimate average treatment effects and explore subgroups with enhanced treatment effects\textsuperscript{10,11}. However, due to the
lack of a control group, popular subgroup methods could not be applied in the database directly. Current researchers
have used subgroup disproportionality to quantify the differential risk of a drug causing an ADE in men or women\textsuperscript{14}
and have predicted the probability of being a female given confounding factors and built balanced cohorts to dampen
the confounding biases existing in FAERS in the meantime\textsuperscript{2}. But existing work is designed for detecting gender dif-
f erences and examine ADEs in only two subpopulations (i.e., male and female). Their methods could not be directly
leveraged for identifying ADEs in multiple age groups (i.e., children, youth, adult and senior).

In this study, we proposed a computational framework to identify the age group of patients who tend to experience
more ADEs after taking a specific drug. Patients were divided into four age groups based on the World Population
Prospects from The United Nations Department of Economic and Social Affairs\textsuperscript{15}. Firstly, we performed Chi-squared
tests to identify drug-ADE combinations which showed significant age differences. Then, for each drug-ADE pair,
we conducted pairwise comparisons and measured disproportionality to detect the age group of patients with the
highest risk of experiencing an adverse event. Finally, we applied logistic regression and likelihood ratio test (LRT)
on the interaction between age group and drug to remove the confounding effects caused by age bias. We applied
our framework on submissions of FAERS from 2004 to the third quarter of 2018. We successfully discovered 2,523
age-related drug-ADE combinations and identified the highest risk age group in each combination. We also conducted
a case study on statin drugs induced ADE in children and youth.

Overall, our contributions can be summarized as follows:

• We developed a new computational framework to quantify the risks of experiencing ADEs for patients in mul-
tiple age subgroups. Our framework allows for the identification of heterogeneous characteristics of patients
that have multiple categories. Through this computational framework, we fully explored age risks in drug-ADE
pairs.

• To discover the highest risk among multiple age groups, we determined exactly which subgroups are signifi-
cantly different in reporting an ADE, then calculated the reporting odds ratios (RORs) and selected the subgroup
with the highest ROR as the significant signal.

• We applied our framework to discover age risks in drug-ADE pairs in FAERS data set. We also analyzed
statin-related ADEs that have the highest risk in children and youth.

Methodology

In our study, the age risks were detected mainly through four steps: discover age differences for specific drugs, discover
age differences for specific drug-ADE pairs, discover the age group with higher risk, remove confounding effect by
age bias. The process is presented in Algorithm 1. The illustration of our framework can be found in Figure 1.

Detect age differences for drugs

First, we detected age differences for a specific drug by considering all adverse drug events appearing with the drug.
The goal was to discover drugs that have an overall age difference in drug-ADE pairs frequency distribution. An
overall Chi-square test was applied to identify if there is an overall shift in drug-ADE pairs frequency distribution.
Because of the large number of drug-ADE pairs, this step could also help avoid testing all individual unique pairs and
then alleviate the testing burden. In addition, there were some drugs that only appeared in one particular age group,
which we are not interested in. Thus, we filtered out drugs that appear only in one group and conducted tests on drugs
that appear in more than two age groups.

For each drug, assume there were $k$ ADEs in total that appeared with the drug. We constructed one of the contingency
tables shown in Figure 1 based on how many age groups the drug appeared in. The value in each cell represents the
total count of reports for each drug and the corresponding $ADE_i$ ($i = 1, \ldots, k$) in each age group.

We applied a Chi-square test for the contingency table, where the null hypothesis is that there is no age difference
for a specific drug. For instance, for the drug occurring in all four groups, $H_0 : P_A^i = P_B^i = P_C^i = P_D^i$, for
Algorithm 1: The algorithmic framework to identify drug-ADE pairs with age risk

**Input:** patient data: drugs X, adverse events Y, age groups Z
**Output:** drug-ADE pairs, higher risk age group, significance

1. for drug $x_i$ in $X$ do
2. Group all ADEs, $y_{ij}$, associated with $x_i$;
3. if $x_i$ exists in at least two age groups then
4. Count the occurrence of $(x_i, y_{ij})$ for each age group;
5. Perform a $\chi^2$ test and compute adjusted $p$-value using Bonferroni adjustment;
6. if adjusted $p$ value $\leq 0.05$ then
7. for ADE $y_j$ in $y_i$ do
8. if $(x_i, y_j)$ exists in at least two age groups and occurrence $\geq 50$ then
9. Count the occurrence of $(x_i, y_j)$ and all other pairs associated with $x_i$ for each age group;
10. Perform a $\chi^2$ test and compute adjusted $p$-value using Bonferroni adjustment;
11. if adjusted $p$ value $\leq 0.05$ then
12. Perform $\chi^2$ tests for pairwise age groups and compute adjusted $p$ values using Bonferroni adjustment;
13. Compute and rank RORs;
14. if adjusted $p$ value $>0.05$ then
15. Combine age groups;
16. Identify the age group $z_i^j$ with the largest ROR;
17. Fit the logistic regression model;
18. Perform a Likelihood Ratio Test on $\beta_3$ and compute adjusted $p$ values using Bonferroni adjustment;
19. if adjusted $p$ value $\leq 0.001$ then
20. return $x_i$, $y_j$, $z_i^j$, adjusted $p$ value

all $i = 1, ..., k$, where $P_i^j$ denotes the probability of being in the age group $j$ and having the $ADE_i$, capital letters represent four different age groups. Bonferroni methods were used for all drugs to adjust P values to correct for multiple testing.

Detect age differences for drug-ADE pairs

The first step filtered out the drugs that illustrated an overall difference in drug-ADE pairs frequency distribution, but not necessarily indicated that all pairs related to those drugs showed age differences. Therefore, this step is to apply a Chi-square test to identify if there exist age differences in a particular drug and adverse drug event pair. We are still only interested in the pairs that occurred in more than two age groups, so filtered out pairs that occurred only in one age group. In addition, in order to ensure the validity of Chi-square tests, we filtered out pairs whose occurrence was less than 50.

For each unique drug-ADE pair, we constructed one of the contingency tables shown in Figure 1 based on how many age groups the pair appeared in. The first column represents the number of reports for the interested adverse drug event after taking the interested drug for each age group. The second column represents the number of reports for all adverse drug events that occurred with the interested drug except the interested one for each age group. Based on the contingency table, we tested the null hypothesis that there is no age difference for a specific drug-ADE pair. The P values were adjusted through Bonferroni methods for all testing pairs to correct for multiple testing.

Detect the age group with higher risk

The above two steps provided drug-ADE pairs that showed different distributions for various age groups, but they did not point out which groups accounted for the differences. We would then conduct pairwise comparisons and make
Figure 1: Illustration of the proposed methodology: In the first step, overall Chi-squared tests for each drug are performed to identify drugs with overall age differences. Then overall Chi-squared tests for each drug-ADE pair are performed to identify drug-ADE pairs with age differences. Next, Chi-squared tests for age group comparisons within each pair are performed. RORs for every two age groups are computed and ranked, which quantifies the age risks. At the end, a logistic model is built for each detected pair and the Likelihood Ratio Test is performed on the interaction of drug and age group to remove age bias.

Disproportionality analyses for subpopulations of drug-ADE pairs obtained from the last step. The objective is to identify one or more age groups that are different from other age groups and are more likely to experience the target adverse event for a specific drug-ADE pair.

Since the pairs might exist in two, three, or four age groups, we conducted at least one and at most six pairwise comparisons. Chi-square tests were conducted by constructing at least one and at most six 2x2 contingency tables for unique drug-ADE pairs. For instance, if a pair existed in all four age groups, then six 2x2 contingency tables shown in Figure 1 would be constructed. In order to ensure the validity of Chi-Square tests, all cells in all contingency tables are greater than five. Therefore, the drug-ADE pairs with occurrence more than zero, and less than five in at least one age group were filtered.

Meanwhile, adjusted ROR for each table was calculated. Adjusted ROR was defined as: $ROR = \frac{a/b}{c/d}$, where $a$ denotes the number of patients in the first age group experiencing the target ADE, $b$ denotes the number of patients in the first age group experiencing all other ADEs except the target ADE, $c$ denotes the number of patients in the second age group experiencing the target ADE, and $d$ denotes the number of patients in the second age group experiencing all other ADEs except the target ADE. A ROR greater than 1 indicated that the first age group is more likely to report the target adverse event.
ADE after taking the specific drug. In contrast, a ROR less than 1 indicated that the second age group is at a higher risk when experiencing the target ADE.

Risks for different age groups can be ordered through the value of ROR. Considering one group as a baseline and fix it, then RORs calculated in terms of that group can be ordered from the least value to the greatest value. On the basis of pairwise comparisons, we then combined the groups which are not significantly different. For groups that were not significantly different, to be more conservative, kept the ROR with the smallest value in the rank. Selected the age groups which have the highest risk as to the significant signals, meaning that certain age groups are susceptible to certain ADEs after taking a drug. For instance, all four group patients taking lamotrigine once experienced convulsion. There did not show a significant difference between children and youth but showed significant differences among the other five comparisons. Combining children group with youth group, and ordered the RORs, children and youth group illustrated the largest risk with the highest odd ratio. Thus we considered children and youth as higher risk groups, meaning that they tend to report more convulsion among the whole population.

In pairwise comparisons, we are only interested in pairs that showed consistent significantly different results within all comparisons. To be more specific, for a specific drug-ADE pair, if group A and group B are significantly different, group A and group C are significantly different, but group B and group C are not significantly different, we wouldn’t continue to analyze it. P values within each drug-ADE pair were adjusted through Bonferroni method for correction.

**Remove age bias in reporting ADEs**

It’s possible that ADEs selected above are caused by age, not necessarily the interested drug. For instance, older adults are more likely to report inadequate pain management than younger adults. The previous Chi-square test might still identify the pairs associated with certain outcomes with the age difference, but these pairs may be not drug-specific. So we would filter out the potential confounding effects in reporting ADEs due to age bias. In other words, we are interested in discovering the ADEs that are caused by the interaction effects of drug and age. Thus, we conducted a logistic regression for each drug-ADE pair identified in the last step and tested on the interaction term. We build such a model:

\[
\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 I(Drug) + \beta_2 Age + \beta_3 Age \times I(Drug)
\]

where Drug is an indicator variable, taking values on 1 if taking the target drug and 0 if not. Target ADE is 1 if experiencing the target ADE and 0 if experiencing other ADEs. \(\beta_0\) represents the log-odds of reporting the target ADE that the patient not taking the drug and in the first age group. \(\beta_1 + \beta_3\) represents the increase of log-odds if taking the drug than not. \(\beta_3\) is the parameter that we are interested in, which quantifies the drug-age interaction effects. Thus, we conducted a likelihood ratio test on \(\beta_3\). The significant results illustrated that the drug effects depend on the identified high-risk age group. Bonferroni adjustment for all candidate drug-ADE pairs was used for multiple testings.

**Experiment**

**Data Set**

In the source FAERS database, there exist multiple versions for an individual report, including one or more follow-up case versions based on the initial case version. In addition, the drug names in FAERS are not normalized, full names, trade names, abbreviations and spelling mistakes are existed instead. We cleaned and normalized FAERS by removing duplicate individual case reports, mapping drug names to RxNorm concepts and outcomes to SNOMED-CT concepts based on a standardization method. Finally, we mapped unique individual case reports with demographic information through ‘primaryid’ and ‘isr’.

In this study, we collected FAERS quarterly submissions from 2004 to the third quarter of 2018. The curated and standardized version consisted of 99,543,819 drug-ADE pairs. After removing missing values, and values out of 100 in age, a total of 75,748,043 pairs remained. We then divided patients into four age groups, 0-14 years old as children, 15-24 years old as youth, 25-64 years old as adult and >65 years old as senior on the basis of the criteria provided by
Results and Analysis

After cleaning and standardization, there remained 4,580,113 unique drug-ADE pairs with 5,661 drugs and 19,198 ADEs in the data set. We applied the methodology to the data set. We removed pairs whose ADE is caused by the misuse of drugs, such as off label use and drug abuse. In the process, by screening drug labels, we also removed some possible false positive pairs caused by co-morbidities. For instance, the atorvastatin-coronary artery disease pair in adult might be spurious since atorvastatin is commonly used in the hyperlipidemia patients for the prevention of heart disease, which is a critical comorbidity. Finally, with confidence level at 0.999, we discovered 2,523 unique age-associated drug-ADE pairs and their highest age risks, among which 408 drugs and 775 adverse events were included. Full results are available in https://github.com/Zhizhen-Zhao/Age-Risk-Identification. Table 1 presents the number of drug-ADE pairs, drugs and ADEs in each age risk group.

<table>
<thead>
<tr>
<th>Age Risk Group</th>
<th>Drug-ADE pairs</th>
<th>Drugs</th>
<th>Adverse Drug Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children</td>
<td>629</td>
<td>208</td>
<td>342</td>
</tr>
<tr>
<td>Youth</td>
<td>672</td>
<td>226</td>
<td>284</td>
</tr>
<tr>
<td>Adult</td>
<td>255</td>
<td>94</td>
<td>143</td>
</tr>
<tr>
<td>Senior</td>
<td>531</td>
<td>175</td>
<td>244</td>
</tr>
<tr>
<td>Children &amp; Youth</td>
<td>325</td>
<td>141</td>
<td>194</td>
</tr>
<tr>
<td>Children &amp; Adult</td>
<td>7</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Children &amp; Senior</td>
<td>22</td>
<td>19</td>
<td>21</td>
</tr>
<tr>
<td>Youth &amp; Adult</td>
<td>47</td>
<td>38</td>
<td>38</td>
</tr>
<tr>
<td>Youth &amp; Senior</td>
<td>15</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>Adult &amp; Senior</td>
<td>20</td>
<td>17</td>
<td>18</td>
</tr>
</tbody>
</table>

Among identified pairs, 1,966 (77.92%) pairs appeared in all age groups and 2,087 (82.72%) pairs have the single age risk group. Children and youth are the most susceptible groups among the whole population. Only 44 (1.74%) pairs show the highest risk in two age groups which are not continuous in classification. We identified the most prominent drug-ADE pairs that pose the highest risk to each single age group, the top 6 pairs are shown in Table 2. For the 10 most prescribed drugs in the U.S., we successfully identified atorvastatin, lisinopril, amlodipine, amoxicillin, omeprazole, losartan and metformin whose ADE pairs have significant age risk. Table 3 shows the most significant risks in a single age group from these drugs.

In addition, we compared the distribution of significant age risks in each system organ class (SOC). All detected drug-ADE pairs are grouped at the SOC level. All 27 SOCs show the significant age risks, where 'General disorders and administration site conditions', 'Injury, poisoning and procedural complications' and 'Vascular disorders' are the top 3 classes that have age risks. Figure 2 presents top 15 SOCs that have all four single age risks. Adult risk is at a lower proportion for all classes. Seniors are at the highest risk for experiencing Infections and infestations, Renal and urinary disorders, Immune system disorders, Musculoskeletal and connective tissue disorders. Children are susceptible to Gastrointestinal disorders, Metabolism and nutrition disorders, Nervous system disorders, Investigations.

Case Study of Statin Drugs Induced ADEs in Children and Youth

In detected drug-ADE pairs with age risk, we filtered statin-related pairs that have age risk in either children or youth shown in Table 4. Our results well complemented current researches on statin-induced ADEs. The results indicated that the number of atorvastatin-related pairs with risk in children and youth seems to be significantly larger than other statin drugs, which demonstrates the variation of ADEs between different statin drugs. We also observed that statin-
### Table 2: Top 6 drug-ADE pairs within each higher age risk group

<table>
<thead>
<tr>
<th>Drug</th>
<th>Adverse Drug Event</th>
<th>ROR</th>
<th>Adjusted P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Children</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>risperidone</td>
<td>emotional disorder</td>
<td>12.51</td>
<td>0</td>
</tr>
<tr>
<td>rivaroxaban</td>
<td>gastrointestinal haemorrhage</td>
<td>1.75</td>
<td>0</td>
</tr>
<tr>
<td>loratadine</td>
<td>overdose</td>
<td>27.25</td>
<td>0</td>
</tr>
<tr>
<td>adalimumab</td>
<td>injection site pain</td>
<td>2.29</td>
<td>4.82e-277</td>
</tr>
<tr>
<td>warfarin</td>
<td>international normalised ratio increased</td>
<td>1.88</td>
<td>3.82e-250</td>
</tr>
<tr>
<td>risperidone</td>
<td>gynaecomastia</td>
<td>8.68</td>
<td>4.38E-250</td>
</tr>
<tr>
<td><strong>Youth</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>misoprostol</td>
<td>haemorrhage</td>
<td>5.01</td>
<td>7.15e-291</td>
</tr>
<tr>
<td>finasteride</td>
<td>erectile dysfunction</td>
<td>9.99</td>
<td>1.48e-179</td>
</tr>
<tr>
<td>mifepristone</td>
<td>abortion incomplete</td>
<td>1.98</td>
<td>3.88e-179</td>
</tr>
<tr>
<td>alprazolam</td>
<td>cardiac arrest</td>
<td>6.60</td>
<td>7.70e-166</td>
</tr>
<tr>
<td>alprazolam</td>
<td>respiratory arrest</td>
<td>7.20</td>
<td>2.29e-129</td>
</tr>
<tr>
<td>doxycycline</td>
<td>haemorrhage</td>
<td>5.75</td>
<td>2.43e-128</td>
</tr>
<tr>
<td><strong>Adult</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aspirin</td>
<td>myocardial infarction</td>
<td>3.06</td>
<td>1.63e-278</td>
</tr>
<tr>
<td>isotretinoin</td>
<td>unintended pregnancy</td>
<td>2.64</td>
<td>1.76e-248</td>
</tr>
<tr>
<td>levonorgestrel</td>
<td>acne</td>
<td>6.96</td>
<td>2.78e-217</td>
</tr>
<tr>
<td>furosemide</td>
<td>cardiac failure congestive</td>
<td>1.91</td>
<td>9.90e-213</td>
</tr>
<tr>
<td>aspirin</td>
<td>cardiac failure congestive</td>
<td>1.92</td>
<td>9.88e-167</td>
</tr>
<tr>
<td>etonogestrel</td>
<td>pulmonary embolism</td>
<td>2.46</td>
<td>1.99e-105</td>
</tr>
<tr>
<td><strong>Senior</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vincristine</td>
<td>febrile neutropenia</td>
<td>1.60</td>
<td>3.21e-256</td>
</tr>
<tr>
<td>etanercept</td>
<td>rheumatoid arthritis</td>
<td>1.51</td>
<td>9.51e-177</td>
</tr>
<tr>
<td>salicylic acid</td>
<td>completed suicide</td>
<td>2.74</td>
<td>1.19e-166</td>
</tr>
<tr>
<td>cyclosporine</td>
<td>eye irritation</td>
<td>6.78</td>
<td>1.30e-139</td>
</tr>
<tr>
<td>ethanol</td>
<td>completed suicide</td>
<td>2.55</td>
<td>2.37e-136</td>
</tr>
<tr>
<td>medroxyprogesterone</td>
<td>breast cancer female</td>
<td>1.99</td>
<td>3.51e-114</td>
</tr>
</tbody>
</table>

*ROR is the odds ratio of the highest risk group over other groups.*

induced cardio and vascular events (bradycardia, cardiogenic shock, electrocardiogram qrs complex prolonged and hypertension) have a higher tendency in children and youth, which were not observed in previous clinical trials\(^{18}\). Meanwhile, a simvastatin-induced muscle symptom (myalgia) was found at a higher risk in the young population. Although this result is not consistent with a meta-analysis in 2008\(^{19}\)which revealed that age (> 65 years old) is a risk factor for myopathy and rhabdomyolysis, almost all studies included in the meta-analysis were conducted in the adult population and they did not find enough data for the young population. In addition, we also conducted the Standardised MedDRA Queries (SMQs) evaluation regarding myopathy. The SMQ terms used in the analysis included Rhabdomyolysis/myopathy (SMQ code: 20000002) that consists of ten narrowly defined PT terms. We obtained a significant association (adjust p-value = 1.56e-06) between simvastatin and Statin associated muscle symptoms (SAMS) in the young population.

Our observation of the higher risk of children or youth in atorvastatin- and simvastatin-induced ADE could be attributed to dose. Current pediatric statin dosage recommendations are extrapolated from existing adult data\(^{20}\). Usual pediatric doses of atorvastatin and simvastatin for familial hypercholesterolemia (FH) patients are similar to those in adults (10–80 mg/day). Although some clinical trials in children proved atorvastatin and simvastatin’s efficacy and safety at these dose levels, there existed many limitations of these findings. The most critical one is the duration of statin therapy in these clinical trials. In clinical practice, patients with FH are subjected to continue with statin treatment for the rest of their lives once the therapy was initiated while the duration of trials could range from only 8 to...
Table 3: Top age risks posted by the most prescribed drugs in the US

<table>
<thead>
<tr>
<th>Drug</th>
<th>Adverse Drug Event</th>
<th>Age Risk Group</th>
<th>ROR*</th>
<th>Adjusted P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>atorvastatin</td>
<td>type 2 diabetes mellitus</td>
<td>Children</td>
<td>3.28</td>
<td>3.86e-10</td>
</tr>
<tr>
<td></td>
<td>cardiogenic shock</td>
<td>Youth</td>
<td>46.24</td>
<td>4.25e-12</td>
</tr>
<tr>
<td></td>
<td>angina unstable</td>
<td>Adult</td>
<td>1.93</td>
<td>2.46e-06</td>
</tr>
<tr>
<td></td>
<td>atrial fibrillation</td>
<td>Senior</td>
<td>4.91</td>
<td>1.17e-08</td>
</tr>
<tr>
<td>lisinopril</td>
<td>renal impairment</td>
<td>Children</td>
<td>4.76</td>
<td>5.74e-04</td>
</tr>
<tr>
<td></td>
<td>shock</td>
<td>Youth</td>
<td>16.03</td>
<td>5.88e-05</td>
</tr>
<tr>
<td></td>
<td>coronary artery disease</td>
<td>Adult</td>
<td>2.30</td>
<td>5.97e-09</td>
</tr>
<tr>
<td>amlodipine</td>
<td>generalised oedema</td>
<td>Children</td>
<td>12.32</td>
<td>5.06e-12</td>
</tr>
<tr>
<td></td>
<td>shock</td>
<td>Youth</td>
<td>11.30</td>
<td>1.85e-83</td>
</tr>
<tr>
<td></td>
<td>myocardial infarction</td>
<td>Adult</td>
<td>2.42</td>
<td>9.82e-20</td>
</tr>
<tr>
<td>amoxicillin</td>
<td>urticaria</td>
<td>Children</td>
<td>3.51</td>
<td>6.95e-08</td>
</tr>
<tr>
<td></td>
<td>deep vein thrombosis</td>
<td>Youth</td>
<td>4.49</td>
<td>6.95e-08</td>
</tr>
<tr>
<td>omeprazole</td>
<td>injury</td>
<td>Youth</td>
<td>12.06</td>
<td>4.5e-18</td>
</tr>
<tr>
<td>metformin</td>
<td>glomerular filtration rate decreased</td>
<td>Children</td>
<td>7.98</td>
<td>1.49e-06</td>
</tr>
<tr>
<td></td>
<td>intentional overdose</td>
<td>Youth</td>
<td>23.39</td>
<td>7.10e-40</td>
</tr>
<tr>
<td></td>
<td>myocardial infarction</td>
<td>Adult</td>
<td>1.35</td>
<td>3.71e-22</td>
</tr>
<tr>
<td></td>
<td>decreased appetite</td>
<td>Senior</td>
<td>1.89</td>
<td>2.72e-49</td>
</tr>
</tbody>
</table>

* ROR is the odds ratio of the highest risk group over other groups.

Table 4: Statin drugs-ADE pair with children or youth risks

<table>
<thead>
<tr>
<th>Statin Drug</th>
<th>Adverse Drug Event</th>
<th>Age Risk Group</th>
<th>Adjusted P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>atorvastatin</td>
<td>cardiogenic shock</td>
<td>Youth</td>
<td>4.25e-12</td>
</tr>
<tr>
<td></td>
<td>electrocardiogram qrs complex prolonged</td>
<td></td>
<td>4.83e-12</td>
</tr>
<tr>
<td></td>
<td>type 2 diabetes mellitus</td>
<td>Children</td>
<td>3.86e-10</td>
</tr>
<tr>
<td></td>
<td>squamous cell carcinoma of skin</td>
<td>Youth</td>
<td>4.63e-10</td>
</tr>
<tr>
<td></td>
<td>actinic keratosis</td>
<td>Youth</td>
<td>1.81e-7</td>
</tr>
<tr>
<td></td>
<td>hypertension</td>
<td>Youth</td>
<td>1.73e-6</td>
</tr>
<tr>
<td></td>
<td>seizure</td>
<td>Youth</td>
<td>8.54e-5</td>
</tr>
<tr>
<td></td>
<td>intentional overdose</td>
<td>Youth</td>
<td>2.02e-4</td>
</tr>
<tr>
<td></td>
<td>bradycardia</td>
<td>Youth</td>
<td>2.69e-4</td>
</tr>
<tr>
<td>lovastatin</td>
<td>completed suicide</td>
<td>Children</td>
<td>2.50e-17</td>
</tr>
<tr>
<td>rosuvastatin</td>
<td>completed suicide</td>
<td>Youth</td>
<td>2.00e-11</td>
</tr>
<tr>
<td>simvastatin</td>
<td>myalgia</td>
<td>Children</td>
<td>4.53e-17</td>
</tr>
<tr>
<td></td>
<td>personality change</td>
<td>Children</td>
<td>1.32e-7</td>
</tr>
<tr>
<td></td>
<td>completed suicide</td>
<td>Children &amp; Youth</td>
<td>3.26e-7</td>
</tr>
</tbody>
</table>

104 weeks. The accumulating dose of statin drugs could not be evaluated in the clinical trials.

On the other hand, pharmacogenetics might also cause a higher risk in children and youth. For instance, Wagner’s non-compartmental analysis of children and adolescent data\textsuperscript{21} demonstrated that each copy of the SLCO1B1 c.521C allele was associated with a 2.5-fold increase in simvastatin acid (SVA, the active form of simvastatin) systemic exposure, which was more pronounced than reported in adult studies. The 9- to the 10-fold range of AUC values noted within the c.521TT and c.521TC SLCO1B1 genotype groups exceeded the between-group variability, implying that additional factors may contribute to inter-individual variability in SVA systemic exposure in children and adolescents. This could induce a higher risk of simvastatin-induced myopathy in the children and youth population.

Thus, although the short- and intermediate-term efficacy and safety of statins have been confirmed by observational studies and meta-analyses\textsuperscript{22}, there still exists some statin-induced ADEs that have not been fully explored due to
biased study population or short evaluation period of drug effects. We found the variation of ADEs caused by different statin drugs and presented statins-induced ADEs that have the highest risk in children and youth. We explained the plausibility of results in terms of drug usage and pharmacogenetics. Our results could take the essential supplement in clinical trials, especially for the statin drugs-related safety issues.

Conclusion

In this study, we developed a computational methodology to explore the differential risks of a drug causing an adverse drug event in different age groups. We performed multiple Chi-squared tests and disproportionality analysis to examine age risks in drug-ADE pairs. We also removed confounding effects due to age bias through logistic regression and likelihood ratio tests. Finally, We applied the methodology on FAERS data set and identified age-associated drug-ADE pairs as well as their highest age risk. FAERS data is subject to biases due to differential prescription and ADEs are usually dependent on physical conditions of the patients. Thus the proposed methodology can be further improved by adjusting possible confounders, such as gender, co-morbidities and co-prescribed drugs. Our results provided a new resource of age-related adverse events for drugs, which is important for the appropriate prescriptions for patients in different age groups. The new methodology could become an efficient tool in the improvement of precision medicine and drug safety supervision.

Acknowledgement

This work was funded in part by the National Institute of General Medical Sciences (NIGMS) of NIH under award number R01GM141279.
References

Gene Set-Level analyses enable Smaller Studies to compare groups in Rare and Infrequent Diseases or in ultra-fine stratification of common disorders

Dillon Aberasturi MS\textsuperscript{1,3,6}, Nima Pouladi PhD\textsuperscript{2,6}, Samir Rachid Zaim PhD\textsuperscript{1,3}, Colleen Kenost EdD\textsuperscript{1,3,6}, Joanne Berghout PhD\textsuperscript{1,2,4}, Walter W. Piegorsch PhD\textsuperscript{1,3,5}, Yves A. Lussier MD\textsuperscript{1,6}

\textsuperscript{1}Ctr for Biomed Informatics and Biostats (CB2), \textsuperscript{2}Dept. of Medicine, \textsuperscript{3}Graduate Interdisciplinary Prog. in Statistics & Data Science, \textsuperscript{4}Ctr for Appl. Genetics and Genomic Medic., \textsuperscript{5}Bio5 Institute; University of Arizona, Tucson, AZ, USA.
\textsuperscript{6}Dept of Biomedical Informatics; University of Utah, UT, USA

Introduction

The rate limiting accrual of subjects suffering from rare and infrequent disorders creates a pressing need for improving the statistical power of transcriptomic studies in small sample sizes and highly heterogenic data\textsuperscript{1}. However, traditional designs fail to fulfill this demand due to large minimum required sample sizes needed to achieve adequate statistical power. For example, transcriptome-wide Generalized Linear Models (GLMs) analyses have a suggested sample size of at least 20 samples per group of subjects for studying main effects\textsuperscript{2} and to adequately power tests for transcript-level interactions the minimum required sample size must be inflated by a factor of four\textsuperscript{3}.

We propose that gene set level methods can attain better performance in smaller cohort sizes for heterogenic data than transcript level analyses and thereby addresses the above problem. To test this hypothesis, we compare within a retrospective dataset for a variety of small cohort sizes two transcript level GLM designs, one gene set level GLM design, and one method called Inter-N-of-1 which uses meta-analysis of gene set level single-subject-study results\textsuperscript{4}. We later compare the two gene set level methods via a simulation with parameters values\textsuperscript{4} based on observed quantities in the retrospective dataset.

Methods

Datasets: We used the same TCGA human breast cancer cohort data and 5,179 gene sets from Gene Ontology Biological Processes (GO-BPs) mentioned within our previous publication\textsuperscript{4} along with the same methods for filtering and preparing the dataset for analysis. This TCGA human breast cancer cohort data was downloaded from The Cancer Genome Atlas (TCGA)\textsuperscript{5} Breast Invasive Carcinoma data collection on 10/22/2019 and consisted of two cohorts of subjects with each subject having a tumoral and a normal breast-tissue sample. The two cohorts were distinguished by whether the subjects had a TP53 mutation (n = 23) or PIK3CA mutation (n = 19) with none having both.

![Figure 1. Inter-N-of-1 design: two-group comparison by meta-analyses of single-subject-studies.](image)

Equations show the effect size (Eq 1) and the dispersion (Eq 2) of a gene set k expression in a patient j of a cohort i. Eq 3 shows the pooling of effect sizes (log[|gene set enrichment Odds Ratio|]) of different members of a cohort I; Eq 4 and Eq 5 pertain to the comparison of two cohorts (i=A, and i=B), which we call Inter-N-of-1.

Analyses: Four different designs were implemented for the analyses: (1) ‘GLM DEGs’ utilized a contrast between the estimated tumoral levels of the two cohorts; (2) ‘GLM DEG Interactions’ utilized an interaction contrast between cohort designation and tumor status; (3) ‘GLM + Enriched Gene Sets’ (GLM+EGS) uses ‘GLM DEG Interactions’ followed by Fisher’s Exact Test\textsuperscript{6}; and (4) Inter-N-of-1 (Fig.1) uses a two-step design: i) single-subject-study (SSS) design that provides p-values and effect sizes for each pathway in each subject (N-of-1-MixEnrich\textsuperscript{7}), followed by ii) two-group comparison via meta-analyses of SSSs using the protocol of our recent publication\textsuperscript{4} at BH-FDR<5%. These results differ from our previous studies as they focus on the increased accuracy of all methods from gene set transformations.
Reference Standards and Evaluation: We applied GLMs via the R software package `edgeR`\(^9\) following each of the GLM designs to the entire TCGA human breast cancer cohort data to create three separate reference standards. Individual test sets were generated by randomly sampling without replacement equal numbers of subjects from each cohort. For each selected cohort size \(\{3, 5, 7, 9, 11\}\), this process was repeated 100 times without ensuring non-redundancy to create 100 random subsets for each of the selected cohort sizes. To each random subset, we applied Inter-N-of-1 and once for each of the three different designs GLMs implemented via the software package `limma`\(^10\). Prior to each of the GLM implementations within the random subsets, the preprocessed data was `limma` TMM normalized\(^11\) and then `voom` normalized\(^12\) via the `limma` function `voomwithQualityWeights` in R.

Each GLM design within the random subsets was evaluated by the full data `edgeR` reference standard of the same design and the gene set level reference standard evaluated Inter-N-of-1. For each method, precision and recall were calculated in the same manner presented in our previous publication\(^8\).

Simulation: For each of the cohort sizes of \(\{3, 7, 10\}\), 30 simulated datasets were generated according to previous published methods\(^4\) leading to a total of 180 simulations. The Inter-N-of-1 and ‘GLM+EGS’ methods analyzed each simulated dataset and were evaluated for each combination of simulation parameter and cohort size based on their overall precision and recall across 30 runs. The software is available upon request for academics.

Results

![Graph showing precision and recall across 30 random subsets](image)

Figure 2. Increasing accuracy via pathway transformations and enabling very small samples sizes. 180 Simulations (Left): the precision and recall for GO terms of size 40 for each combination of method, cohort size, and proportion of subjects with coordinated DEGs is shown. The proportion of subjects with coordinated DEGs is shown in parentheses in the legend and was set at either the upper quartile (0.75) or the lower quartile (0.25) observed in the breast cancer data. Breast Cancer study (right panel): the precision and recall across 100 randomly sampled subsets of the TCGA human breast cancer cohort data is shown for each of the combinations of method and cohort size.

Discussion

In sample sizes of 3 to 11 subjects per group, gene set-level analyses were more accurate than transcript-level ones in the retrospective dataset and in simulations. This may provide one avenue for improving performance within small cohorts of heterogenic data and thereby enable smaller studies, such as those pertaining to rare genetic disorders for which accrual is rate limiting. However, such an approach is limited in two ways that require future studies: (i) case-controls must be well matched for confounding covariates (e.g., sex, age, ancestry), and (ii) they result in altered gene sets (e.g., biomolecular pathways) rather than the conventional DEGs.

References

GRACE: A Smoothing Algorithm for Removing Heterogeneous Noise and Implausible Fluctuations of Observed Disease Curves

Anushka U. Abeysekara, PhD\textsuperscript{1,2}, Christopher R. Wilson, MS\textsuperscript{1,2}, Scott L. DuVall, PhD\textsuperscript{1,2}, Evan P. Carey, PhD\textsuperscript{1,3}

\textsuperscript{1} VA Informatics and Computing Infrastructure, VA Salt Lake City Health Care System, Salt Lake City, UT, USA; \textsuperscript{2} Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT, USA; \textsuperscript{3} University of Colorado Denver, Denver, Colorado

Abstract

Understanding the state of a disease curve represented as discrete points in a time series, such as publicly reported COVID-19 daily case counts and deaths, is challenging due to the presence of unstructured and irregular noise elements. Limits and delays in the mechanics of underlying data-generation processes such as collection, processing, and reporting, result in cyclic trends of varying lengths and include implausible data points given the context of expected disease dynamics. Standard smoothing algorithms that seek to remove noise (e.g., 7-day average) are not resilient to these traits and are insufficient for smoothing reported disease case curves. This article presents a novel algorithm (GRACE) to smooth noisy disease curves. The input parameters of the GRACE smoother have an explicit relationship with disease dynamics, and the algorithm dynamically adjusts for heterogeneous and cyclic noise.

Introduction

Time series data play a vital role in healthcare, and its importance most recently occurred during the COVID-19 pandemic. Daily reported new cases and deaths are widely used to understand the current status of the pandemic, forecasting spread, and even informing policy. However, the observed data are noisy, which can degrade the accuracy of downstream analyses and it is crucial to remove the noise before further use. Smoothing algorithms such as Simple Moving Average (SMA), Exponential Moving Average (EMA), and others are widely used to remove noise from observed time series data.

All these techniques require user-defined input parameters to operate. For example, SMA requires the user to define the window size. However, the smoothness of the resulting denoised curve will not only depend on the user-defined window size, but also on many other factors within the data itself, such as periodicity, curvature of the trend component, and the existence of sudden spikes. Furthermore, selecting a wider SMA window could over-smooth the input sequence concealing valuable information. Selection becomes even more challenging when the amplitude of the noise and the periodicity of periodic noise components vary over time. Therefore, we propose a novel smoothing algorithm, Growth RAte ConstrainEd (GRACE) smoother, to address the limitations of these standard smoothers.

The GRACE algorithm calculates the observed exponential growth rate ($r_t$) and requires the user to input the allowed maximum curvature of the growth rate ($m$). This hyperparameter, $m$, controls the smoothness of the resulting denoised curve independent of the properties of the noise in the data. As such, the user can control the smoothness of the resulting curve based on actual disease dynamics.

The GRACE smoother is an open-source Python package available on GitHub. Refer to the GitHub repository for the user manual and documentation (https://github.com/udaraabeyskara/GRACE).

Methods

The GRACE smoother computes weights to detect rapid rises and dips in the data that are implausible. It uses these weights to dampen the effect of noise from these spikes. The data are then passed through a Kolmogorov-Zurbenko (KZ) low-pass filter, initially parameterized with a small window. It then removes high frequency components with wavelet filtering. GRACE iterates with more aggressive smoothing by increasing the KZ window until it satisfies the
maximum curvature of the growth rate, $m$, constraining the curve. For each time series, the observed exponential growth rate $r_t$ is calculated, as defined in Error! Reference source not found., where $N_t$ and $N_{t-1}$ are the daily reported cases (or deaths) on day $t$ and $t-1$, respectively.

$$r_t = \frac{N_t - N_{t-1}}{N_{t-1}} \quad \text{Equation 1}$$

The red lines in Figures 1 and 2 illustrate two scenarios: a theoretical Susceptible-Infectious-Recovered (SIR) curve with 0.6% Gaussian noise added and daily reported new COVID-19 cases in the United States between April,07, 2020 and April,13,2021, respectively. The green lines show the corresponding calculated $r_t$ distributions. The distribution for the observed data is filled with fluctuations, unlike the theoretical SIR curve. The curvature of these sharp fluctuations is measured as $r_t curvature$, as shown in Equation 2. Here, the maximum observed $r_t curvature$ is 11.86. The maximum $r_t curvature$ for the theoretical SIR curve with 0.6% noise added is 0.006. The larger $r_t curvature$ quantitatively represents that the observed data fluctuates at a higher frequency and amplitude than the noise added SIR curve. Therefore, to smooth the observed data until it reaches the smoothness of the noise added SIR curve, GRACE iteratively smooths more aggressively and recalculates $r_t curvature$ until the maximum is less than the user-defined $m$ input parameter.

$$r_t curvature = \frac{d^2 N_t}{dt^2} \quad \text{Equation 2}$$

The maximum $r_t curvature$ constrains how abruptly $r_t$ is allowed to change at any point on the curve. When analyzing real-world data such as in scenario 2, the user can define the $m$ input parameter using the theoretical or observed effective reproduction number or growth rate of the infectious disease. In this way, the smoothing of the curve is explicitly linked to the way the disease can practically spread in a community.

**Discussion**

A prominent issue with real-world reported disease data that the GRACE smoother addresses is implausible rises and dips. One extreme example is the spike reported in India’s death data in June 2020. It was reported that there were 2,003 deaths on June 16th, but on 15th and 17th, the reported daily deaths are only 334 and 380. This phenomenon occurs regularly across weekends where data may not be reported followed by catch-up reporting early in the week and for other reasons that generate the characteristic saw-tooth patterns such as in Figure 2. Many smoothers overestimate the effect of these spikes and generate box-like formations surrounding the irregularities. The GRACE smoother is robust against this behavior by dampening the effect spikes have on the smoothed curve.

Many smoothers must use impractically wide windows to approximate disease dynamics. For example, if the SIR curve in Error! Reference source not found. is smoothed using SMA and constrained to a maximum growth rate of 0.006, the window size would grow to 61 days. In contrast, the GRACE smoother can achieve the same curve with only a window of 14 days.

Because the GRACE smoother’s hyperparameter directly relates to disease dynamics, it is more robust against irregularities and can produce a more representative curve. By using both KZ-filter and wavelet-based high-frequency filtering, the smoothness of the resulting denoised curve is independent of the noise properties in the data. These features mean that real-world disease data smoothed with GRACE can support more robust downstream analysis.

**References**

Multimorbidity Patterns Across Race/Ethnicity Stratified by Age and Obesity: A Cross-sectional Study of a National U.S. Sample

Manal Alshakhs, MS\textsuperscript{1}, Bianca Jackson, MPH\textsuperscript{1}, Davina Ikponmwosa, MHIM\textsuperscript{1} Rebecca Reynolds, EdD, MHA\textsuperscript{1}, Charisse Madlock-Brown, PhD, MLS\textsuperscript{1}

\textsuperscript{1}University of Tennessee Health Science Center, Memphis, Tennessee

Introduction
Multimorbidities become more prevalent as individuals age, and they have been associated with substantial burden and increased mortality\textsuperscript{1}. The objective of our study is to assess differences in prevalence of multimorbidity by race. This research will support the development of patient-centered care approaches by stratifying patients with obesity-associated multimorbidity into groups with similar characteristics. Our study is the first to identify distinct prevalent multimorbidities by race stratified by obesity and age.

Methods
We applied the FP-growth algorithm on middle-aged and elderly cohorts using 2016-2017 data from the Cerner HealthFacts\textsuperscript{®} data warehouse. Our outcome variables were multimorbidity patterns prevalent within each racial/ethnic category stratified by age and obesity status. We identified disease combinations above a threshold of 5% prevalence within each cohort. We identified prevalent combinations shared by all races/ethnicities, those shared by some, and those unique to one group for each age/obesity level. We used the Clopper-Pearson method to generate binomial proportion confidence intervals to identify which races/ethnicities may have similar prevalence. To compare prevalence rates by race, we used the g-test of independence. Sensitivity analysis included an ANOVA for medians of BMI averages across race for each age/weight class.

Results
Our criteria matched 1,208,687 patients in the middle-aged cohort and 26,046 patients in the elderly cohort. Due to low patient counts in the Asian/Pacific Islander, Native American, Hispanic and Biracial elderly groups, we combined those patients into one category, Other, in the elderly cohort. Table 2 shows a summary of the results. For each race category, the middle-aged obese category had more multimorbidities than the elderly non-obese. African Americans had the most multimorbidities for each age/weight group. Middle-aged non-obese cohorts had the lowest number of multimorbidities for each race, and elderly obese had the most.

Table 2. The number of multimorbidity patterns stratified by race/ethnicity at 5% threshold across weight categories.

<table>
<thead>
<tr>
<th>Race</th>
<th>Middle Aged Non-Obese</th>
<th>Middle Aged Obese</th>
<th>Elderly* Non-Obese</th>
<th>Elderly Obese</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall</td>
<td>Distinct</td>
<td>Overall</td>
<td>Distinct</td>
</tr>
<tr>
<td>Caucasian</td>
<td>7</td>
<td>0</td>
<td>23</td>
<td>1</td>
</tr>
<tr>
<td>African American</td>
<td>20</td>
<td>5</td>
<td>50</td>
<td>21</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>4</td>
<td>0</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Native American</td>
<td>16</td>
<td>3</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>Hispanic</td>
<td>6</td>
<td>0</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>Biracial</td>
<td>2</td>
<td>0</td>
<td>9</td>
<td>0</td>
</tr>
</tbody>
</table>

We created heatmaps for multimorbidity patterns shared by all groups, shared by some, and those distinct to one race/ethnicity for elderly and middle-aged cohorts. Figure 2 displays a sample representing the multimorbidity patterns shared by only some races/ethnicities for the middle-aged cohort. All combinations included at least one of the following diseases: E11:Diabetes, I10:Hypertension, E78:Lipidemia, M54:Dorsalgia, or M25:other joint disorders. In the middle-aged with obesity cohort, the multimorbidities with the highest prevalence for each group included hypertension. Additionally, I25:Chronic ischemic heart disease and K21:Gastro-esophageal reflux disease appeared in several combinations. There are numerous multimorbidity combinations comprised of 3 morbidities for the elderly obese cohort. Caucasians had fewer multimorbidity patterns in patients with obesity than the African American and

* For elderly, Asian/Pacific Islander, Native American, Hispanic, and Biracial cohorts are combined into “Other races.”
Other cohorts in the elderly. There were no shared multimorbidity patterns between Caucasians and Other (other races) in the elderly cohorts.

**Figure 2.** Figures 2a and 2b show shared multimorbidities across two or more races in middle-aged non-obese and patients with obesity, respectively, with 95% confidence and at a 5% prevalence threshold. The g-test of independence was calculated, resulting in significant results with p<.05.

**Discussion**

This is the first study to identify multimorbidity disease prevalence of any size by racial/ethnic category, stratified by age and obesity status. Our findings demonstrate that even after controlling for age and obesity, there are differences in multimorbidity prevalence across races, and some combinations are distinct to racial groups. Although we are not the first to use a frequent itemset algorithm to assess multimorbidity, our study adds to the current knowledge by examining the prevalence of specific multimorbidity patterns by racial/ethnic category, stratified by age and obesity status. Our findings also demonstrate that there are multimorbidity combinations distinct to racial groups, particularly amongst the middle-aged cohort. Multimorbidity increased with age in both the obese and the non-obese groups. Many identified disease combinations have likely not been studied, as identifying unique patterns among races is unprecedented. Multimorbidity prevalence was the highest among African Americans and lowest among Asian/Pacific Islanders. Even when factoring in age and weight class, the differences remain. Common morbidities present in disease combinations across all races were lipidemia, hypertension, and diabetes regardless of age or obesity level. This research supports the development of patient-centered care approaches. Providers should be aware of the most likely multimorbidity combinations to develop and coordinate appropriate care plans when treating patients with these disease groupings. Guidance on how best to manage specific patients based on their multimorbidity profile could be beneficial for patients and useful for the providers providing care.

**References**

Lessons Learned from Large Scale Curation of EHR Data

Nadia S.P. Bahadur, MSCR¹, Andrew Niederhausern¹, Neil J. Shah, MD¹, Chelsea Nichols, MPH¹, Fenil Gandhi, MBBS¹, Kimberly Chan¹, Laura E. Barton, CCRP¹, Alysha Insinga, MS¹, Grace Capreol¹, Lauren Esposito, MS, MPH¹, Latasha Estrada, MPH¹, Joshua Goldberg¹, Ralph Nelson¹, Jasme Lee¹, Samantha Brown¹, Jessica A. Lavery, MS¹, Molly Maloy, MS¹, Katherine S. Panageas, DrPH¹, Theodora A. Bakker, MSLIS¹ and John Philip, MS¹
¹ Memorial Sloan Kettering Cancer Center (MSKCC), New York City, NY, United States

Introduction
MSK-IMPACT is a custom genomic assay of 505 genes to identify actionable targets to help drive precision treatment. MSKCC currently has about 62,000 IMPACT patients assayed and continues to accrue more every day. Disease management teams (DMTs) lack the structured data from the patient’s clinical journey to analyze alongside the genomic data. Thus, researchers and clinicians spend a significant amount of time extracting data from the EHR. We use a vendor called VASTA (www.vastaglobal.com) to extract key clinical elements and store them in MSKCC’s enterprise Unified Data Fabric (UDF). Here we present key lessons learned as the project was slowly increased in scale and moved into full operation.

Methods
The curation cohort includes any cancer patient with IMPACT testing and a minimum of 1 year of follow up after their assay. Key steps in the curation project are:

1. Establish a team and third-party vendor for curation. VASTA is contracted from September 2020 to April 2022 for 38,000 IMPACT patients at MSKCC. Staffing for this project includes a VASTA team, and a MSKCC team which consists of a manager, quality control (QC) MD lead, 5 research project associates (RPAs) for QC and a bioinformatics engineer.

2. Establish the data model: ASCO’s mCODE, NAACCR data model and AACR GENIE PRISSMM model for progression of disease are all pan-cancer data models. However, each model is missing some component of the full cancer journey. Thus, we combined all three models, merging equivalent fields and keeping the differences, resulting in a model we call Core Clinical Data Elements (CCDE). It covers patient characteristics, comorbidities, cancer diagnosis (TNM staging), medication history, external cancer surgeries, radiation oncology history, pathology reports, imaging reports and tumor marker information. All resulting data elements were reviewed and approved by a clinical advisory board of MSKCC MDs.

3. Establish data sources: The team identified structured data sources from the patient’s EHR and associated it with the CCDE fields. The structured data was piped from the source to REDCap. The structured data is used to help provide context during manual curation.

4. Train the curators: From each DMT, we asked a clinician or research manager to teach the disease-specific nuances of that cancer type to our QC team of five MSKCC RPAs. The team made training slides for VASTA on the curation guidelines. The RPAs and VASTA each conducted pilot curation of the same 15 patients. The RPAs performed quality control review on VASTA’s pilot patients and addressed any re-training issues with VASTA.

5. Establish quality control: For each cancer type the team defined major and minor errors per REDCap form. Once curation began, VASTA conducted 10% review with a separate team of QC managers. The MSKCC QC team conducted their own separate 10% QC by reviewing the weekly QC release from VASTA. The biostatistics teams, using custom R Shiny application, summarized the major and minor errors found by both QC teams. Any errors were corrected by VASTA and then reviewed by the RPA team.

6. Store the final QC reviewed data in an enterprise data repository: After curation and quality control were completed, we migrated the abstracted discrete/structured dataset into the UDF.

Results
The VASTA team, consisting of 25 curators, has completed 1,427 bladder cancer patients, across 94,794 individual forms, including 1,596,477 curated fields from December 2020 to May 2021. After completion of the bladder cancer dataset, QC queries resulted in approx. 3,000 findings. Once QC is complete, the ingestion of the curated discrete dataset will be about 14 million fields.
Discussion

Key Lessons Learned: Applying the lessons learned during curation of the bladder cancer cohort to the next cohort, prostate cancer, has allowed us to decrease the total time from setting up each disease subset, through curation, to QC and data use.

• How to monitor and improve curation time.

In the first cohort, 1,427 bladder cases were curated from December 2020 to May 2021 (5 months) for 25 curators. For the prostate cancer cohort, we monitored the average curation time and number of completed patients per week using real-time updates in Tableau (figure 1). Dashboard monitoring offered points of improvement. We inquired with VASTA on increased curation trends on certain forms to ensure curation time steadily decreases as VASTA becomes accustomed to each cohort. For instance, in figure 1 average curation time spiked to 505.9 minutes when it was around 400 minutes the previous week. We were able to quickly notify VASTA and they informed us that they had newly trained curators started that week that accounted for this jump in curation time. With increased monitoring, we were able to provide real-time feedback to VASTA to investigate potential issues.

• How to improve Quality Control

For the first cohort we monitored the amount of major and minor violations per week and average number of patients completed per week (figure 1). For the Prostate cohort, we added complex queries during curation to identify errors sooner and correct common mistakes ahead of time rather than waiting until the end of curation.

• How to show clinicians the value of the curated data.

Early on, we noted the importance of having SMEs to provide buy in from their DMTs. This is necessary to provide accurate guidelines for VASTA. For bladder, prostate and subsequent cohorts we continued to refine our messaging to demonstrate how DMTs will save time by using our pan-cancer defined fields. The QC from the tableau dashboards gave the DMTs confidence in our dataset. The initial DMT took 6 months to fully support the project. Later DMTs have onboarded in half the time.

General Advice and Practical Guidelines

We suggest 1 QC person per 11 VASTA curators. Our QC medical doctor spends 30% FTE effort to oversee the QC team of 5 RPAs. Start up for the project takes the most time. Creating the data model, identifying discrete data sources to non-curated fields, hiring and training staff took 6 months. Then transitioning between cancer types, creating their data models, meetings with MD experts and training on new cancers take 3-6 months. It is important to have parallel discussions with multiple DMTs. Small curation pilots are needed to assess curation time of the data model for different cancers to create an estimated timeline for each cohort. When curation time was too large, we cut certain fields (tumor size, number of regional lymph nodes examined) to reduce curation time.

Conclusion

We originally mitigated some of the pitfalls of large-scale curation by monitoring data quality and curation by creating real-time tableau dashboards and notifying VASTA of performance issues. More importantly, we learned how to establish trust and commitment with the clinical leadership of DMT teams to have the next disease type ready for curation and to ensure the accurate curation of these cancers. Future work includes using AI/ML or NLP tools to prepopulate the manually curated fields. We are also building a curation assistant tool (CAT) that displays the EHR source data quickly to the curators.

References

Using machine learning to predict the negative patient experience using the HCAHPS survey

Vitej Bari, MS¹, Jamie S. Hirsch, MD¹, Joseph Narvaez, MBA¹, Robert Sardinia, MS¹, Kevin Bock, MD¹, Michael Oppenheim, MD¹, Marsha Meytlis, PhD¹
¹Northwell Health, Lake Success, NY, USA

Abstract

Patient experience, an essential quality metric, is influenced by numerous contributing factors including physician communication. We built ten machine learning models using clinical data to predict patients’ inpatient hospital experience. Responses to the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey, were predicted, with AUC range 0.647 – 0.707. Machine learning models can accurately identify patients at risk of a negative experience, in near real-time, in support of proactive service recovery efforts.

Introduction

In addition to underlying quality of care delivery, a major goal of healthcare systems is to provide a positive experience to every patient throughout their medical journey. Positive patient experience is a healthcare quality aim proposed by the Institute of Medicine (IOM) and patient experience is positively associated with clinical effectiveness and patient safety. Patient experience insights can therefore be used to drive strategies for clinical practice as well as overall system transformation. The HCAHPS survey is the first national, standardized, publicly reported survey of patients’ perspectives of hospital care, and represents an objective measure of patient experience. The Centers for Medicare & Medicaid Services (CMS) developed HCAHPS star ratings to assess excellence in healthcare quality. The HCAHPS survey is comprised of 19 questions in 10 domains: doctor communication, nurse communication, response to staff, discharge information, care transition, communication about medicine, quietness, cleanliness, food and respect of religion. In this work we build a separate machine learning model for each domain. Patients’ responses are grouped into positive and negative. A patient’s response is considered positive if they have a top box response across all individual questions within a domain, all other responses are considered negative.

Methods

The patient cohort consisted of 84,309 patients who were admitted to 11 hospitals within the Northwell Health system between January 1, 2018, and December 31, 2020, and responded to the post-discharge HCAHPS survey. The significance of model AUC differences was evaluated by using k-fold cross validation. We considered the mean AUC of each model over 5 validation folds. XGBoost models were trained on 70% of available data and their predictive ability tested on the remaining 30%. Area under the receiving operator characteristic curve (AUC) was used as the primary metric, with 95% confidence intervals calculated by bootstrapping with 1,000 iterations. The distinct features that were included in the model are age, gender, race, marital status, language, religion, admission through the emergency department (ED), Charlson comorbidity index, hospital length of stay, number of ED visits in the prior 6 months, pain score, allergy count, ability to hear, chronic pain, Nursing Unit the patient stayed in, hour of day patient was admitted, the final hospital the patient was in, grouped ICD-10 code, and order for blood culture. We also tested the following features but did not include them in the model: if there were any surgical site incisions at admission, vision impairment, order for supplementary oxygen, order for Xray, order for CTMR. These features were not included because they did not show any relationship with patient experience. A threshold of 0.5 was used in all the models. Separate models were created to predict a negative experience to each of HCAHPS domain. All clinical data were collected from the enterprise inpatient electronic health record database (Sunrise Clinical Manager: Allscripts, Chicago, IL). At Northwell Health, all HCAHPS surveys are administered by Press Ganey Associates (South Bend, IN), and results are returned to the health system.
Results:
We measured the prediction performance of ten machine learning models, which represented the experience of patients across 10 domains. The performance of each model is shown in Table 1 below. Response to Staff was the highest performing model with AUC of 0.707 while Discharge Information was the lowest performing model with an AUC of 0.647.

<table>
<thead>
<tr>
<th>Domain</th>
<th>AUC</th>
<th>Confidence Interval</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>NPV</th>
<th>PPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doctor Communication</td>
<td>0.692</td>
<td>0.686 - 0.697</td>
<td>0.172</td>
<td>0.956</td>
<td>0.724</td>
<td>0.63</td>
</tr>
<tr>
<td>Nurse Communication</td>
<td>0.68</td>
<td>0.674 - 0.687</td>
<td>0.156</td>
<td>0.961</td>
<td>0.71</td>
<td>0.652</td>
</tr>
<tr>
<td>Response to Staff</td>
<td>0.707</td>
<td>0.670 - 0.713</td>
<td>0.844</td>
<td>0.407</td>
<td>0.627</td>
<td>0.688</td>
</tr>
<tr>
<td>Discharge Information</td>
<td>0.647</td>
<td>0.667 - 0.681</td>
<td>0.063</td>
<td>0.988</td>
<td>0.741</td>
<td>0.653</td>
</tr>
<tr>
<td>Care Transition</td>
<td>0.673</td>
<td>0.667 - 0.680</td>
<td>0.955</td>
<td>0.146</td>
<td>0.602</td>
<td>0.705</td>
</tr>
<tr>
<td>Communication about medicine</td>
<td>0.653</td>
<td>0.646 - 0.660</td>
<td>0.74</td>
<td>0.457</td>
<td>0.585</td>
<td>0.63</td>
</tr>
<tr>
<td>Quietness</td>
<td>0.706</td>
<td>0.701 - 0.712</td>
<td>0.6</td>
<td>0.683</td>
<td>0.643</td>
<td>0.642</td>
</tr>
<tr>
<td>Cleanliness</td>
<td>0.701</td>
<td>0.695 - 0.707</td>
<td>0.17</td>
<td>0.977</td>
<td>0.743</td>
<td>0.749</td>
</tr>
<tr>
<td>Food Domain</td>
<td>0.701</td>
<td>0.695 - 0.707</td>
<td>0.906</td>
<td>0.309</td>
<td>0.656</td>
<td>0.693</td>
</tr>
<tr>
<td>Respect Religion</td>
<td>0.702</td>
<td>0.696 - 0.708</td>
<td>0.22</td>
<td>0.958</td>
<td>0.709</td>
<td>0.725</td>
</tr>
</tbody>
</table>

Table 1. XGBoost model performance of the domain models of the HCAHPS survey

Conclusion
An XGBoost model using data routinely available in medical records provided guidance for improving patient experience by accurately predicting patients at risk of having a negative experience. This machine learning model is currently being placed into production to facilitate communication and service recovery with the patient experience team.

References
Secondary Academic Appointments: A Hidden Dimension of Disparity?
Suresh K. Bhavnani PhD1, Weibin Zhang PhD1, Tho Ayala BS2, Craig K. Cassidy MHA3, Sandra Hatch MD4, Alfredo G. Torres PhD MS5, Charles Mouton MD MS2,6, Randall J. Urban MD7
1Preventive Medicine and Population Health, 2Office of the Provost, 3Office for Research Advancement, 4MD Anderson Cancer Ctr. 5Microbiology and Immunology, 6Family Medicine, 7Internal Medicine, University of Texas Medical Branch, Galveston, TX

Introduction

Although equal-employment opportunities have been protected by the Civil Rights Act for more than half a century, specific races/ethnicities, women, and the economically-disadvantaged, continue to be under-represented in biomedical research. For example, while African Americans, Hispanics, Native Americans, and Pacific Islanders are together expected to form a majority of the US population by 2050, they currently form less than 10% of the biomedical research workforce.1 While such employment disparities are easily measurable within organizational units such as in a department, school, or university, other potential disparities as in secondary academic appointments occur across organizational units, requiring more powerful analytical methods. Furthermore, while secondary appointments are valuable markers of multidisciplinary expertise providing access to a wider pool of collaborations, funding and publications, such appointments are not required to be reported when they are without salary (WOS), and remain under-scrutinized. Such analytical and reporting reasons could result in undetected disparities within secondary appointments. Here we demonstrate how a bipartite network analysis and visualization can help to reveal disparities concealed in secondary appointments, enabling the design of targeted interventions.

Method

Research Question. What is the nature of diversity within primary and secondary faculty appointments?

Data. We extracted all deidentified records of paid faculty (n=879) from the administrative database of an academic medical center. Variables included appointment type (primary, secondary), school (School of Medicine [SOM], School of Health Professions [SHP], School of Nursing [SON]), department (d=33, e.g., Neurology), rank (assistant, associate, professor), gender (male, female), and race (White, African American, Hispanic, Pacific Islander, American Indian, Asian). All 879 faculty members had paid primary appointments, of whom 140 also had WOS secondary appointments.

Analysis. The analysis consisted of 3 steps: (1) Bipartite Network Analysis and Visualization. (a) Represented the data as a bipartite network (Fig. 1), where nodes represented faculty members (circles) or departments (triangles). The edges (lines) connecting faculty members to departments represented a primary appointment (dark gray), or a secondary appointment (light gray), and weighted based on their importance (primary=2, secondary=1). Diameter of the faculty member nodes represented rank (large=professor, medium=associate, small=assistant). (b) Used bicluster modularity maximization2 to identify the number and boundaries of department-faculty biclusters, and the degree of biclustering (Q). (c) Measured the significance of Q by comparing it to a distribution of Q generated from 1000 random permutations of the network. (d) Used the force-directed algorithm Kamada-Kawai to layout the nodes within each department, and then grouped by school (SOM, SHP, SON) to

Fig. 1. Bipartite network visualization of primary and secondary appointments across departments, organized by school.
improve comprehension. (2) **Node and Edge Enrichment Analysis.** Used chi-squared with Bonferroni correction to measure: (a) the difference in proportion of race (white non-Hispanic vs. the rest of the races), and gender within each department compared to the US population; (b) the difference in proportion of race, gender, and rank of faculty members with secondary appointments, compared to those with only primary appointments. Used a binomial distribution to measure the significance of the proportion of inter-departmental edges (representing secondary appointments between pairs of departments) within biclusters in both directions, compared to the expected number of those edges (taking into consideration the department size) between all department pairs in the network. (3) **Interpretation.** The results were presented to a multidisciplinary team of data and governance administrators from the provost’s office of the university to determine: (a) interpretability of the network and whether the results aligned with their administrative experience; (b) potential causal mechanisms underlying the observed disparities; and (c) potential interventions to address disparities.

**Results**

**Structure of Primary and Secondary Appointments.** The bipartite network analysis and visualization revealed: (a) **departmental clustering** (same-colored nodes in Fig. 1), which was significant (Q=0.86, mean of random Q=0.59, p<.005, two-tailed). This resulted in 4 multi-departmental clusters within a school (e.g., SOM:Otolaryngology and SOM:Neurosurgery), 4 multi-departmental clusters that spanned 2-3 schools (e.g., SOM:PMH, SHP:Nutrition/Metabolism, and SON:Nursing PhD Program [Fig. 1A]), and 15 clusters that contained a single department; and (b) **reciprocity** (Fig. 1B) in secondary appointments within clusters existed between 2 pairs of departments, which were significant compared to the expected across all departmental pairs. For example, SOM:Microbiology/Immunology (A) and SOM:Pathology (B) had significantly higher secondary appointments with each other (A→B=5, p<.001; B→A=7, p<.001, two-tailed), and therefore were reciprocal. In contrast, SOM:PMH (A) and SHP:Nutrition/Metabolism (B) had significantly higher secondary appointments only in one direction (A→B=5, p<.001; B→A=1, p=.9; two-tailed), and therefore were not reciprocal.

**Diversity in Primary and Secondary Appointments.** Enrichment analysis revealed: (a) **intra-departmental disparity** which consisted of significantly higher males within 2 SOM departments (SOM:Internal Medicine, SOM:Surgery), significantly higher females within two SOM departments (SOM:Baccalaureate-Prg., and SON:Master’s-Prg.), and significantly higher non-whites within SOM:Internal Medicine (p<.001, department=58%, US population=40%), compared to the US population; (b) **inter-departmental disparity in secondary appointments** (Fig. 1C) consisted of a significantly higher number of faculty with secondary appointments that were white (p<.05, secondary=66%, primary=55%), and full professors (p<.01, secondary=72%, primary=22%), compared to those with only a primary appointment. In contrast, there was no significant difference in secondary appointments based on gender (p=.05, secondary=63% male, primary=53% male).

**Interpretation of Structure and Diversity.** The executive leadership found the results to be interpretable, despite some of the findings being unexpected. The expected results included the known disparity in race within primary appointments in SOM:Internal Medicine, and known gender disparity within SON departments. They inferred that such disparity might exist because SOM departments tend to hire from a limited pool of international graduating students in the US, whereas the SON departments tend to hire from a pool of female nurses. Both imbalances could be addressed by drawing from more diverse pools, coupled with inclusion strategies to increase their retention. In contrast, as secondary appointments are rarely scrutinized, disparities in those results were unexpected. They reasoned that faculty with secondary appointments tend to be mostly white and full professors, either because such appointments were offered as administrative roles after they were promoted, or were offered as part of a start-up package during recruitment. While more data is needed to identify specific mechanisms precipitating such disparities, there was consensus that secondary appointments needed more scrutiny, with a focus on how they impact the trajectory of junior faculty. Such analyses would enable the design of targeted interventions such as offering meaningful secondary appointments to junior faculty. Furthermore, as junior faculty tend to be more racially diverse, such an intervention has the potential to address the current race and rank disparity across secondary appointments.

**Conclusions and Future Research**

While diversity in primary appointments is frequently and easily analyzed using conventional methods, disparity in secondary appointments can remain under the radar as they are not required to be reported when they are WOS, and require advanced methods to analyze and comprehend complex associations across departments. Here we demonstrated how bipartite networks enabled (1) the quantitative identification of significant disparities in both primary and secondary appointments, and (2) the visual interpretation of disparity patterns by a team of multidisciplinary administrators, enabling them to collaboratively reason about potential underlying mechanisms, and design potential targeted interventions. Our future research aims to test the generality of the approach, and replicability of the results using data from other medical schools, in addition to measuring the impact of diversity interventions on career outcomes of junior faculty using longitudinal data.

**Acknowledgements.** Funded in part by UTMB CTSA (UL1 TR001439), and MD Anderson Cancer Center.

**References**

Healthcare assisting chatbot for personalized air quality notification and health recommendation using IoT: design and implementation

Chiraphat Boonnag, MD 1, Piyapong Khumrin, MD PhD 1,2
1Biomedical Informatics Center, Faculty of Medicine, Chiang Mai University, Thailand; 2Department of Family Medicine, Faculty of Medicine, Chiang Mai University, Thailand

Introduction

Thailand’s air quality regularly reaches dangerous levels between January and April, also known as the “burning season.” Smog and haze from neighboring countries and annual forest and agricultural burning are the primary reasons for poor air quality in most cities in Thailand’s northern area. Fine particulate matter (PM2.5) has been identified as a substantial cause of mortality worldwide1. PM 2.5 is extremely small and can easily be deposited into alveoli and even into the circulation, triggering an inflammatory response and affecting many parts of a human physiological system.

Many air quality stations based on IoT technology have been created to efficiently monitor air quality and communicate data to a web server over the Internet in real-time. The Internet of Things (IoT) has become the fundamental backbone of all systems in recent years, and it is already having an impact on smart cities. Because they have the ability to accept and exchange data linked to existing internet infrastructure, chatbots are already being utilized as IoT interfaces. Therefore, we developed an IoT-enabled chatbot with personalized air quality notification and health recommendations to increase data accessibility and personal health awareness.

Methods

We developed a chatbot named “Smokealert” which was designed to enhance air quality information and personal health recommendation on top of an existing online instant messaging application, LINE. We chose this app because of its popularity in Thailand in comparison to other apps, as well as its strong user interface and user experience.

System design and architecture

This end-to-end system was deployed on the cloud. The detail of specific informatics tools we built to support the IoT-enabled chatbot system were shown in figure 1. Cloud solutions can facilitate and simplify system expansion or support applications with enormous data and users2.

Figure 1. Architecture of the IoT-enabled chatbot system deployed on google cloud

Features

We utilized natural language processing (NLP) to interpret the user’s intents, which entails identifying the user’s motivation to initiate each dialogue and deliver the correct response. This IoT-enabled chatbot had four main contributions as follows:

A) Personal history taking; When users added this chatbot, the system greeted them while inquiring about their health background, including age group, location, and underlying conditions related to PM2.5, as shown in
This information was utilized to categorize users in order for them to be informed of hazards and given personalized health recommendations for their behavior.

B) Air quality notification; Every day, Smokealert sent out an air quality message with user-specific suggestions based on users’ underlying conditions in the morning and evening to raise awareness of PM 2.5, as illustrated in figure 2B.

C) Air quality checker; If the user wanted to explore or inquire about the air quality at that time or location, Smokealert would report US AQI, PM2.5 volume from the nearest measurement station, and the user’s specific location, as shown in figure 2C.

D) Symptom checker; Furthermore, users may inquire the chatbot about symptoms related to air pollution, as illustrated in figure 2D.

Figure 2. (A) personal history taking, (B) air quality notification, (C) Air quality checker and (D) Symptom checker.

Results

The service from January 1, 2021, to June 30, 2021, had 2,042 users. Users consisted of 1,324 females (64.8%) and 718 males (35.2%). Furthermore, the participants were mostly 50 years old and older (658 or 32.2%), followed by 20-29 years old and 40-49 years old, with 23.8% and 20.2%, respectively. It was found that most of the users (40.3%) were in normal health, 32.7% more allergic to air pollution than the general population, and 26.9% had underlying condition related and were at risk of health to PM 2.5.

Discussion and conclusion

We applied the notion of combining chatbots, IoT, and personal health profiles to solve health issues such as real-time air quality notifications and personalized health advice. Smokealert is a helpful IoT-based tool to enable rapid and widespread notification and health recommendations.

Challenges included the communication need and coordinating the needs of multiple stakeholders. Future development would involve creating a multi-purpose chatbot to make it more accessible to people. Improving the user interface and user experience would be another complex undertaking because it is essential that people can utilize the chatbot efficiently.

References

Comparing EHR data collected from FHIR API with an OMOP Database

Robert Carroll, Ph.D.\textsuperscript{1}, Karthik Natarajan, Ph.D.\textsuperscript{2}, Alex C. Cheng, Ph.D.\textsuperscript{1}, Francesco Delacqua\textsuperscript{1}, Michelle Jones\textsuperscript{1}, Colleen Lawrence, Ph.D.\textsuperscript{1}, Paul A. Harris, Ph.D.\textsuperscript{1}

\textsuperscript{1}Vanderbilt University Medical Center, Nashville, TN
\textsuperscript{2}Columbia University, New York City, NY

Introduction

Electronic health record (EHR) data have played a significant role in identifying disease trends and have led to discoveries that improved health for many patients. However, data from many patients with diverse backgrounds and locales are required to make these discoveries possible. This demand for large and diverse EHR data is challenging to satisfy since EHR data at different institutions is not standardized. Initiatives such as All of Us sponsored by the NIH have used the Observational Medical Outcomes Partnership (OMOP) common data model to standardize the formatting of EHR and other data across all institutions that contribute data to the initiative. While OMOP has provided a standard template to which institutions must conform, converting EHR data to OMOP is a time-consuming task requiring technical infrastructure and clinical expertise. Many low-resourced institutions that serve underserved populations lack the time and expertise to convert EHR data to OMOP. These are the very institutions that initiatives such as All of Us need to include.

The 21st Century Cures Act has led to rules requiring EHR companies to make patient data available through application programming interfaces (API). Moreover, these APIs must conform to a data standard called HL7 Fast Healthcare Interoperability Resources (FHIR). Using APIs decreases the infrastructure needed to collect, store, and transform data from the EHR into OMOP. Before we start developing such an application, we must determine whether the data available through the EHR FHIR APIs has sufficient coverage (FHIR data available for many OMOP domains) and completeness (data for a given patient and given domain in OMOP is also in FHIR).

Methods

We randomly selected 100 patients from Vanderbilt University Medical Center who had at least two encounters at least 60 days apart in our 30-month study period of January 1, 2018, to June 30, 2021. For those patients, we extracted information from two sources: 1) the FHIR API REDCap application (called Clinical Data Interoperability Services)(\textsuperscript{1}), and 2) Vanderbilt’s Research Derivative (RD), a curated OMOP-formatted data warehouse.

We compared Measurements between the two data sources, which included OMOP Measurement data and FHIR Observation resources with Labs and Vital Signs profiles. For this analysis, we compared data from both sources at the individual level across our study period, assessing the coverage of structured concepts in each source and identifying any discrepancies in values where applicable. For the review, “Match” indicates both sources have that individual, date, code, and value (unless noted); “No match” indicates the individual, date, and code was not present in both sources.

This study was approved by the Vanderbilt Institutional Review Board as an exempt research study.

Results

Our sample of patients consisted of 19 males age ≤ 18 years, 15 males age 19 to 40 years, 12 males age ≥ 41 years, 20 females age ≤ 18 years, 13 females age 19 to 40 years, and 21 females age ≥ 41 years.

Comparison of the two sources identified several categories of information that differed among the sources. For Measurement data, we identified four major contributors of differences: (1) Data sources captured, (2) Precision of measurements, (3) Granularity of concepts, and (4) Accuracy of concept assignment.

We noted that there were some differences in coverage of Measurements for these sources: eGFRs, Microbiology results, and Blood Type were captured in these FHIR resources, EKGs, high-frequency inpatient vitals, computed BMI, and antibiotic resistance test results were available in the OMOP Measurements table. For (2) above, there were different levels of precision for some values, for instance automatically converted weights had one decimal place in FHIR and two in OMOP.

For (3) and (4), some of the same events were reported with different LOINC codes. This varied from small changes that could be accounted for with the hierarchy, e.g., a Method Dimension with variations including no method, “test
strip”, and “automated test strip”, to moderately impactful changes, e.g., an incorrect Property of Presence instead of Mass/Volume, to harmful changes, e.g., an incorrect Component Dimension.

![Figure 1: Visual representations of measurement agreement](image)

Panel (A) shows the agreement of Measurements on a per-person basis by count. Panel (B) shows the agreement of Measurements on a per-measurement basis by percent.

**Discussion**

These results show promise for integrating data from various sources to support research data interoperability. While Figure 1 indicates some significant mismatches between FHIR and OMOP data at VUMC, using LOINC’s hierarchy and aligning inpatient vitals would address the majority of both codes and individual measures. Specifically, the individuals with 1,000+ measurements are inpatients that had minute-level data not present in our FHIR extract. Identifying the appropriate source for this data in our FHIR implementation would resolve much of the current No matches by count.

The variation in data provided by the FHIR endpoints and OMOP tables reveals some interesting impacts of data siloing in the EHR. The Big Data team at VUMC supports the incorporation of various sources, which has provided the advantage of integration of EKG data, for instance. The data provided by FHIR, specifically Blood Type and the Microbiology results, are more integrated here, though also available through other data flows in the RD. Ensuring that all the expected data is acquired from various FHIR implementations will ideally stabilize over time; for example, EKGs are specified as a part of the United States Core Data for Interoperability (USCDI) V2.

The alignment of concepts used is where there is a clear value add of manual data curation. Some LOINC codes reported in FHIR are deprecated, and these artifacts can be corrected in the framework of the updated codes. While not directly investigated here, older data captured without LOINC codes can also be standardized to support research use cases.

**Funding:** This work is supported by NIH/NLM contract #75N97019P00279 and NLM OD 1 U2C OD023196.

**References**

Pragmatic Design and Application of a Named Entity Recognition Pipeline to Assist Contact Tracers during the COVID-19 Pandemic

John Caskey, PhD; Iain McConnell, PhD; Madeline Oguss, MS; Edwin Nyakoe-Nyasani; Dmitriy Diligach, PhD; Rachel Kulikoff; Brittany Grogan; Crystal Gibson, MPH; Elizabeth Wimmer; Traci E DeSalvo, MS; Matthew Churpek, MD, MPH, PhD; Majid Afshar, MD, MSCR;

University of Wisconsin-Madison, WI; Loyola University Chicago, IL; Public Health Madison Dane County, Madison, WI; Wisconsin Department Health Services, WI

Introduction
As of August 1, 2021, the state of Wisconsin (WI) has confirmed 622,676 cases of SARS-CoV-2 (COVID-19). At the county level, health departments have found that free text fields from the COVID-19 Initial Case Interview (contact tracing) forms contain information to identify potential organizations and locations where transmission of the virus occurred and when individuals were infectious. Public health workers have encountered a high caseload and are overwhelmed with an abundance of free-text information in the interview forms. Current methods to mine the free text fields are manual, without rapid and systematic methods for finding transmission clusters. We employed methods in natural language processing with pre-trained neural language models for named entity recognition (NER) and a novel address search algorithm using the Wisconsin Electronic Disease Surveillance System (WEDSS) to assist in contact tracing efforts and reduce the burden on the health departments.

Methods
WEDSS is a secure, web-based system designed to facilitate reporting, investigation, and surveillance of communicable diseases including data on testing for COVID-19 cases since the outbreak began in February 2020. A total of 281 fields were extracted from WEDSS for our analyses, including 26-character string fields from the county-level data related to information on potential contacts.

For the pipeline development, all 26-character string fields were concatenated into one document with model runs at the case-level. This text was preprocessed to remove non-meaningful annotations such as the contact tracer. The WordPiece tokenizer was then used to build groups of 512 tokens from each document that were fed into the Bidirectional Encoder Representations from Transformers (BERT) neural language model. We employed a prior BERT-base-cased model that was fine-tuned on the English version of the standard CoNLL-2003 Named Entity Recognition (NER) shared task. The pipeline was downloaded from HuggingFace and reports an F1 score of 91.3 on the CoNLL-2003 test dataset. Post-processing of the named entities included the removal of frequently occurring named entities (i.e., “Wisconsin”, “GMT”) identified from 12 months of post-testing case interview forms and removal of duplicate named entities. For all case IDs that had the same named entity reported, the average predicted probability was provided as the score for the likelihood of identifying it as a person, organization, location, or miscellaneous.

Weekly reports were built using the date nearest to contracting the virus or testing positive. The goal of the report was to identify named entities associated with a cluster outbreak. A cluster was defined as two or more cases associated with the same named entity in a seven-day period. Each cluster in the report also included the associated case IDs to guide contact tracers. The NER tool was validated against known outbreaks that were reported in WEDSS. All confirmed and probable cases with known outbreaks in Dane County between July 2020 and June 2021 served as the validation dataset.

Many of the name entities contained common business names that can have multiple locations within a county (i.e., “Culvers,” “Walmart”); therefore, we developed a novel address search algorithm into the pipeline. The latitude/longitude coordinates for each case ID in the cluster were extracted from WEDSS, and a k-means approach was used to identify the centroid coordinates for the cluster of case IDs for a particular name entity. Data from the US Census Bureau for 2018 showed that the commute distances for over two-thirds of the businesses in three major metropolitan areas in Wisconsin were between 0 and 24 miles (https://onhemap.ces.census.gov), and the Google Places API has a 30 km radius as a hard limit when conducting a search. Therefore, we set a search radius of 19 miles from the centroid coordinates, and our location mapping pipeline ran the Google Places API to perform a search for business names matching the named entities within this radius that matched our internal database of organizations. This pipeline provided the most likely business address for the named entity. If an exact match could not be made, then the top three results were filtered using fuzzy string matching. The address results were merged into the report to further aid contact tracers. We selected the month interval of October 2020 as a high case load representation to test the accuracy and precision of the address search algorithm.
Named entities that matched a known outbreak were true positives. If the address search algorithm produced an address match with a named entity but without an outbreak match in WEDSS then it represented a potential novel outbreak. Named entities that were produced without an address match were false positives. The Precision, Recall, and F1 scores were reported for the NER tool and address search algorithm. The Institutional Review Board of the University of Wisconsin approved this study, and a Data Use Agreement was established between the Wisconsin Department of Health Services and the University of Wisconsin - Madison.

Results
The validation dataset was composed of 46,898 probable or confirmed cases of SARS-CoV-2 with 4,183,273 total BERT tokens and 15,051 unique BERT tokens across the free text fields from the case interview forms. The longest field was InvestigationNotes with a median token count of 126.5 (Interquartile Range 67.0-232.5). Across 12 months of data, the median recall and precision were 0.59 (95% CI ±0.06) and 0.46 (95% CI ±0.05), respectively. The NER tool performed best during months with a high volume of cases. In October 2020, the recall and precision were 0.73 and 0.51, respectively. However, by June 2020 when case load dropped, the recall and precision also dropped to 0.43 and 0.36, respectively. A de-identified sample report is shown in the Table demonstrating a match across named entities from the NER pipeline with entities from known outbreaks (outbreak entity) as well as a new fast food restaurant that may be a novel outbreak.

Table. Weekly deidentified report for contact tracers by county

<table>
<thead>
<tr>
<th>Named Entity</th>
<th>Type</th>
<th>Iterations</th>
<th>Score</th>
<th>IncidentIDs</th>
<th>Outbreak Entity</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun prairie</td>
<td>Location</td>
<td>12</td>
<td>0.67</td>
<td>[12345*, 79900*]</td>
<td>sun prairie</td>
<td>sun prairie (1) Local Retailer, 555 Local St, Madison</td>
</tr>
<tr>
<td>local retailer</td>
<td>Organization</td>
<td>7</td>
<td>0.54</td>
<td>[23456*, 79895*]</td>
<td>local retailer</td>
<td>local retailer (1) Big Box Store, 123 Sales St, Middleton</td>
</tr>
<tr>
<td>big box store</td>
<td>Organization</td>
<td>3</td>
<td>0.45</td>
<td>[12345*, 55821*, 60512*]</td>
<td>big box store</td>
<td>big box store (1) Fast Food Place, 100 Food Ave, Madison</td>
</tr>
<tr>
<td>fast food place</td>
<td>Organization</td>
<td>2</td>
<td>0.71</td>
<td>[23456*]</td>
<td>fast food place</td>
<td>fast food place (1) Fast Food Place, 100 Food Ave, Madison</td>
</tr>
<tr>
<td>*restaurant</td>
<td>Organization</td>
<td>1</td>
<td>0.91</td>
<td>[23456*]</td>
<td>restaurant</td>
<td>restaurant (1) Fast Food Place, 100 Food Ave, Madison</td>
</tr>
<tr>
<td>*named place</td>
<td>Miscellaneous</td>
<td>1</td>
<td>0.16</td>
<td>[67111*]</td>
<td>named place</td>
<td>named place (1) Fast Food Place, 100 Food Ave, Madison</td>
</tr>
</tbody>
</table>

Named Entity = result from NER pipeline; Iterations = unique mentions of NER across available case IDs from the reporting period; Score = average predicted probability from the classifier for the type of named entity. IncidentIDs = unique case IDs for lookup by contact tracer; Outbreak entity = known outbreak exposures; Address = matched named entity using longitude/latitude for address from k-means clustering from Google Places API; *Named entities only qualified as cluster outbreaks if they had ≥2 case IDs associated with them.

Output from the address search algorithm was analyzed to assess how many of the outbreaks correctly or incorrectly matched to the named entities. For the month of October 2020, a total of 1,840 named entities were identified. Of the outbreaks recorded in WEDSS (n=519), a match was made between the NER tool named entity and the named entity from the address search algorithm in 78% of the cases. Among the remaining named entities from the NER tool without a match to known outbreaks (n=1321), an additional 73% (n=963) had an address identified as potentially novel named entities to assist contract tracing efforts. The search algorithm had a precision and recall of 0.83 and 0.90, respectively.

Discussion
Our NER tool and address search algorithm were able to extract large amounts of surveillance data and summarize a report to highlight potential outbreaks and their associated addresses. The report was designed in weekly intervals by county so a systematic approach can be shared for any county in the state of Wisconsin. We demonstrated our pipeline performs best during high case volume periods when automated methods for contact tracing efforts are most needed. In addition, our tool has the potential to identify novel cluster outbreaks not identified by conventional methods. We are currently piloting our pipeline with Dane County for prospective validation and to gather qualitative data if any meaningful improvements are made in improving accuracy and saving time during contact tracing efforts.

Our work derives from a statewide database so it may be scaled across counties to assist other health departments in Wisconsin. Our pipeline is also publicly available at https://github.com/disulfidebond/APOLLO for other states to adapt to their workflow. Further, our pipeline may also be applied for other communicable disease and surveillance efforts.

References
Iterative improvement and validation of an algorithmic software application for remote medication management of hypertension

Kira Chaney, MPH¹, Hunter Nichols, PharmD², Naomi DL Fisher, MD², Alexander J Blood, MD³, David Zelle¹, Thomas Venman¹, Jiyeon Shin¹, John Segal¹, John St. Laurent¹, Jacqueline Chasse, FNP-C², Jennifer Crossen, PhamD², William J Gordon, MD, MB¹,²,³, Kavishwar B Wagholikar, PhD, MBBS¹, Benjamin M Scirica, MD² ¹Mass General Brigham, Boston, MA ² Brigham and Women’s Hospital, Boston, MA ³ Harvard Medical School, Boston, MA

Introduction

Professional society guidelines provide guidance on how to treat hypertension, but many patients do not receive guideline-directed medication care which results in wide variation in patient care and missed opportunities to reduce cardiovascular risk¹. Implementation of guideline-based clinical decision support (CDS) tools are challenging due to ambiguity and complexity of the clinical logic described in the guidelines². Inaccuracies in system-generated recommendations lead to alert fatigue and perpetuate variation in care. Our objective was to evaluate the accuracy of the recommendations from the algorithmic CDS software application within an active hypertension medication management program in order to inform strategic changes to the CDS software tool. In December 2018, a document-based algorithm with demonstrated effectiveness in the remote management of hypertension³ was transitioned into a software application previously used in other remote patient management programs. The integrated algorithmic software generated recommendations for medication management for hypertension patients based on the individual patient characteristics and blood pressure goals⁴.

Methodology

One year after initiation, a team of clinicians conducted a comprehensive review of the CDS system logic. This qualitative evaluation of the clinical algorithm specification was performed by the clinical pharmacist, cardiologists, and the endocrinologist. The cardiologists and endocrinologist provided guidance regarding where the algorithm could be adjusted and still be consistent with the guidelines and with appropriate clinical practice.

A secondary analysis of the specific recommendations that were rejected by the clinicians was conducted. The medication recommendation from the CDS application was reviewed by a clinician at each medication step in the algorithm and captured as either accepted or rejected. If rejected, it captured the override recommendation. A subsequent, additional group review of these rejected recommendations further validated the individual decision of the clinician to reject the recommendation.

Results

The comprehensive review of the hypertension algorithm identified 5 sets of initial improvements. Three sets of improvements were completed. The first set of improvements, refining the program contraindications was deployed in April 2020. Some of the key concerns that were identified were insufficient safety stops at the program level for medications that a patient was currently taking when lab values should indicate monitoring. The second set of improvements, refining the calcium channel blocker (CCB) medication class logic, was implemented in June 2020. The group of clinicians identified a need for more dose specific logic for CCBs.

The third set of algorithm functionality improvements, refining the angiotensin-converting-enzyme inhibitor (ACEi) medication class logic and angiotensin receptor blocker (ARB) medication class logic, was completed in November 2020 and these improvements were informed by both the systematic review of the logic as well as by the group review of rejected recommendations.

The accuracy of the algorithm was measured by the percentage of rejected recommendations of the total medication recommendations made by the software. In April 2020, prior to any functionality changes, 54% of the CDS medication recommendations (over the previous 3-month period) were rejected. In the 9-month period of evaluation and iterative improvements in the CDS algorithm, the percentage of rejected recommendations was reduced from 54% to 16% (Figure 1).
Conclusions:

Although the CDS algorithm was based on a treatment framework defined by guidelines, after over a year in use, we found that more than half of the recommendations were not being accepted by a clinician at the time of medication change and required modification to make them more applicable. Effective improvement of this medication management CDS algorithm was only achieved through a process of continuous qualitative evaluation of the algorithm, quantitative measures of its accuracy, and implementation of data driven functionality changes. Iterative, data informed improvements decreased the percentage of rejected recommendations and made the CDS algorithm more accurate. The process for validation and iterative improvement of this hypertension CDS algorithm may be relevant to CDS implementations in other clinical domains. Regardless of whether algorithm functionality improvements can be done in the short term, collection of data around when, why, and how the algorithm can be optimized is key for any future improvements in the accuracy of a CDS algorithm.

References

**Magnified Convolutional Enrichment Representation Model**

Guocai Chen¹, Herbert Chen², Yuntao Yang¹, Abhisek Mukherjee³
Shervin Assassi⁴, Claudio Soto³, W. Jim Zheng¹*

¹ School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX, USA  
² Krieger School of Arts and Sciences, Johns Hopkins University, Baltimore, MD, USA  
³ Department of Neurology, McGovern Medical School, University of Texas Health Science Center at Houston, Houston, TX 77030  
⁴ Department of Medicine, Division of Rheumatology, University of Texas Health Science Center at Houston, USA

**Introduction**

Feature representation and learning process formed the cornerstones of machine learning. Feature representation mathematically characterizes domain entities by encoding input features followed with feature normalization and feature extraction. It may immensely reduce the cost of the learning process and improve the performance. A typical example is the BERT model, a self-supervised language model, which led to significant improvement in almost all natural language processing downstream tasks. In this work, we introduced an advanced model to represent entities with a dynamic mechanism, which was further enhanced by fusing the local and global over-representation on the basis of encyclopedic knowledgebases.

**Methods**

We designed a mathematical model, Magnified convolutional enrichment representation model (MaCER), to evaluate the over-representation [2] of a disease and genes as the controlled vocabulary. The model is introduced in detail in Figure 1. The collection of PubMed abstracts is divided into several proportions with partial overlap, where each proportion will serve as the background knowledge for the analysis of overrepresentation of the terms and the controlled vocabulary. Specifically, the human genes are applied as the controlled vocabulary to represent the human diseases in this project. To ensure the representative and the differential power, we select 60% of the abstracts for each proportion, although it is presumably not sensitive to the model. The number of effective iterations vary from term to term (aka. the diseases) and the elements of the vocabulary (aka. the genes) to another, for example, the number of effective iterations for Breast cancer and GHST or PRR12 are 30 or 34 respectively, while it is 9 for Dentin dysplasia, type II and DSPP. The non-overrepresented proportions are referred to as ineffective iterations and discarded. The model mimics the rapid development of the PubMed database. With the

**Figure 1.** Magnified convolutional enrichment representing model. Phase 1. The over-representation between a disease and a gene is evaluated on a number of subsets of PubMed citations, the p-values for the subsets are then adjusted with false discovery rate Benjamini-Hochberg (*fdr_bh*) [1], the median of the adjusted p-values is selected to represent the enrichment between the disease and the gene. Phase 2. The feature set of the human diseases are processed with deep denoising Autoencoder for feature extraction and dimension deduction. Phase 3, the extracted features are fused with the word embedding for human diseases.
dynamic mechanism, the overfit problem and aggregation effect may effectively be eliminated. Furthermore, by leveraging the contexture information with the word embedding [3] and the global enrichment information, the model is adopted to represent the complex human diseases.

Results

In this project, PubMed (ncbi.nlm.nih.gov) is used as the background knowledge. The citations in PubMed were collected by Sep 7, 2020, which contains 41,010,490 abstracts. The mapping of gene to PubMed (gene2pubmed, ncbi.nlm.nih.gov) dated on Sep 8, 2020, contains 12,575,651 tuples, where 1,518,029 maps labeled for human genes are used in this project. The total number of human genes labeled in the NCBI database is 38,671.

With the Human disease network [4] as the gold standard dataset, The diseases we extracted from the Human disease network are encoded with Fingerprint model, the Convolutional enrichment representation model, and the MaCER model, and are compared to other existing methods, including glove [5], word2vec-chow [6], fasttext [7], and word2vec-1gram [6]. The 10-folder cross-validation results are measured with accuracy, recall, precision, and F1-score. As shown in Figure 2, MaCER has the best performance among these approaches.

Discussion

In this work, we introduced a novel Magnified convolutional enrichment representation model to model the entities in an encyclopedic background knowledge base to integrate the local and global encoding. The model was evaluated with the Human disease network to predict the disease-disease associations. The performance of the MaCER model is up to 97.8% on the Human disease network, which demonstrated the good fitness of the MaCER model for predicting the associations of complex diseases and the advantage of the MaCER model over the sole embedding models or Convolutional enrichment representation model. The MaCER model is demonstrated based on the most significant biomedical knowledgebase. The statistical nature of the model implies that sufficient resources of information are a necessity to ensure the performance. On the contrary, with a prolific knowledge, for example, stock market database, clinical database, an event model built with MaCER will be a good performer.

Acknowledgement

This work is partly supported by the Cancer Prevention and Research Institute of Texas through grant RP170668 (Zheng), and the National Institutes of Health (NIH) through grants 1UL1TR003167 and R01AG066749 (Zheng). We also would like to thank Texas Advanced Computing Center (TACC) for providing the computing resource.

Leveraging novel ePRO and existing EHR datamarts to assess clinical outcomes in ambulatory oncology

Jessica L. F. Cleveland, MSOR¹, Sunyi Zhang MS¹, Sherry Lee¹, Christine Cronin¹, Jason Johnson¹ PhD, John Orechia¹, Deb Schrag MD MPH¹,², Michael J. Hassett MD MPH¹,², Nadine J. McCleary, MD MPH¹,²

¹Dana-Farber Cancer Institute, Boston, MA; ²Harvard Medical School, Boston, MA

Introduction

Studies have demonstrated that the collection of electronic Patient Reported Outcomes (ePROs) improves quality of life, reduces ED visits/hospitalizations, and prolongs survival. Although full integration of ePROs into the electronic health record (EHR) has been of great interest recently in the ambulatory oncology setting, methods and tools needed to support the systematic extraction, validation, and analysis of novel ePRO data in the EHR are limited. Leveraging ePRO data enables operational and research teams to a) direct quality improvement (QI) efforts, b) assess patient questionnaire burden, c) drive population health management and d) use ePROs as part of alternative payment models. Previous work has demonstrated that the use of audit and feedback can be a successful mechanism for source data validation and ad-hoc analysis of ePRO data. Here we describe the process of coalescing novel ePRO datamarts and analyzing ePRO data combined with additional demographics and healthcare utilization datamarts to enable assessment of clinical outcomes and quality measures in an ambulatory oncology setting at one NCI-designated cancer center.

Methods

To start, ePRO data were curated as part of a questionnaire datamart that followed a one-level ‘root’ and “branch” structure (Figure 1). Data validation of the ePRO responses included an assessment of missing data and out of range or inappropriate values. ePRO data included 15 symptom domains leveraging the PRO-CTCAE tool, PROMIS-10 global health, Social Determinants of Health, and more. The designated data stewardship group then identified two additional source datamarts that were integral to the analysis of ePRO data: demographics and healthcare utilization. Curation of demographics data included patient level details such as birthdate, primary language, race, ethnicity, sex at birth, and self-identified gender. The healthcare utilization datamart included encounter type, time, department and provider (Table 1). A previously curated treatment plan datamart was also used in the resultant systematic analysis of outcomes. Leveraging the treatment plan, ePRO, demographics, and healthcare utilization datamarts, we were able to compute the number of telephone encounters within 30 and 60 days and the number of emergency department visits or hospitalizations within 30 and 60 days for all patients with ePRO responses. Figure 2 outlines an example of how clinical outcomes can be derived relative to treatment start and made available.

Results

The pragmatic ePRO datamarts and resultant clinical outcomes data have been leveraged by both research and operational teams. Since becoming available across the ambulatory center in May 2020, the number of operational and research data requests and subsequent dissemination through presentations, publications and QI initiatives has been used to ascertain the impact of the datamarts. To date, there have been over 30 data requests made by clinical teams at Dana-Farber Cancer Institute for operational and research purposes. From these requests, more than 10 research abstracts and publications have been submitted or are in progress. Quality improvement initiatives focused on ePRO response rates and demographics data collection have been piloted, and alternative payment contracts have been negotiated with the institution’s two largest payers.

Discussion

ePRO data are novel and complex, and therefore required special consideration to effectively source, validate, steward, and analyze. Due to the time-dependent nature of treatment plans, ePROs, and outcomes, it is critical to curate and collate key data in this case ePROs with demographics and healthcare utilization. Demographics data showed significant gaps in data collected, triggering efforts to improve how this information is gathered over the course of care to fully understand gaps in health equity. Prior to this effort, ePRO data existed in silos and were not accessible or explorable. This work demonstrates the ability to enable translation of evidence-based ePRO practice by making the data useful and usable.
Illustrations

<table>
<thead>
<tr>
<th>ePRO Questionnaire Structure</th>
<th>ePRO Questionnaire Datamart Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>None &gt; 1</td>
<td>Root</td>
</tr>
<tr>
<td></td>
<td>Branch</td>
</tr>
<tr>
<td></td>
<td>Question</td>
</tr>
<tr>
<td>Frequency Pain?</td>
<td>ePRO</td>
</tr>
<tr>
<td>Severity Pain?</td>
<td>Pain</td>
</tr>
<tr>
<td>Interference Pain?</td>
<td>Frequency</td>
</tr>
<tr>
<td>Frequency Fatigue?</td>
<td>ePRO</td>
</tr>
<tr>
<td>Severity Fatigue?</td>
<td>Pain</td>
</tr>
<tr>
<td>Interference Fatigue?</td>
<td>Severity</td>
</tr>
</tbody>
</table>

Figure 1. Questionnaire datamart one-level ‘root’ and ‘branch’ datamart structure

<table>
<thead>
<tr>
<th>Questionnaire Datamart</th>
<th>Demographic Datamart</th>
<th>Utilization Datamart</th>
</tr>
</thead>
<tbody>
<tr>
<td>Root Questionnaire Name</td>
<td>Birth Date</td>
<td>Encounter Type</td>
</tr>
<tr>
<td>Branch Questionnaire Name</td>
<td>Primary Language</td>
<td>Appointment Time</td>
</tr>
<tr>
<td>Question</td>
<td>Race</td>
<td>Admit Time</td>
</tr>
<tr>
<td>Answer</td>
<td>Ethnicity</td>
<td>Discharge Time</td>
</tr>
<tr>
<td>Response Status</td>
<td>Sex at Birth</td>
<td>Encounter Department</td>
</tr>
<tr>
<td>Response Date/Time</td>
<td>Self-identified Gender</td>
<td>Encounter Provider</td>
</tr>
</tbody>
</table>

Figure 2. Example of systematic longitudinal outcomes

References


Variability in EHR Data About Race and Ethnicity as Observed in the National COVID Cohort Collaborative Data Enclave

Lily Cook, MA1, Juan Espinoza, MD2, Nicole G. Weiskopf, PhD1, Nisha Mathews, PhD3, David Dorr, MD4, Kelly L. Gonzales, PhD4,5,6, Adam Wilcox, PhD7, Charisse Madlock Brown, PhD8

1Oregon Health & Science University School of Medicine, Department of Medical Informatics and Clinical Epidemiology, Portland, OR; 2Department of Pediatrics, Children's Hospital Los Angeles, CA; 3Associate Degree Nursing Department, San Jacinto College Central, Pasadena, TX; 4Enrolled Citizen of the Cherokee Nation; Oregon Health & Science University-Portland State University joint School of Public Health, Portland, OR, USA; 5BIPOC Decolonizing Data Council; 6Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA; 7Tennessee Clinical and Translational Science Institute, University of Tennessee Health Science Center, Memphis, TN

Abstract

We analyzed data from 6.5 million patients tested for COVID-19 contributed by 56 healthcare institutions and found that “No matching category” was the second largest harmonized racial group. We evaluated whether this data conformed to the federal standards for patient information about race and ethnicity. 20.7% of the data about race met our definition of non-conformance; the largest category was data that was missing. Hispanic patients were over-represented in the non-conforming racial data.

Introduction

Our objective was to explore variations in how healthcare systems collect and report information about the race and ethnicity of their patients. To this end, we sought to assess the quality of ethnicity and race EHR data as reflected in National COVID Cohort Collaborative (N3C) Data Enclave; we focused on conformance to the standard federal definitions, as well as missingness, and misclassification. A significant technical challenge related to integrating race and ethnicity data across EHR systems is the lack of consistency in how data about race and ethnicity is collected and structured by healthcare organizations. The N3C provides a unique opportunity to assess how different healthcare systems collect and conceptualize information about their patients’ race and ethnicity, and to examine efforts to integrate these categories across different data models. Because the COVID-19 pandemic disproportionately impacts communities of color and exacerbates existing health inequities, the accurate identification of these cohorts is crucial to understanding these disparities.

Methods

The N3C Enclave contains health records from 6.5 million patients tested for COVID-19 from 56 healthcare institutions and networks across the country. The data have been harmonized from four Common Data Models (CDMs): ACT, PCORnet, and TriNetX CDMs have been mapped to OMOP, while institutions using OMOP have their data ingested directly. We defined race and ethnicity data as “conforming” if they had been mapped to one of the five minimum categories for race congruent with the federal standard: (a) American Indian or Alaska Native, (b) Asian, (c) Black or African American, (d) Native Hawaiian or Other Pacific Islander, and (e) white. Ethnicity data were conforming if they were harmonized into one of the two standard categories for ethnicity: “Hispanic or Latino” or “Not Hispanic or Latino”. For the purposes of collecting and organizing patient demographic data, race and ethnicity are considered distinct concepts, and ethnicity refers only to Hispanic or Latino origin. We used a multi-step process that included data processing, terminology harmonization, and descriptive analyses. We first sorted the values in the harmonized, mapped race field into primary race categories, then aggregated each category by source values and tabulated these results. Finally, we cross tabulated the resulting categories by three factors: patient ethnicity, the number of data partners using each code, and which data models utilized those particular encodings.

Results

There are a total of 25 harmonized categories for race available in the N3C Data Enclave. The top three categories--white, “No matching concept,” and Black or African American-- account for 96.97% of the data. No patients with
race of American Indian or Alaska Native were found (as per request of the NIH Tribal Health Research Office); during ingestion these data were relabeled as being of race “other.”

We found that “No matching category” was the second largest harmonized racial group in the N3C, and that 20.7% of the data about race met our definition of non-conformance. As shown in Figure 1, non-conforming racial data could be grouped into seven categories: Missing, Other, Refused, Multi-racial, Misclassified (i.e., incorrectly mapped), Uninterpretable encodings, and Granular data.

The most common reason for race data to be non-conforming was because it was missing from the record (n = 384,089). This was found in data from 31 of the contributing healthcare institutions and reported 29 distinct ways. The second largest category of non-conforming race data was patients labelled as “Other” race. The majority of these (61.0%, n = 217,476) were of Hispanic or Latino ethnicity. Patients who declined, or refused, to answer questions about race represented 8.0% (n = 107,445) of the non-conforming data. Multiracial patients represented 7.4% of the non-conforming data although they were 1.5% of all the patients in the N3C Data Enclave. Of the 257 different codes used by systems to represent race, 119 of them were distinct codes used to represent multiracial patients. Examining the source data revealed that 3.1% of the non-conforming race data, 0.6% of all data in the N3C, appears to have been incorrectly mapped, or misclassified. Although only 14.8% of the patients in the Enclave are Black or African American, 47.1% of these misclassified patients should have had their race mapped to Black or African American. For 18,738 patients, the source institution had provided a code such as “@” or “z” that were specific to certain sites and did not conform to those recognized by any of the known data models. Data from 2,477 patients did not map to one of the five categories because they were represented by a more granular racial subcategory that had not been rolled up into one of the five main race categories. There were nine granular subgroups of race available in the non-conforming data; the largest was Asian Indian (n = 1,534). For patients whose ethnicity was Hispanic or Latino, “non-conforming” was the single largest racial category (51.7%; n = 399,831).

Discussion

Our analyses of the N3C Enclave revealed a number of facts that are important for researchers to consider when drawing conclusions based on these data, or data from similar large-scale data resources. By delving into the N3C data ingestion process, we were able to identify specific points at which the sites could potentially improve adherence to the standard when preparing their data. Finally, the impact of these data quality issues was not equal across all races and ethnicities, which has the potential to introduce substantial bias in analyses and conclusions drawn from these data. We hope to utilize these results to prevent errors and to identify ways to correct the race and ethnicity data post-ingestion.

References
1. Hong SS. N3C Data Ingestion and Harmonization Workstream Race Mapping Crosswalk [Internet]. National COVID Cohort Collaborative; 2020 [cited 2021 Aug 23].
2. HL7 International. Structure Definition: US Core Race Extension [Internet]. [cited 2021 Aug 23].
Identifying Traumatic Brain Injury in Females from Electronic Health Records: Considerations for Defining a Cohort

Bernadette A. D’Alonzo, MPH1, Rebecca Morse1, Abigail C. Bretzin, PhD, ATC1, Andrea L.C. Schneider, MD, PhD2, Silvia Canelón, PhD1, Mary Regina Boland, MA, MPhil, PhD, FAMIA.1,3

1Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, 2Department of Neurology, University of Pennsylvania, Philadelphia, PA, 3Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA.

Abstract
This study investigates different case definitions for female patients with traumatic brain injury (TBI) using electronic health records (EHR). Survival time differed by TBI case definition (p<0.0001). Patients meeting DOD/VA TBI definition alone demonstrated longest survival, followed by those meeting CDC-only definition. Patients who met criteria for TBI using DOD/VA + CDC definitions together had shortest survival. Findings have implications for informaticians; case definition can greatly alter patient populations retrieved for TBI in the EHR.

Introduction
Traumatic Brain Injury (TBI) is an acute injury, defined as an alteration in brain function as a result of mechanical energy transmitted to the head from external forces1. TBI is prevalent in the United States (US), with an estimated 1.5-2 million Americans experiencing a TBI annually, roughly 80% of whom are treated in the Emergency Department (ED)2. TBI is a significant public health concern, and yet, our understanding of its epidemiology is limited. Surveillance of TBI has focused mainly on athletes and combat veterans3-5 resulting in differences in definitions and classifications of TBI. Since both of these populations are predominantly male, our understanding of the incidence of TBI among females is limited. Applying these male-centric (androcentric) TBI classification systems to females might reveal differences in terms of the quality of the existing TBI classification systems. These challenges may result in patients with TBI being missed, misunderstood or misdiagnosed6. Yet, electronic health record (EHR) data remains an underutilized resource for better understanding TBI in the community setting. Both the US Centers for Disease Control and Prevention (CDC) and Department of Defense (DOD)/Veterans Affairs (VA) have developed their own surveillance case definitions of TBI in the EHR using ICD-9 and ICD-10 code sequences7. The CDC code sets were designed for Emergency Admissions to estimate the rate of TBI, and they exclude patients seen outside the hospital setting, such as primary care, urgent care, and specialty clinics8. CDC reports of TBI have historically not presented results stratified by sex; more recently, TBI statistics appear skewed-male. The DOD/VA surveillance case definition defines codes in terms of TBI severity (mild, moderate, severe, penetrating, unclassifiable), however, these have been validated in the military health system, whose primary patient population (predominantly male) may not reflect the community setting. Therefore, a reexamination and potential broadening of these definitions to consider females with TBI is warranted. Our hypothesis is that when using EHR data to describe TBI in the community setting, neither CDC nor DOD/VA case definitions alone may fully capture and elucidate female patients with TBI, and so it may be prudent to examine both. The purpose of this study was to compare and highlight potential differences in the case definitions for TBI from the CDC and DOD/VA in the context of a cohort of females from the EHR at Penn Medicine.

Table 1. Patient characteristics by TBI case definition (n=4466).

<table>
<thead>
<tr>
<th>Total cohort</th>
<th>CDC*</th>
<th>DOD/VA*</th>
<th>CDC+DOD/VA*</th>
<th>Unclassified*</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient age (median, IQR)</td>
<td>55 (36, 70)</td>
<td>43 (27, 58)</td>
<td>51 (31, 68)</td>
<td>55 (37, 70)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age group (%)</td>
<td>420 (50)</td>
<td>652 (50)</td>
<td>729 (56)</td>
<td>677 (51)</td>
<td></td>
</tr>
<tr>
<td>&lt;55</td>
<td>420 (50)</td>
<td>679 (71)</td>
<td>729 (56)</td>
<td>677 (49)</td>
<td></td>
</tr>
<tr>
<td>≥55</td>
<td>420 (50)</td>
<td>284 (29)</td>
<td>562 (44)</td>
<td>695 (51)</td>
<td></td>
</tr>
<tr>
<td>Race/Ethnicity (%)</td>
<td>26 (3.1)</td>
<td>21 (2.9)</td>
<td>41 (3.2)</td>
<td>33 (2.4)</td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>210 (25)</td>
<td>235 (24)</td>
<td>301 (23)</td>
<td>547 (40)</td>
<td></td>
</tr>
<tr>
<td>Black/African American</td>
<td>23 (2.7)</td>
<td>39 (4.0)</td>
<td>53 (4.1)</td>
<td>39 (2.8)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>24 (2.9)</td>
<td>31 (3.2)</td>
<td>61 (4.7)</td>
<td>56 (4.1)</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>557 (65)</td>
<td>637 (65)</td>
<td>835 (65)</td>
<td>697 (51)</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>17 (2.0)</td>
<td>26 (2.7)</td>
<td>33 (2.6)</td>
<td>49 (3.6)</td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>233 (28)</td>
<td>223 (23)</td>
<td>527 (41)</td>
<td>409 (30)</td>
<td></td>
</tr>
<tr>
<td>Accident (%)</td>
<td>77 (9.2)</td>
<td>853 (89)</td>
<td>693 (54)</td>
<td>378 (28)</td>
<td></td>
</tr>
<tr>
<td>TBI Severity (%)</td>
<td>11 (1.3)</td>
<td>96 (10.0)</td>
<td>544 (42)</td>
<td>92 (6.7)</td>
<td></td>
</tr>
<tr>
<td>Mild</td>
<td>1 (0.1)</td>
<td>6 (0.6)</td>
<td>47 (3.6)</td>
<td>45 (3.3)</td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>751 (89)</td>
<td>8 (0.8)</td>
<td>7 (0.5)</td>
<td>857 (62)</td>
<td></td>
</tr>
<tr>
<td>Severe/Penetrating</td>
<td>32 (3.8)</td>
<td>23 (2.4)</td>
<td>80 (6.2)</td>
<td>172 (13)</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>31 (3.7)</td>
<td>47 (4.9)</td>
<td>239 (20)</td>
<td>288 (21)</td>
<td></td>
</tr>
<tr>
<td>Patient class (%)</td>
<td>869 (96)</td>
<td>916 (95)</td>
<td>1038 (80)</td>
<td>1084 (79)</td>
<td></td>
</tr>
</tbody>
</table>

*Categories are mutually exclusive: CDC=CDC-only, DOD/VA=DOD/VA-only, CDC+DOD/VA=CDC and DOD
Methods
We obtained EHR data for 1,060,100 female patients with inpatient or outpatient clinic visits to Penn Medicine from 2010-2017. We extracted patient demographic information from encounter records. We used International Classification of diseases version 9/10 (ICD-9/10) codes from both CDC and DOD/VA surveillance case definitions to identify patients with TBI. We applied the DOD/VA TBI severity matrix (mild, moderate, severe/penetrating, unknown) to determine and assign TBI severity. We applied a decision-tree algorithm of TBI date to extract the most severe TBI diagnosis and labelled that as TBI severity level across all visits. We took the first date of TBI as date of the most severe diagnosis. We examined demographic characteristics among females who met definitions for TBI in three groups: 1) CDC-only, 2) DOD/VA-only, 3) CDC and DOD/VA. Patients with TBI that did not meet either CDC or DOD/VA definition were listed as 4) Unclassified (generally includes ICD-9/10 codes used in the literature for TBI not included in either of the two other definitions). We conducted crude survival analysis with outcome time to death (and survival time restricted to 5 years) and inspected Kaplan-Meier curves with log-rank tests to test for differences among the three groups.

Results
We identified 4466 patients with TBI at Penn Medicine between 2010 and 2017 with information on mortality. There were significant differences across TBI case definitions for each demographic characteristic (age, race, mechanism of injury coded as accident, severity, death status, inpatient/outpatient visit), except Hispanic ethnicity (p=0.2) (Table 1). Log-rank tests revealed that survival time, on average, significantly differed across patients by source of case definition (p<0.0001). Patients who met DOD/VA surveillance case definition only demonstrated the longest survival time on average, followed by CDC definition only, while patients who met both appeared to have significantly shorter survival time, on average. Patients who met neither surveillance case definition appeared to have the shortest average survival time (Figure 1).

Conclusion
Demographic characteristics and time to survival differed significantly among female patients identified with TBI according to surveillance case definition. The choice of using either CDC or DOD/VA definition of TBI can significantly affect the cohorts derived. Therefore, care must be taken when extracting patients with TBI from EHRs. Future work will involve elucidating differences between the CDC and DOD/VA TBI definitions and these cases, which may inform revisions and expansions of current case definitions of TBI in the community setting. This work is important to informaticians to understand that case definition can greatly alter the patient populations retrieved for TBI in a large EHR system. In addition, it lays the foundation for next studies to examine whether these male-focused TBI definitions may need to be recalibrated for use on female-only populations.

References
Machine Learning for Developing a Prediction Model of Hospital Admission of Emergency Department Patients: Hype or Hope?

Anne A.H. de Hond, MSc¹, Wouter Raven, MD¹, Laurens Schinkelshoek, Bsc¹, Menno Gaakeer, MD², Ewout ter Avest, MD, PhD³, Ozcan Sir, MD⁴, Heleen Lameijer, MD, PhD⁵, Roger A.P.A. Hessels, MD⁶, Resi Reijnen, MD⁷, Evert de Jonge, MD, PhD¹, Ilse Kant, PhD¹, Ewout Steyerberg, PhD¹, Christian Nickel, MD⁸, Bas de Groot, MD, PhD¹

¹Leiden University Medical Center, Leiden, the Netherlands; ²Adrz Hospital, Goes, the Netherlands; ³University Medical Center Groningen, Groningen, the Netherlands; ⁴Radboud University Medical Center, Nijmegen, the Netherlands; ⁵Medical Center Leeuwarden, Leeuwarden, the Netherlands; ⁶Elisabeth-TweeSteden Hospital, Tilburg, the Netherlands; ⁷Haaglanden Medical Center, The Hague, the Netherlands; ⁸University Hospital Basel, Basel, Switzerland

Introduction

Emergency department (ED) crowding is a well-known problem affecting the quality of care and patient safety [1]. Long ED length of stay (LOS) is associated with reduced patient satisfaction, negative effects on staff, and poorer patient outcomes, including increased in-hospital mortality [2]. ED patients who ultimately need to be admitted contribute disproportionately to the occurrence of crowding [3].

Our recently published study [4] had two aims. First, we investigated whether machine learning (ML) models could predict hospitalization of ED patients more accurately than logistic regression. Second, we investigated the trade-off between the potential to improve the predictive performance of the models when including more variables and the potential to reduce time to decision-making by developing models at triage, at ~30 min (when vital signs are available) and ~2 hours (when blood test results are available).

Methods

We analyzed consecutive ED patients of three hospitals between 1 January 2017 and 31 December 2019 using the Netherlands Emergency Department Evaluation Database (NEED, for more information, see www.stichting-need.nl). Hospital admission was defined as admission to a normal ward, admission to a medium care or coronary care unit, transfer to another hospital, admission to an intensive care unit, and the patient dying in the ED. The remaining cases were categorized as the patient being discharged. Demographics, urgency, presenting complaints, disease severity and proxies for comorbidity, and complexity were used as covariates. We developed prediction models for hospitalization using an increasing number of data available at triage, ~30 minutes (including vital signs) and ~2 hours (including laboratory tests) after ED registration, using ML (random forest, gradient boosted decision trees, deep neural networks) and multivariable logistic regression analysis (including spline transformations for continuous predictors). Performance was based on discrimination (Area Under the Receiver Operating Characteristic Curve also AUC) and calibration. Confidence intervals (CI) were calculated via bootstrapping. Shapley Additive exPlanations (SHAP) were added to estimate the importance of individual variables for the model predictions.

Box 1. Calculations of mean theoretical time to decision making

The theoretical mean (relative) time to decision making reduction was based on the number of patients in the test data receiving an actionable decision (admitted or sent home) according to best performing model (XGBoost).

A patient received an actionable decision from the model when: P(hospitalization) > 95% PPV threshold or P(hospitalization) < 95% NPV threshold.

Mean time to decision making (TDM) and mean relative TDM reduction in minutes are calculated as:

mean(TDM patient – TDM patient model) and mean(100x(TDM patient – TDM patient model) / TDM patient).

TDM patient model is set to 15 minutes (triage model), 30 minutes (30-minute model), or 2 hours (2-hour model) for patients with an actionable decision. TDM patient model is set to TDM patient when the patient did not receive an actionable decision.
To assess the potential clinical value of these models, we calculated the Mean theoretical reduction in time to decision making based on thresholds corresponding to the 95% positive and negative predictive value (for the calculations see Box 1).

Results

We included 172,104 ED patients of whom 66,782 (39%) were hospitalized. The patients were on average 50 years old and 48% was female.

The AUC of the multivariable logistic regression model was 0.82 (0.78-0.86) at triage, 0.84 (0.81-0.86) at ~30 minutes and 0.83 (0.75-0.92) after ~2 hours. The best performing ML model over time was the gradient boosted decision trees model with an AUC of 0.84 (0.77-0.88) at triage, 0.86 (0.82-0.89) at ~30 minutes and 0.86 (0.74-0.93) after ~2 hours. The best performing ML model performed similarly to the logistic regression model. The calibration of all models was generally excellent, with calibration slopes close to 1. The strongest predictors based on SHAP values were age and treating specialty. For the model at triage, a Mean theoretical time to decision-making reduction of 33 minutes (25%) could be realized based on both thresholds across the whole population. At the 30-minute time point, this increased to 40 minutes (26%), which fell back to 31 minutes (21%) at the 2-hour point.

Conclusion

Our study showed that machine learning models had an excellent but similar predictive performance as the logistic regression model for predicting hospital admission. In comparison to the 30-minute model, the 2-hour model did not show a performance improvement. In line with recent correspondence on this study [5], we intend to further investigate when and how to incorporate interpretability in prediction models that will be implemented in clinic care. Further, we aim to investigate how to introduce these models (interpretable or not) in a responsible way to clinical practice, supporting the intended use.

References

1. Van Der Linden, M.C., M. Khursheed, K. Hooda, J.M. Pines, and N. Van Der Linden, Two emergency departments, 6000km apart: Differences in patient flow and staff perceptions about crowding. Int Emerg Nurs, 2017. 35: p. 30-36.
Assessing Clinical Site Readiness for EHR-to-EDC Data Collection

Eric L. Eisenstein, DBA1, Meredith N. Zozus, PhD2, Maryam Y. Garza, PhD, MPH, MMCi3, Holly J. Lanham, PhD2, Bhargav S. Adagarla, MS1, Felicia L. Graham, MBA1, Christie Milleson1, Anita Walden, MS,4, Kanecia O. Zimmerman, MD, MPH1, Karan R. Kumar, MD, MS1

1Duke Clinical Research Institute, Durham, NC; 2University of Texas Health Science Center at San Antonio, San Antonio, TX; 3University of Arkansas for Medical Sciences, Little Rock AR; 4Oregon Health & Science University, Portland, OR

Introduction
eSource software is used to transfer patient electronic health record data into clinical study databases. While there has been considerable interest in this information technology, the actual number of eSource studies conducted to date is quite small. A recent systematic review found only 14 clinical studies that used direct EHR data extraction through eSource. Eight of the 14 studies were conducted at a single site with a single EHR. Four of the remaining six multi-site studies were part of the same European pilot study. Thus, while there is some evidence of eSource use in sites with sophisticated information technologies, there is little evidence for sponsors to use in selecting sites for multi-center eSource studies. We conducted a study to determine whether it is possible to identify clinical research site characteristics that would make eSource use feasible for clinical studies.

Methods
Investigators from the University of Arkansas for Medical Sciences (UAMS), the University of Texas Health Science Center at San Antonio (UTHSCSA), and the Duke Clinical Research Institute (DCRI) developed the Site EHR-to-eCRF eSource Questionnaire for assessing a clinical site’s readiness to participate in an eSource study. The eSource site readiness survey contains three components that are completed by site clinical research coordinators (CRCs), principal investigators (PIs), and informatics leadership, such as chief research information officers or organizational informatics leaders (CRIOs). This work was funded by the National Institute of Child Health and Human Development (HHSN275201800003I) and the National Center for Advancing Translational Sciences (U24TR001608), approved by the Duke University Medical Center Institutional Review Board (Pro00102679), and is submitted on behalf of the Best Pharmaceuticals for Children Act – Pediatric Trials Network Steering Committee.

The site readiness survey is divided into eight sections that evaluate different aspects of site readiness to utilize eSource software in multi-center clinical studies. All respondents (CRC, PI, and CRIO) complete a set of core questions with additional questions addressed to one or more respondent types. Total questions differed by role: CRC (n=318), PI (n=191), and CRIO (n=136). Pediatric Trial Network (PTN) leadership sent an email to study site PIs describing the present study and soliciting their involvement. Study sites received $833 for each respondent ($2500 maximum per site). There was no direct compensation for individual survey respondents. After each survey was completed, the study team reviewed the responses and contacted respondents to schedule a follow-up interview to clarify their survey responses and obtain additional information that would be used in qualitative analyses. We used elements from the Organizational Readiness To Change Assessment (ORCA) to highlight factors that could support or hinder clinical research site eSource adoption.

Results
Surveys were completed by 57 respondents (CRC, 21; PI, 18; and CRIO, 18). The average respondent experience in their role at their present institution was greater than 7 years (CRC, 7.3 years; PI, 12.7 years; and CRIO, 7.9 years). Sites participated in more than ten clinical trials during the previous 12 months, on average (CRC; 12.1 clinical trials and PI, 12.4 clinical trials), with most having participated in at least one pediatric network study (CRC, 81%; PI, 89%).

CRCs reported that, in the previous 12 months, 54.7% of their study data was first documented in their EHR (considered the source by regulators). However, the percent of data first documented in the EHR varied greatly by site. Nonetheless, the amount of study data first documented in EHRs effectively sets a limit on the amount of participant data that can be transferred from the site’s EHR to a study’s EDC without changing clinical site data collection procedures. CRCs and PIs ranked medication administration, medication order, laboratory and vital signs
data types as having the highest priority for automation (ranked 1 through 4) and as required for inclusion in any eSource solution (also ranked 1 through 4). Medication orders and medication administration also required the most data collection time.

Many eSource technologies access EHR data via FHIR (HL7 Fast Healthcare Interoperability Resources). However, only 44% of site EHRs were FHIR enabled, with 22% developing this capability. Similarly, only 22% of sites were using FHIR to exchange patient data with other institutions for routine patient care, with 33% developing this capability. Lastly, 44% of sites required that software requesting data from an EHR FHIR server be installed behind the institution’s firewall. Given that many eSource solutions are hosted by a sponsor or third-party, this policy could inhibit eSource utilization.

Most respondents were affiliated with organizations that used EHR research functions (CRC, 76%; PI, 72%; and CRIO, 89%). However, the research functions used differed by respondent type and institution. Generally, CRIOs reported greater EHR research function availability than did CRCs or PIs. For example, only 45% of CRCs used data extracted from their EHR to assess study feasibility; whereas, 89% of CRIOs said this function was available. Similarly, only 33% of CRCs used EHR decision support to find potential study patients; whereas, 67% of CRIOs said this function was available. Thus, while having EHR research functions is important when evaluating sites for eSource studies, having CRCs and PIs who have actually used those functions also is important.

Generally, opinion leaders, clinical investigators, and research staff were perceived as supportive of change; whereas, senior leadership/clinical management and IT staff members were less supportive. Instances where senior leadership/clinical management had neutral or unfavorable assessments typically occurred in organizations where the clinics/hospitals were not owned by the medical school or where the organization did not have a CTSA (Clinical and Translational Science Award). In these cases, senior leadership/clinical management tended to focus on patient care and had less interest in supporting clinical research. CRC and PI perceptions of IT staff members typically differed depending upon how their organization’s health IT and EHR support groups were organized. When clinical research was supported by a centralized IT group that focused on patient care and business systems, PI and CRC perceptions were less favorable. However, when there was a separate IT research support group, PI and CRC perceptions were generally positive. This was the case when IT research support was limited to assisting study teams in setting up EHR research functions for specific studies or creating REDCap databases for investigator-initiated studies.

Discussion

Site readiness to participating in eSource studies is not merely the technical problem of extracting study data from a site’s EHR. Computer capabilities clearly are important. However, whether the organization prioritizes clinical research and the site study team’s previous use of specific EHR clinical research functions are equally important considerations.

References


565
Identifying Patients with Interstitial Lung Disease in Electronic Health Records: Development and Validation of Machine Learning Algorithms

Erica Farrand¹, Oksana Gologorskaya², Lakshmi Radhakrishnan², Hunter Mills², Atul Butte²

¹Department of Medicine, University of California San Francisco, CA; ²Bakar Computational Health Sciences Institute, University of California San Francisco, CA.

Introduction: Patient cohort identification is an essential, time-consuming, and costly first step in clinical research studies, particularly in rare diseases. Increasingly investigators leverage large digital health databases and data science to make the process more constructive, efficient, and affordable. Successful computationally enhanced models have been tested in high prevalent diseases. However digital health data stands to transform the study of rare diseases, for which there are critical evidence gaps and challenges identifying patients for clinical trials. We evaluated whether a novel computational model could be developed to support cohort identification in interstitial lung disease (ILD), a rare, complex lung disease for which prior algorithms have relied exclusively on diagnostic codes with limited success.

Methods: We leveraged an electronic health record (EHR) database, known as the de-identified Clinical Data Warehouse (CDW) within the UCSF Information Commons. During our study period (1982-2020) there are over 5 million distinct patients in CDW and data extending across 2334 data domains covering diagnoses, medications, procedures, flow sheets, inpatient, outpatient, and emergency department visits. We used the UCSF ILD database, a registry of ILD patients to train the models. A screening EHR cohort was derived by identifying adult patients with a minimum of 5 clinical encounters and therefore sufficient EHR data to inform the ILD classification model (~1 million adult patients). The training dataset was constructed from 3,500 labeled patients from the ILD registry (i.e. ~3000 confirmed/diagnosed ILD cases and ~500 confirmed non-ILD cases based on multidisciplinary review) and enriched with 10,000 randomly selected patients without ILD ICD 10 diagnostic codes from the screening EHR cohort (i.e. additional negative cases). The training set was then split into training (80%) and test sets (20%) using stratified K-fold cross-validation. The remaining patients from the screening EHR cohort were reserved for future use, including post-training model evaluation. A gradient boosting algorithm was used to train the initial ILD classifier (ILD yes/no) in a fully supervised approach, in which the model features were selected based on clinical domain expertise (i.e. radiology, pathology, laboratory and demographic data) and structured data elements available in EHR. After evaluating the initial ILD classifier performance (F-score, precision, recall and area under the receiver operating characteristic curve), we applied the ILD classifier to the screening EHR cohort, generated predicted probabilities for ILD (ranging from 0-1) and then randomly selected the first validation set (comprised of both yes and no cases, n=164) for blinded manual case review. These patients were manually labeled as ILD yes/no and this newly labeled data set was added to the training set to retrain the model. Explanatory model analysis was performed, including evaluation of feature importance to identify top influencer features and features associated with noise. This second iteration of refined models (with one or more features removed) and augmented training data were evaluated using a second validation set (Figure 1). Finally, we tested alternative algorithms to the Gradient Booster (initially chosen based on interpretability, which is crucial for user buy-in) to evaluate other machine learning techniques for classification.

Results: The initial model had a baseline recall (sensitivity) of 0.94 and precision (positive predictive value) of 0.86 (Model V1). This model was evaluated on an initial validation set comprised of 164 patients from the target EHR cohort that the model had not previously seen. Within the validation set, the model had a recall of 0.79 and precision of 0.69. The 164 patients from the target EHR cohort underwent blinded manual case review and classification, and then were used to retrain the model (Model V2), with improved performance (Table 1). Prior to the second iteration of the model, we identified three variables that were associated with noise (false positives) in the model – prednisone, isoniazid, and cough (Figure 2). Model versions V3a-c represented feature refinement with removal of (a) prednisone, (b) isoniazid, and (c) cough and isoniazid from the model. Model V3c was selected for its optimal performance metrics (Table 1) and tested on a new, second validation set of 422 patients. Within the validation set, model V3c had a recall of 0.85 and precision of 0.77. Finally, we compared the use of a Gradient Boosting machine learning algorithm to a Multiple Classifier System (MCS) as an alternative for increasing accuracy in pattern recognition. We used the
Table 1. ILD Classifier Performance Metrics

<table>
<thead>
<tr>
<th>Model Version</th>
<th>F-Score</th>
<th>Precision</th>
<th>Recall</th>
<th>Area Under the Curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Iteration</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model V1</td>
<td>0.899 ± 0.01</td>
<td>0.862 ± 0.00</td>
<td>0.940 ± 0.00</td>
<td>0.995 ± 0.0015</td>
</tr>
<tr>
<td>Model V2</td>
<td>0.919 ± 0.00</td>
<td>0.904 ± 0.00</td>
<td>0.937 ± 0.00</td>
<td>0.996 ± 0.0013</td>
</tr>
<tr>
<td>Second Iteration</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model V3a</td>
<td>0.922 ± 0.00</td>
<td>0.898 ± 0.00</td>
<td>0.946 ± 0.00</td>
<td>0.995 ± 0.0014</td>
</tr>
<tr>
<td>Model V3b</td>
<td>0.926 ± 0.00</td>
<td>0.921 ± 0.00</td>
<td>0.926 ± 0.00</td>
<td>0.996 ± 0.0013</td>
</tr>
<tr>
<td>Model V3c</td>
<td>0.926 ± 0.00</td>
<td>0.921 ± 0.00</td>
<td>0.932 ± 0.00</td>
<td>0.996 ± 0.0013</td>
</tr>
</tbody>
</table>

**Dynamic Selection (DS) approach of MCS and a pool of classifiers generated by Random Forest method, Bagging Classifier, Stacked Classifier for our study (Figure 3)**. Of these alternative models, the DS algorithm with Random Forest method yielded the best results, with a F-score = 0.895, precision = 0.884, and recall = 0.907, results similar to those observed using the original gradient booster algorithm prior to model enhancement (Model V1, Table 1).

**Discussion:** We have demonstrated that automated EHR algorithms can be developed using machine learning methods to identify a cohort of rare, complex ILD cases with high PPV using readily available, structured data fields. The models were optimized to identify ILD cases, with an ILD prevalence in the training and test sets that mirror real-world ILD prevalence. This is novel in the field of ILD which has thus failed to develop automated methods for cohort identification which could be used to improve clinical trial recruitment, identify patterns in health care utilization and associations with patient outcomes, and evaluate evidence-practice gaps. Serial training of the models by applying a combination of active learning, feature analysis, and feature modification led to iterative improvement in the performance metrics. Internal evaluation of the models using new validation sets demonstrated significantly improved performance with model refinement, outperforming all published ILD algorithms. Next steps will prioritize external validation of the model as well as development of future fusion models that combine wider range of structured data and incorporate unstructured data to improve patient representation and reduce errors.

**References:**
Measuring and Controlling Medical Record Abstraction Error Rates in an Observational Study

Maryam Y. Garza, PhD, MPH, MMCi1, Zhuopei Hu, MS1, Jeannette Lee, PhD1, Song T. Ounpraseauth, PhD1, Jessica Snowden, MD1, Anita Walden, MS2, Alan E. Simon, MD3, Barbara McClaskey, PhD, APRN, RNC4, Sandra Bauman, BS5, Sarah Sanders, RN, MS5, Lori A. Devlin, DO, MHA6, Leslie W. Young, MD7, Meredith N. Zozus, PhD, CCDM8

1University of Arkansas for Medical Sciences, Little Rock, AR; 2Oregon Health & Science University, Portland, OR; 3National Institutes of Health, Bethesda, MD; 4Pittsburg State University, Pittsburg, KS; 5University of New Mexico, Albuquerque, NM; 6University of Louisville School of Medicine, Louisville, KY; 7Larner College of Medicine, University of Vermont, Burlington, VT; 8University of Texas Health Science Center at San Antonio, San Antonio, TX

Abstract

Studies have shown that medical record abstraction (MRA) is among the most significant sources of error for data collection in clinical studies. Yet, the quality of data collected using MRA is seldom assessed. We employed a novel, theory-based framework for data quality assurance and control of MRA in clinical research.

Introduction

Medical record abstraction (MRA) has traditionally been, and continues to be, one of the most common forms of data acquisition for clinical research studies. However, the quality of MRA has often been questioned, as it is highly prone to human error and often adds to the overall complexity of clinical research. It is critical that the quality of the data collected through MRA be closely monitored and managed. However, attempts to improve the quality of MRA in clinical research have not been formally evaluated in the literature. Thus, improvements in MRA data quality have been limited. Our research aims to address this gap by implementing and evaluating a standardized process for MRA training and continuous quality control (QC) within the context of a clinical research study. Accordingly, the objective for this study was to determine the effects of formalized MRA training and continuous QC processes on data quality over time.

Methods

We conducted a retrospective analysis of QC data collected during a cross-sectional medical record review of mother-infant dyads with Neonatal Opioid Withdrawal Syndrome. Approximately 1,800 cases were abstracted across all sites, of which a subset of cases (over 200) underwent a formalized QC process to identify data quality errors and determine the association between MRA and data quality. The Error Rate Calculation framework, outlined in the Good Clinical Data Management Practices (GCDMP) guidelines, was utilized to describe error rates, the distribution of the error rates, and the error rates over time. A confidence interval approach was used to calculate crude (Wald’s method) and adjusted (generalized estimating equation) error rates across sites and over time. We calculated error rate using the number of errors divided by total fields (“all-field” error rate) and populated fields (“populated-field” error rate) as the denominators, to provide both an optimistic and a conservative measurement, respectively. The overall error rates were also compared to rates from the literature.

Results

The all-field error rate was 1.24%, 95% CI [1.14, 1.34], and the populated-field error rate was 3.04%, 95% CI [2.81, 3.30], across the full QC dataset (across all case types, across all sites). This translated to 124 and 304 true errors per 10,000 fields, respectively. Accounting for the clustering of cases within sites, the study total all-field adjusted error rate was 1.17%, 95% CI [0.91, 1.50] and the adjusted populated-field error rate was 2.87%, 95% CI [2.21, 3.74]. As expected, the 95% CIs for adjusted error rates were much wider compared to the crude estimates. For both the crude and adjusted populated-field error rates, there was a statistically significant downward trend among the sites with additional QC encounters. More specifically using the crude error estimates, the error rate decreased by 0.51% (p=0.017; 95% CI: [-0.88, -0.14%]; R²= 0.71) for each QC Events encounter. Similarly, the error rates accounting for clustering decreased by 0.46% (p=0.016; 95% CI: [-0.80, -0.13%]; R²=0.72) for each QC encounter. The overall
error rate from the literature meta-analysis was 6.57% (95% CI: 5.51%, 7.72%) compared to 1.04% (95% CI: 0.77%, 1.34%) based on the all-field ACT NOWS CE meta-analysis. The difference between the two error rate estimates were statistically significant based on the Wald-type z-test and meta-regression (p<0.0001). The error rate estimate for the populated-field ACT NOWS CE meta-analysis was 2.55% (95% CI: 1.88%, 3.35%). Similarly, the error rate was statistically significant compared to the 6.57% error rate based on the literature (p<0.0001).

Discussion

Through this work, we have quantified the results of the formalized MRA training and continuous QC process within the context of a multicenter clinical research study, and provided a baseline measure for traditional MRA error rates. More importantly, we have demonstrated that use of a standardized training program and ongoing data quality monitoring processes can yield positive results. The decisions to employ (1) formalized abstraction training prior to study start, and (2) continuous QC throughout the course of the study resulted in significantly lower error rates overall, with continued improvements in data quality observed over time. Importantly, the results presented here are of immediate use in informing investigators and research teams as they plan future research studies. The framework used here is a proven mechanism for controlling for data errors in studies relying on MRA for data collection that can be leveraged by other researchers to inform future study design and quality assurance processes for clinical studies relying on MRA for data collection.

Funding

This work was supported in part by the Environmental Influences on Child Health Outcomes (ECHO) program and the National Institutes of Health (NIH), and by the Translational Research Institute (TRI) at the University of Arkansas for Medical Sciences (UAMS), grant UL1TR003107, through the National Center for Advancing Translational Sciences (NCATS) of the NIH. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Developing A Precision Health, Patient Centered, Self-Management Framework

Stephanie Gilbertson-White, PhD, APRN-BC1, Heather Davis, MS, MLIS, PMP2,3, Chris Ortman, BA2,3, Jamie Sorensen, BA4, Kimberli Garcia, BS1, Chi Yeung, PhD1,5, Aaron Seaman, PhD3,4

1 College of Nursing, University of Iowa, Iowa City, Iowa; 2Institute for Clinical and Translational Science, University of Iowa, Iowa City, Iowa; 3Carver College of Medicine, University of Iowa, Iowa City, Iowa, 4College of Public Health, University of Iowa, Iowa City, Iowa; 5Memorial Sloan Kettering Cancer Center, New York, New York

Introduction
Adolescents and young adults (AYAs) with a history of cancer, defined as cancer patients ages 15-39, have many population-specific needs, including access to supportive care specialists, improved health literacy, self-advocacy for their medical needs, and improved methods of communication. It has been posited that AYAs have not seen improvements in cancer survival because the types of cancers most frequently diagnosed in these patients are both rare (and therefore understudied) and pathologically more aggressive.1 AYAs frequently travel long distances to academic medical centers to receive treatment from oncologists with expertise in their cancer. Access can be even more challenging for AYAs living in small towns or rural areas who often lack community-based supportive care specialists. Expectations for AYAs to self-manage stressors and symptoms from cancer have outpaced the development of effective interventions to ensure that AYAs have the knowledge and skills required. In addition, AYAs’ confidence in the efficacy of self-management strategies may interfere with their willingness and ability to follow through with strategies when they are provided. The need for supportive care interventions tailored to the unique needs of AYAs is critical. eHealth interventions that are mobile-ready, provide evidence-based self-management strategies, and target behavior change have the potential to address the delivery and uptake challenges and may ultimately decrease distress and improve quality of life (QOL) in AYA cancer survivors.

Learning to self-manage stressors and symptoms is a critical skill that AYAs must develop in order to minimize symptom distress, maintain optimal physical functioning, and maximize QOL. Engaging patients in self-management encourages them to prioritize problems and provides opportunities to obtain information, build skills, receive support and counseling, and implement strategies that are tailored to their individual needs.2 In response, we developed Oncology Associated Symptoms & Individualized Strategies (OASIS).3 OASIS is a theory-guided4 supportive care intervention designed to enhance self-management skills. The goal of OASIS is to help patients self-manage complex and interacting symptoms, access peer support, and communicate with their health care provider. OASIS 1.0 was a web-based intervention consisting of a library of educational content and self-tracking of self-management activities along with real-time telehealth visits over Zoom with a clinician. A review and meta-analyses of existing symptom self-management apps5 revealed no existing apps have integrated the elements of education, self-regulation, and social support. Based on these results with OASIS 2.0 we sought to add a social support community, allow video and/or 1:1 messaging with the clinician, and convert the intervention to a fully functional mobile app. The purpose of this study is to describe the results of our stakeholder engaged research conducted to develop and refine OASIS to be a patient-centered, self-management framework for a tailorable supportive care application.

Methods
A framework (Figure 1) was developed to map the application features and user cognitive and behavioral activities onto the theoretical underpinnings of the intervention. This framework provides a disease agnostic structure to the application allowing it to be tailored to the needs of users from anywhere along the cancer diagnosis/treatment/survivorship trajectory. For this study, AYAs with a current or past history of cancer were recruited to participate in two phases of data collection. Phase 1) We conducted online focus groups to gather perspectives and perceived concerns about the proposed elements of OASIS 2.0. Phase 2) Based on the results of the focus groups, a pilot study was conducted to gather user feedback about application features. The educational content library from OASIS, along with two

Figure 1. A patient-centered, self-management framework for a tailorable supportive care application

Application Features
- Tracking
- Metrics
- Data Visualization
- Diary
- 1:1 Clinician Messaging
- Peer Group Message Board
- Content Library
- Data driven recommendation

Theoretical Underpinnings
- Self Regulation
- Education
- Social Support

User Activities
- Reflection
- Action
- Self-efficacy
- Tailored Interactions
- Clinical Care
- Interaction
- Community
- Network Building
- Knowledge
existing mobile apps were provided to AYAs for a period of one month. Six cohorts of AYAs were recruited. No control group participants were included. Activities included tracking symptoms and strategies, recording thoughts in an open-ended diary, reviewing educational information in a content library, and participating in a peer group message board.

Results

Phase 1 (Focus Groups)
The sample consisted of (n=37) participants across 7 focus groups. The mean (SD) age was 26.78 years (+/-6.8). The focus groups were divided by age: 15-18 years (5 participants), 18-29 years (18 participants), and 30-39 years (14 participants). Most of the sample self-identified as women (65%) and post-treatment (67%). Major themes of the perceived concerns from the focus group were identified: general, look and feel, educational content, tracking their data, and peer group dynamics. General concerns about apps were push notifications, ads, apps that take up storage, difficult navigation, and unclear purpose. Look and feel qualities included: ease of navigation, visually interesting, professional look, and appropriate for various learning styles (text versus videos). Desired qualities for the educational content were: current content, readability, not too much, and not too little. Regarding tracking data comments included: wanting reminders, concerns about privacy, and connectivity to electronic health record. Finally, regarding peer group dynamics concerns included: who would moderate, age differences, disease status differences (active treatment versus post-treatment), bullying, inaccurate comments, and message overload.

Phase 2 (Pilot with existing apps)
AYAs ages 15 to 39 were recruited (n=61) in six waves: group 1, n=10, group 2 n=10, and group 3 n=16, group 4 n=7, group 5 n=14, and group 6 n=4. The mean (SD) age across the adult groups was 25.5 years (+/-6.5) range 18-40 and 16.2 years (+/-0.9) range 15-17 in the teen group. Most of the sample self-identified as women (70%). Tracking and diary use: Participants tracked their mood an average of 27.6 days (out 30). The mean mood rating was 3.82 (on a 1 to 5 scale with 1 = awful and 5 = rad). The most frequently tracked strategies were: Social Activity, Medications, Physical Activity, Exercise, Distraction, and Family Activities and. The diary was used by 70% of participants at least once. The average number of diary entries across the 30 days was 10.5 (range 0 to 30) with a word count that ranged from 2 to 300. Content library use: The pages that received the most visits were: Fatigue, Memory Problems, and Pain. The pages that had the longest visits were: Sleep Hygiene, Fatigue, and Memory Problems. Message board participation: Participation in the group messaging varied across groups. Excluding posts by the moderator, the average number of comments or replies by participants per group was 90.5 (range of 39 to 168). Groups with older participants (i.e. late 30s) tended to have more posts than groups with younger participants. Message board topics included: Introductions, podcast and book recommendations, cancer and social gatherings, cancer misconceptions, scan anxiety, symptom management (e.g. pain, sleep problems), and use of other health apps. Clear patterns emerged regarding social leadership in the message board. Groups with 2-3 strong message posters had greater overall message board participation. In addition, groups with participants who modeled highly personal sharing early on had more personal sharing overall.

Conclusions
Based on the findings of this study (focus groups and pilot with existing apps), we have developed an innovative intervention that will provide tailored symptom management care to patients with cancer across the disease trajectory. Our next steps are to evaluate engagement metrics from the phase 2 data and their associations with demographic and clinical characteristics as well as with the outcomes of symptom burden, self-efficacy, and perceived social support. In addition, we will recruit AYAs to pilot test OASIS 2.0 and to evaluate various app features designed to promote engagement with the intervention. This work has the potential to be quickly scaled up for efficacy testing and then rolled out into clinical practice.

References
The Challenge of Phenotyping Acute Adverse Events

Benjamin A. Goldstein PhD 1, Jillian Hurst PhD 1, Amanda Brucker PhD 1, Haley Hostetler MD 1, JJ Strouse MD 1, Cindy Ke Zhou PhD 2, Azadeh Shoaibi PhD 2, Hui-Lee Wong PhD 2

1Duke University, Durham, NC; 2Food and Drug Administration, White Oak, MD

Introduction

Health informatics research using both electronic health records (EHR) and insurance claims data requires the development of computable phenotypes to identify specific disease states. To date, most computable phenotype development has focused on the identification of chronic diseases. Notably, specificity is typically quite high 1 for these disease states, as patients will have multiple encounters and data points that can be used to define the disease state, and research to improve these phenotypes is typically focused on increasing sensitivity. In contrast, acute events, such as adverse events that can occur after a drug administration or procedure, typically have fewer data points and encounters that can be used for the development of computable phenotypes, and there is a need to develop highly specific computable phenotypes for their identification in various data sources.

Population-level evaluation of adverse events following biologic drug administration have been conducted using health insurance claims databases. Such databases provide robust coverage of medical encounters, regardless of where care is received. However, they have the limited information necessary to observe an exposure (i.e., administration of a biologic product) and outcome (i.e., diagnosis of a specific medical condition). A well-recognized limitation of such databases is that they lack granular information to accurately characterize a patient’s health condition and health history, including data related to chronic conditions, laboratory values, and timings of health care delivery.

In contrast to administrative claims data, health system-based electronic health records (EHRs) contain more granular information on patient encounters, allowing for detailed phenotyping of the patient’s condition and health history. However, unlike claims databases, EHR data lack information on encounters that occur outside of the health system, creating a challenge of observability.

In this study, conducted in partnership with the Food and Drug Administration (FDA) Biologics Effectiveness and Safety (BEST) Initiative, we compared different approaches for phenotyping adverse events (AEs), highlighting the opportunity to leverage granular EHR data.

Methods

Setting: We used data from the Duke University Health System (DUHS) EHR system. DUHS is an academic medical system with an integrated EPIC EHR system. Uniquely, DUHS serves as the primary provider (>85%) for residents in Durham County, allowing for (presumed) full capture of a range of medical services, including primary care, hospitalizations, and specialty referral care.

Exposure of Interest: For this study our exposure of interest was administration of intravenous immune globulin (IVIG). IVIG is a blood product prepared from batched donor serum that is used across medical specialties to treat a variety of conditions, including immune deficiencies, autoimmunity disorders, inflammatory disorders, and infectious diseases. IVIG is given both acutely and chronically, in both inpatient and outpatient settings, depending on the indication. We used RxCUIs to identify administrations. For each administration we abstracted time of administration and identified the administration setting, i.e., outpatient (OP) or inpatient (IP).

Adverse Event of Interest: Administration of IVIG has been reported to be potentially associated with a variety of AEs 2. We focus on three immediate (proximal) AEs (tachycardia, bradycardia and anaphylaxis) and two distal AEs (thrombotic events and hemolysis). For the proximal AEs they had to occur at the same encounter of the administration; for the distal AEs they had to occur within 7 days of the administration.

Phenotype Definitions: For each AE we considered three forms of a phenotype: ICD-9-/10-CM based, lab/vitals/medication based or the combination of the two.

Chart Review: We performed a RedCap based chart review for phenotyped AEs. For each suspected AE we assessed whether the AE occurred and whether it was deemed to be associated with the IVIG administration. This work was approved by the DUHS IRB and supported by an FDA BAA grant.

Results

We identified 3898 patients who had an IVIG administrations for a total of 29,973 administrations (median [Q1, Q3] of 4 [2, 8] per person). Over half (68%) of administrations were outpatient. The median (Q1, Q3) time between administrations was 2 (1, 28) days.
Phenotyping Results
1.4% of administration encounters had any suspected AE (30% OP, 70% IP). The ICD + contextual EHR data based definitions had the greatest PPVs. However, because cases that met this definition were the least prevalent, the definition would also have the lowest capture (sensitivity) in the full data, highlighting the trade-off between different definitions. Based on the chart review, many of the ICD-only false positives were for patients with a “history of thrombosis.” Therefore, we refined the ICD-based phenotype to exclude patients with any previous thrombotic events.

Table 1: Anaphylaxis Phenotyping Results

<table>
<thead>
<tr>
<th>No</th>
<th>Yes</th>
<th>PPV</th>
<th>#/ % patients in EHR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD</td>
<td>3</td>
<td>3</td>
<td>50%</td>
</tr>
<tr>
<td>ICD + Medicines*</td>
<td>0</td>
<td>2</td>
<td>100%</td>
</tr>
<tr>
<td>Meds Only</td>
<td>19</td>
<td>11</td>
<td>57%</td>
</tr>
</tbody>
</table>

*Based on Administration of Epinephrine with 6 hrs of IVIG administration

The administration of Epinephrine alone was not specific enough for AE detection. Even when epinephrine was administered it was often thought due to other reasons and not anaphylaxis. While the administration of medication alongside an ICD code was accurate, this was also quite rare.

Table 2: Tachycardia or Bradycardia Phenotyping Results

<table>
<thead>
<tr>
<th>No</th>
<th>Yes</th>
<th>PPV</th>
<th>#/ % patients in EHR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD</td>
<td>16</td>
<td>24</td>
<td>60%</td>
</tr>
<tr>
<td>ICD + Vitals*</td>
<td>4</td>
<td>16</td>
<td>80%</td>
</tr>
<tr>
<td>Vitals Only</td>
<td>33</td>
<td>7</td>
<td>18%</td>
</tr>
</tbody>
</table>

*Based on mean Heart Rate in 6 hrs after IVIG administration

Heart rate alone is not specific enough for AE detection. ICD codes alone were reasonably good indicators of the presence of tachycardia or bradycardia. However, since most of these were inpatient administrations, the AE was rarely thought to be caused by the IVIG (4% of the time).

Table 3: Thrombosis Phenotyping Results

<table>
<thead>
<tr>
<th>No</th>
<th>Yes</th>
<th>PPV</th>
<th>#/ % patients in EHR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD</td>
<td>27</td>
<td>23</td>
<td>46%</td>
</tr>
<tr>
<td>ICD + Labs*</td>
<td>2</td>
<td>16</td>
<td>89%</td>
</tr>
<tr>
<td>Labs Only</td>
<td>43</td>
<td>7</td>
<td>14%</td>
</tr>
</tbody>
</table>

*Based on presence of D-Dimer lab or Troponin > 0.01 ng/mL in 7 days following IVIG administration

ICD codes led to an over-diagnosis of thrombotic events, typically due to a “history of thrombosis”. The inclusion of a lab test increased the specificity of the phenotype at a cost of sensitivity. Based on chart review results, in future work we refined this phenotype to exclude people who had any history of thrombosis.

Table 4: Hemolysis Phenotyping Results

<table>
<thead>
<tr>
<th>No</th>
<th>Yes</th>
<th>PPV</th>
<th>#/ % patients in EHR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD</td>
<td>39</td>
<td>1</td>
<td>3%</td>
</tr>
<tr>
<td>ICD + Labs*</td>
<td>9</td>
<td>11</td>
<td>55%</td>
</tr>
<tr>
<td>Labs Only</td>
<td>30</td>
<td>10</td>
<td>25%</td>
</tr>
</tbody>
</table>

* Based on haptoglobin < 30 mg/dL or lactate dehydrogenase (LDH) > 200 U/L in 7 days following IVIG administration

ICD codes alone performed meaningfully worse than other phenotypes. This was mainly due to the presence of hemolytic disease as a general comorbidity. Similar to thrombosis, for future work we refined this phenotype to exclude individuals with a history of disease.

Discussion
This work highlights some of the opportunities and challenges in using EHR data to detect AEs. The more granular EHR data allow for more specific definitions of patient cohorts and outcomes. However, this specificity comes at a loss of sensitivity resulting in small sample sizes.

References
Evaluation of Google Search Trends for COVID-19 Vaccination Administration Efforts within a Public Health Department

Sae H. Han, MPH¹, Nicholas D. Soulakis, PhD¹²
¹Northwestern University, Chicago, IL; ²Chicago Department of Public Health, Chicago, IL

Introduction

During the COVID-19 pandemic, public health officials needed timely data to make actionable, evidence-based decisions. The newly curated Google COVID-19 Vaccination Search Insights (VSI) dataset may provide public health departments with up-to-date insights into community searches surrounding COVID-19 vaccinations. With these insights, public health departments may be able to develop tailored communication messaging and vaccination distribution strategies. The primary objective of this study was to evaluate the Google VSI dataset [1] by assessing whether Google vaccination search trends correlate with the number of COVID-19 vaccination doses administered in Chicago available in Chicago Data Portal [2], as well as assess the timeliness of any correlation. The secondary objective was to explore the possibility of utilizing Google vaccination search categories to predict vaccination doses administered.

Methods

The Google VSI dataset provides weekly time-series with relative frequency of searches for the following vaccine search categories: 1) COVID-19 Vaccination (all searches related to COVID-19 vaccination; overall search), 2) Vaccination Intent (searches related to eligibility, availability, and accessibility of vaccines; intent search), and 3) Safety and Side Effects (searches related to the safety and side effects of the vaccines; safety/side effect search), with the overall search being a superset of the intent and safety/side effect categories. Epidemilogic time-series and correlations analyses of the Google VSI dataset and the number of COVID-19 vaccinations administered to Chicago residents were conducted for the weeks between January 4, 2021 and December 20, 2021. Pearson correlation coefficients were calculated between the relative search frequencies and the population-normalized number of COVID-19 vaccination doses administered. Lag correlation analyses were performed to assess for temporal relationships between search frequencies and vaccinations administered. Additionally, time-series predictions for vaccination doses were conducted using Seasonal Autoregressive Integrated Moving Average (SARIMA) models with and without exogenous variables (i.e., Google vaccination search frequencies). The models’ performances were evaluated using root means squared error (RMSE). Analyses were stratified by the following two Chicago COVID-19 Community Vulnerability Index (CCVI) ZIP code categories: 1) High C CVI ZIP codes (communities that have more barriers to COVID-19 vaccine uptake) and 2) Low/Medium C CVI ZIP codes.

Results

Matching the two datasets on weekly date (N=51) and ZIP code (N=51) led to a total of 2,601 weekly instances of COVID-19 vaccination search frequencies and doses included in the study. This exploratory study found strong correlations between Google COVID-19 vaccination search frequencies and weekly COVID-19 vaccination doses administered in Chicago (r=0.58-0.83). When stratified by C CVI category, the correlations were stronger for the Low/Medium C CVI category (r=0.59-0.85) as compared to the High C CVI category (r=0.42-0.71) (Table 1). Strong correlations were found to persist between each of the three search frequency categories preceding the vaccinations administered up to three weeks (r>0.65) (Table 2). Results of the SARIMA and SARIMAX modeling indicate that including the three Google COVID-19 vaccination search categories, lagged by at least two weeks, create models that are better at predicting vaccination doses as compared to the SARIMA model that only includes the vaccination doses (Table 3).

Discussion

The study found strong correlations between Google COVID-19 vaccination search frequencies and weekly COVID-19 vaccine doses administered within Chicago. Additionally, the SARIMAX models that include the Google COVID-19 vaccination search categories were found to perform better at predicting vaccination doses in Chicago as compared to the univariate SARIMA model that only includes the dose data. The study demonstrates the value of Google’s new COVID-19 VSI dataset as a useful tool for public health departments to inform their
COVID-19 vaccination distribution plans. Furthermore, the lag patterns found in this study provide promising evidence for utilizing Google’s new dataset to allow for better COVID-19 vaccine resource allocation and ensure that public health departments are better prepared for increases in vaccination demand.

Table 1. Overall Correlation Coefficients for Google COVID-19 Vaccination Search Categories and Normalized Weekly COVID-19 Vaccination Doses Administered in Chicago.

<table>
<thead>
<tr>
<th>Search Category</th>
<th>Overall Correlation Coefficient [95% CI]</th>
<th>High CCVI Correlation Coefficient [95% CI]</th>
<th>Low/Medium CCVI Correlation Coefficient [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Search</td>
<td>0.68 [0.66, 0.70]</td>
<td>0.51 [0.46, 0.57]</td>
<td>0.71 [0.69, 0.73]</td>
</tr>
<tr>
<td>Intent Search</td>
<td>0.58 [0.55, 0.60]</td>
<td>0.42 [0.35, 0.48]</td>
<td>0.59 [0.56, 0.62]</td>
</tr>
<tr>
<td>Safety/Side Effect Search</td>
<td>0.83 [0.82, 0.84]</td>
<td>0.71 [0.67, 0.75]</td>
<td>0.85 [0.84, 0.87]</td>
</tr>
</tbody>
</table>

Table 2. Lag Correlation Coefficients for Google COVID-19 Vaccination Search Categories and Normalized Weekly COVID-19 Vaccination Doses Administered in Chicago.

<table>
<thead>
<tr>
<th>Time Lag (Week)*</th>
<th>0</th>
<th>-1</th>
<th>-2</th>
<th>-3</th>
<th>-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Search</td>
<td>0.68</td>
<td>0.72</td>
<td>0.748</td>
<td>0.747</td>
<td>0.72</td>
</tr>
<tr>
<td>Intent Search</td>
<td>0.58</td>
<td>0.67</td>
<td>0.74</td>
<td>0.779</td>
<td>0.778</td>
</tr>
<tr>
<td>Safety/Side Effect Search</td>
<td>0.83</td>
<td>0.80</td>
<td>0.74</td>
<td>0.67</td>
<td>0.57</td>
</tr>
</tbody>
</table>

* Time lag refers to the week(s) that the Google search categories precede the vaccination doses when calculating correlations.

Table 3. RMSE Values for SARIMA(X) Models that Predict Weekly COVID-19 Vaccination Doses Administered in Chicago.

<table>
<thead>
<tr>
<th>Time Lag (Week)*</th>
<th>0</th>
<th>-1</th>
<th>-2</th>
<th>-3</th>
<th>-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base + Overall Search SARIMAX (0,1,0)(2,1,2,7)</td>
<td>0.05071</td>
<td>0.01064</td>
<td>0.00924</td>
<td>0.01119</td>
<td>0.01042</td>
</tr>
<tr>
<td>Base + Intent Search SARIMAX (0,1,0)(2,1,2,7)</td>
<td>0.04332</td>
<td>0.01036</td>
<td>0.00961</td>
<td>0.01052</td>
<td>0.01101</td>
</tr>
<tr>
<td>Base + Safety/Side Effect Search SARIMAX (0,1,0)(2,1,2,7)</td>
<td>0.02691</td>
<td>0.01013</td>
<td>0.00760</td>
<td>0.00713</td>
<td>0.00769</td>
</tr>
<tr>
<td>Base + All Three Searches SARIMAX (0,1,0)(2,1,2,7)</td>
<td>0.03434</td>
<td>0.01021</td>
<td>0.01021</td>
<td>0.00912</td>
<td>0.00958</td>
</tr>
<tr>
<td>Base Model** SARIMA (0,1,0)(2,1,2,7)</td>
<td>0.00983</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Time lag refers to the week(s) that the Google search categories precede the vaccination doses when making predictions.
** Univariate base model only includes vaccination doses.

References

Development of a Smartphone-Based Adaptive Decision Support Tool for Predicting Patients at Risk of Chemotherapy-Induced Nausea and Vomiting using Decision Tree Induction

Humayera Islam\textsuperscript{1,2,3}, Md Kamruz Zaman Rana\textsuperscript{1,2,3}, A. Mosharraf Hossain, MD, MPH\textsuperscript{4}, Illhoi Yoo, PhD\textsuperscript{2,3}, Abu S M Mosa, Ph.D., MS, FAMIA\textsuperscript{1,2,3*},
\textsuperscript{1}NextGen Biomedical Informatics Center; \textsuperscript{2}Department of Health Management and Informatics; \textsuperscript{3}Institute for Data Science and Informatics; University of Missouri School of Medicine, Columbia, Missouri; \textsuperscript{4}Department of Hematology and Medical Oncology, BayCare Health System South Florida Baptist Hospital, Plant City, Florida

*Corresponding Author: mosaa@health.missouri.edu

Introduction

Chemotherapy-Induced Nausea and Vomiting (CINV) are the two most unpleasant concomitants of chemotherapy, which can lead to detrimental consequences on patients' health, increasing the socio-economic burden on overall healthcare. Although several antiemetic guidelines exist for CINV management, the guidelines are built on limited patient-related risk factors. Consequently, the physicians utilize their own experiences to manage CINV leading to inconsistencies in management and decision-making. Artificial Intelligence (AI) based smartphone applications developed from rigorously validated AI models and data can help predict the risk of CINV at the point of care. Moreover, access to robust clinical decision support can lead to better evidence-based treatment and improved CINV management. This study aims to develop a smartphone application for clinical decision support to recommend the risk of CINV at the point of care using a decision tree induction method trained on several patient-related CINV risk factors.

Methods

We collected data from the electronic medical records (EMR) used at the University of Missouri Ellis Fischel Cancer Center by retrospective record review. The detail of patient and variable selection process is described in Mosa(2020)\textsuperscript{1}. There were six independent datasets of patients based on emetogenicity (low, medium, and high), two phases of CINV (acute [AP] and delayed [DP]), and 14 patient-related risk factors. We used a decision tree model for predicting CINV for both phases of each emetogenicity level. The performance measures for the decision tree approach were compared with other popular machine learning algorithms. We developed the smartphone app called "CINV Risk Prediction Application" using the ResearchKit in iOS based on the predictions of the decision tree algorithm. We developed an adaptive approach (AD) instead of asking all the fourteen questions on patients one by one, which might have made the task more cumbersome for the clinicians. In AD, the question list adapts itself depending on the answers to the previous ones; hence, it utilizes the different paths depending on the response of the parent node of the decision trees. Figure 1 shows the AD approach, which first follows a single path from an AP decision tree flowchart to generate a questionnaire for the clinician and saves all the answers in a database. Later, when recommending CINV risk for the

Figure 1. Adaptive approach workflow for CINV prediction.
DP, the application chooses a DP decision tree flowchart and asks only previously unanswered questions. A step generator feature determines the next step, either a new unanswered question or a generated recommendation using the answered questions. This ensures only a minimum number of questions to be included for faster and effective decision-making.

Results

Decision tree performed well in both phases of high emetogenic chemotherapies with a significant margin than the other algorithms. The accuracy measure for the six patient groups ranges from 79.3%-94.8%. The flow of the adaptive approach for a single path is illustrated in Figure 2. This case shows the user-selected MEC as the emetogenicity level for AP, and the model chose the relevant decision tree. According to this flowchart, the first question was "whether the patient had anxiety during the chemotherapy", for which the user selected "no" as an answer. Following this answer, the next question was "the history of previous chemotherapy". Again, the user selected "yes", which led to the next question about "dehydration". Since the answer was "yes" for dehydration, the next question was "smoking status". By asking these four questions, the system identified that the patient is at high risk of CINV. Though there are 14 risk factors, our adaptive approach only asks the necessary questions by choosing one pathway from the flowchart. To provide a recommendation for DP, the app skipped the questions already being answered and asked only a single question to generate the final recommendation.

Conclusions

This study aims to address a real clinical problem that can help reduce the gap between clinical practices and evidence-based guidelines for CINV management. We used a rule-based app for its simplicity in explainability and software implementation. Furthermore, we developed the app using “an adaptive approach” that can be both time and energy-efficient at the point of care. This app also provides the flexibility to provide personalized care to meet individual patient needs; thus, promoting the notion of precision medicine in CINV management. Moreover, this app can be specifically beneficial for clinicians from hospitals without standard EMR access; hence, it can effectively reduce the existing inconsistencies of CINV management and decision-making across different geographic locations worldwide.

References

Large Scale Whole Genome Sequencing Analysis of Autism Spectrum Disorder

Jae-Yoon Jung, PhD 1, Brianna Chrisman 2, Kelley Paskov 3, Nate Stockham 4, Peter Washington 2, Sepideh Maleki 5, Kevin Tabatabaei 6, Dennis P. Wall, PhD 1,3

Departments of 1Pediatrics, 2Bioengineering, 3Biomedical Data Science, 4Neuroscience, Stanford University, Stanford, CA 94305, USA
5Department of Computer Science, University of Texas, Austin, TX 78712, USA
6Faculty of Health Sciences, McMaster University, Hamilton, ON L8S 4L8, Canada

Introduction

Autism Spectrum Disorder (ASD) is a neuro-developmental condition characterized by repetitive behaviour or interest, and early deficits in social communication and interaction1. While hundreds of ASD risk gene candidates have been reported from genome-wide association2 and next generation sequencing3 studies, to date recurrent CNVs are the only established factor for inherited risk4. Most other studies focus on de novo variant risk in simplex (only one child and no other siblings are affected in a family) families, which account for estimated 3 - 5% of ASD cases5 but in theory does not account for high heritability of ASD. Here we report the mega analysis result of the three largest ASD whole genome sequencing (WGS) data sets to identify and estimate the effect of the inherited variants across different variant types, combined with the effect of de novo variants.

Methods

The Hartwell Autism Research and Technology Initiative (iHART.org)6 data set includes 1,875 affected children and 418 unaffected siblings from 1,004 multiplex families, selected from the Autism Genetic Resource Exchange (AGRE)7 consortium. The MSSNG8 data set consists of 2,032 affected children from simplex families (with 169 unaffected siblings) and 838 affected children from 387 multiplex families, in total of 9,621 samples. The Simons Simplex Collection (SSC)9 data set consists of 2,408 affected children in simplex quad (two parents, one affected and one unaffected children) families, in total of 9,160 samples. The thousand genome project samples from the phase 3 WGS (N=3,202, 609 trios) are used to estimate the base ratio of rare and de novo mutations in controls. Each data set was sequenced in Illumina HiSeq or later platform, and jointly genotyped following GATK10 best practice pipeline.

After extracting rare (AF < 0.01) variants, we used VEP11 tool to group them into the following categories12: protein-truncating variants (PTV), probably-damaging misense (MIS3), H3K27ac peaks in fetal and adult samples (regulatory), SPIDEX peaks (splicing), and other regulatory regions defined by VEP. Next, for each trio family, we grouped the extracted rare variants into 3 classes: de novo, transmitted, and parent-only (not-transmitted). After removing any variants or samples out of more than 3 standard deviation from the mean in any category or classes, we performed burden tests and TADA-A analyses12 with 3 individual sets, simplex vs. multiplex data, and all merged data sets.

Results

First, we re-confirmed 22 genes with FDR < 0.01 from our previous work6. This is interesting in a sense that while the total number of samples is more than doubled-up in the current test, we did not use the previously-included recurrent CNV data sets with strong signals, implying de novo plus inherited variants can independently support the candidate risk gene findings reported from CNV studies. Second, simplex vs. multiplex data sets showed significantly different risk gene sets, while there are a few gene clusters shared across different family relationship models. Specifically, simplex (mostly SSC) data set has signals in de novo variants, while multiplex (mostly iHART and MSSNG) data set showed signals across de novo and inherited variants. In addition, the burden tests also showed that the relative risk of de novo variants are not the same across data sets. For example, while SSC data set has the similar risk (23) for the de novo PTVs, MSSNG and iHART sets show much less relative risk, between 12 to 17. Third, we found new risk candidate genes based on regulatory or splicing category signals (FDR < 0.1) that would not be counted as candidate genes without these non-coding variant signals. Re-confirmed variants in these categories include CHD8 target variants, that are among the strongest signals and consistent with previous findings3.
Conclusion

In this work, we demonstrated that for core ASD risk genes, inherited variants have a significant role in multiplex families to the extent of the role of de novo variants in simplex families, which explains how the same core candidate genes can appear in both family types. However, such distinction between different family structures becomes less clear for other ASD risk candidate genes. We will examine whether two ASD risk candidate gene sets from simplex vs. multiplex families will lead to significantly different biological mechanisms, as suggested in our previous report6. Finally, we found new candidate risk genes based on non-coding region variants, implying the complementary role of WGS in rare variant analysis.

References


mHealth-4-Mhealth (mobile health for migrant health) Surveillance Program

Ellen K Kerns, PhD MPH1, Chad J. Abresch, PhD1, Jana Broadhurst, MD PhD2, Michelle Warren, PhD3, Russell J. McCulloh, MD1

1Dept. of Pediatrics; 2Global Center, University of Nebraska Medical Center, Omaha, NE, United States
3Dept. of Modern Languages, University of Nebraska Kearney, Kearney, NE, United States

Introduction

Migrant children experience pronounced consequences when in-person school is closed, or when required to quarantine/isolate at home following a documented positive SARS-CoV-2 test in their household. Furthermore, they and their families are more vulnerable to socioeconomic costs when daily attendance at school or work is interrupted. This presentation will cover the implementation of our mHealth-4-Mhealth program that is designed to equip migrant school children’s households with (1) at-home COVID-19 symptom screening with optional salivary SARS-CoV-2 testing, (2) school attendance decision support, and (3) access to a community navigator for assistance with responding to social determinants of health challenges. This program is intended to reduce student absenteeism among migrant children in the face of the unique challenges presented by the COVID-19 pandemic.

Methods

Children and families will be contacted and enrolled through self-referral and recommended referral by the Nebraska Migrant Education Program (MEP). MEP staff will report aggregate monthly school absences among their enrollees who are and are not participating in the mHealth program. Participants will complete surveys covering the National Institutes of Health (NIH) Rapid Acceleration of Diagnostics-Underserved Populations (RADx-UP) tier-1 common data elements1. Participants households will complete two screening tools delivered via mobile application (app): 1) a COVID-19 symptom screening survey (completed twice weekly); and 2) a social determinants of health challenges screening survey (completed every other week). The adult head of each household enrolled in the study will download the app during enrollment. The other members in the household enrolled in the study can add it to their phones as well using the same email the head of household used to register.

COVID-19 household symptom screening: Households will complete the COVID-19 Symptom Screener during the school week (goal: twice weekly). A household member with the app installed will answer questions about household members’ exposure risk history and whether any household members are experiencing any symptoms of illness. Based on CDC and local health guidelines the app will then either inform the household member that: 1) SARS-CoV-2 testing is not recommended and household members can go to work or school; or 2) household members should perform at-home SARS-CoV-2 testing and stay home pending test results. Household members may choose to decline testing and select this option within the app. The app will provide CDC and local guidance on stay at home/isolation recommendations when testing results are not available. Household members will then be asked for reasons why they are choosing to decline testing and will be invited to participate in qualitative interviews exploring reasons for declining testing.

SARS-CoV-2 testing: Households choosing to obtain SARS-CoV-2 testing will collect saliva samples from each enrolled participant in the household and mail them via overnight shipper. Prior to shipping, an adult participant will complete a specimen manifest sheet to document the date and time that specimens were obtained from each participant. Positive test results will be reported to households by email and/or a phone call from the program’s test coordinator. Positive test results will also be reported automatically to the Nebraska Department of Health and Human Services. Households will be sent additional test kits and offered daily at-home testing for negative family members to identify additional possible infections. Households with a confirmed infection will also be asked to continue completing symptom screenings to determine if other household members develop symptomatic infection.

Use of the Household Challenges Tracker and Family Navigator: When respondents complete the Household Challenges Tracker within the mHealth tool, they will record challenges faced in the last two weeks based on questions derived from the NIH PhenX toolkit. Households that report at least one challenge will be asked to answer two additional questions: 1) was assistance available for all the challenges your household faced in the past two weeks; and 2) would you like additional assistance with any current challenges. For households indicating they would like to receive assistance, the app will direct their contact information, along with a report of challenges faced, to a family navigator who will contact the family to connect them with aid and resources. The navigators will
log the number of households requesting assistance, number of households contacted, and resolution date of their interactions.

Figure 1 details the data flow from participants to final reports/dashboards. mHealth tool (app) data is written to a FERPA/HIPAA compliant database. Every interaction made within the app is tagged and tracked, along with the registered household that used it, the date, time, and location of each session. Participant surveys are assigned by email and recorded via REDCap. Participant test results will be synced to the study SharePoint folder. The study SharePoint folder is accessible through an institutional login by an authorized employee and 2-factor authentication. All three data sources will be accessed through their respective gateways/APIs to create real-time reports and dashboards including: 1) individual-level test result lists organized by household and MEP program for use by the Test Result Coordinator, 2) household-level challenge lists among those requesting assistance for use by the Family Navigator, and 3) aggregate-level registration demographics and counts of symptoms, challenges, and positive tests over time by demographics for MEP staff to review.

Results

We deployed a pilot of the symptom screener (June 1st) and challenges tracker (August 1st) to 2 after school programs in summer 2021. The after-school program enrolled 119 students in 72 households (60% of enrolled) over 2.5 months. 7 households reported challenges with 3 requesting assistance in the first 2 weeks post deployment. The MEP mHealth program will begin recruitment in October with plans to enroll 800 participants (migrant children and their fellow household members) across 200 households. We will compare student absenteeism among MEP children whose families are participating in our program and those that are not. We also will compare the demographics of the children enrolled and the children whose households participate in each individual aspect of the program to those in the MEP to evaluate any disparities in representation. Finally, we will track program engagement over time to determine the proportion of households testing when prompted, recording weekly symptom screenings, recording monthly challenges tracking, and requesting assistance when challenges are reported.

Conclusion

Study participants already face the daily risks of job loss, school absence, and severe illness, disability, or death attributable to the SARS-CoV-2 pandemic. Should the mHealth-for-Mhealth program prove effective it will provide proof of effectiveness of an mHealth-directed, home-based disease surveillance and social risk mitigation program for migrant children and their families, which could directly benefit both study participants as well as migrant families across the United States.

References

An Open-Source Toolkit for Enriching Patient Records with Social and Environmental Determinants of Health

Paul Kingsbury, PhD, Beau MacDonald, MA, Praveen Angyan, MS, Mark Abajian, BA, Amy Chuang, MS, Daniella Garofalo, MPA, Neil Bahroos, MBA, Daniella Meeker, PhD, John Wilson, PhD, Juan Espinoza, MD

1Keck School of Medicine of USC, Los Angeles, CA; 2USC Spatial Sciences Institute, Los Angeles, CA; 3Children’s Hospital Los Angeles, Los Angeles, CA

Introduction.

The integration of social and environmental determinants of health (SEDoH) with clinical information has been identified as a priority to improve healthcare delivery and outcomes. Many important efforts have leveraged high-quality geocoded data aggregated from surveys and sensors to make that data usable for research and clinical practice. These efforts have tended to incorporate a limited range of data elements or aggregate several data elements into a single index. Previous efforts have also been difficult to replicate at new sites due to technical barriers, knowledge gaps, and governance concerns.

We announce here the release of the GIS Toolkit, a resource for incorporating neighborhood-level demographic, socioeconomic, and environmental data as well as model-based indices from geocoded sources. It incorporates the following novel characteristics:

1. Open-source code and documentation for ingesting datasets available nationally, such as the US Census American Community Survey (ACS);
2. A standards-based approach to leverage common data models;
3. Exposed error measurements and confidence intervals;
4. Accommodation to overlapping temporal changes in patient and environmental features.

Methods.

The GIS Toolkit enables institutions to enhance physical addresses with a set of curated SEDoH variables from several domains, including social and community contexts, economic contexts, education, physical infrastructure, and environmental conditions. Variable measurements are drawn from a variety of publicly sources, outlined in Table 1.

Table 1. Example SEDoH Variables

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Example Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Community Survey, 2013-2017 and 2014-2018</td>
<td>GINI Inequality Coefficient, Housing Median Year Built, Population Density, Old Age/Child Dependency Ratio, Federal Poverty Level, % Receiving SNAP, Median Household Income, Unemployment Rate</td>
</tr>
<tr>
<td>California Environmental Protection Agency Office of Environmental Health Hazard Assessment CalEnviroScreen 3.0 June 2018</td>
<td>O3, PM2.5, Drinking Water Quality, Asthma ER Visits, SB535 Disadvantaged Community</td>
</tr>
<tr>
<td>Centers for Disease Control</td>
<td>Social Vulnerability Index</td>
</tr>
<tr>
<td>Esri Business Analyst</td>
<td>Alcohol Access</td>
</tr>
<tr>
<td>Southern California Environmental Health Sciences Center</td>
<td>Average Ozone by year, Average NO2 by year, Average PM10 by year, Average PM2.5 by year</td>
</tr>
<tr>
<td>US Department of Agriculture</td>
<td>Low Food Access</td>
</tr>
</tbody>
</table>

Each variable is mapped to standard terminologies, when available, for ease of inclusion in applications such as i2b2; the Toolkit comes with instructions for integrating with both the OMOP and i2b2 Common Data Models. A tool is included for extending local i2b2 ontologies for the SEDoH variables. Some variables overlap with similar patient
variables, but care should be taken with interpretation. For example, a patient’s neighborhood may be identified as 53% Black population, but this is not the same as the patient themselves identifying as Black—although the neighborhood statistic may be useful for predicting the race for a patient who has not provided their demographic information.

Along with the SEDoH facts, we record metadata on the provenance and accuracy of the facts. These metadata include the upper and lower 95% Confidence Intervals and the start and end dates. The latter are especially important to capture patients who may not be present at an address at the same time as a data source measures a particular variable.

**Results.**

SEDoH facts are stored in the patient record separate from the patient address. The GIS Toolkit calls the US Census Geocoder API to associate the address to a geospatial polygon (eg, Census tract) and the data source facts about that polygon are applied to the patient. This allows environmental data about patients to be used and shared without exposing PHI in the form of patient addresses. Additional metadata records the validity dates for patients (when the patient used a particular address), measurements (eg air quality measured in a particular year), and even the geospatial polygon (since Census tracts are revised after each decennial Census). Figure 1 below shows a subset of variables and values for five synthetic addresses.

**Figure 1.** A subset of variables assigned to patient records

To evaluate the accuracy of the GIS Toolkit, a set of 1000 synthetic addresses (ie, well-formed and plausible but not associated with any human being) was drawn from openaddresses.io. The addresses are geolocated into Census tracts using both the GIS Toolkit and ArcGIS Pro and the resulting tracts compared. Agreement is reported with a simple concordance measure averaged over the 1000 addresses. Informal evaluation with a smaller number of synthetic addresses have suggested a nearly perfect agreement.

**Discussion.**

While most of SEDoH variables included with the GIS Toolkit come from national data sources, others are specific to southern California befitting USC’s patient population. Leveraging the Toolkit as a template, institutions in other regions can easily adapt to comparable local datasets.

**Conclusion.**

The GIS Toolkit offers a boost to institutions seeking to add contextual SEDoH variables to their patient analyses but lacking the resources to develop extensive capabilities on their own. The open-source license encourages institutions to incorporate the Toolkit into larger applications, as well as to contribute to the continued development of the Toolkit.

**Acknowledgement.**

This project was supported in part by NCATS UL1TR001855 and a subaward from NCATS 5U24TR002306.

**References**

5. Gardner BJ, Pedersen JG, Campbell ME, McClay JC. Incorporating a location-based socioeconomic index into a de-identified i2b2 clinical data warehouse. JAMIA. 2019;26(4):286-93.

Jeffrey G. Klann, PhD; Griffin M. Weber, MD, PhD; John H. Holmes, PhD; Gilbert S. Omenn, MD, PhD; Gabriel A. Brat, MD; Shawn N. Murphy, MD, PhD

†Harvard Medical School; †Research Information Science and Computing, Mass General Brigham; ‡Laboratory of Computer Science, Massachusetts General Hospital; ‡Beth Israel Deaconess Medical Center – All in Boston, MA; ‡University of Pennsylvania Perelman School of Medicine, Philadelphia; ‡University of Michigan, Ann Arbor, MI

Introduction

The coronavirus disease 2019 (COVID-19) pandemic has highlighted both the promise and current challenges in electronic health record (EHR) research. The Consortium for Clinical Characterization of COVID-19 by EHR (4CE) is an international consortium of over 350 hospitals across nine countries on four continents. Delineation of severe disease is central to understanding risk factors and treatments. Due to data inadequacies in existing severity definitions (e.g., data such as respiratory rate are unavailable in most EHR data warehouses), our network previously developed and validated a computable phenotype of COVID disease severity that focused on invasive mechanical ventilation for acute respiratory failure and vasoactive medication infusions for shock.[1]

The 4CE severity phenotype does very well at detecting severe disease, but it is not limited to only detecting COVID-19 disease. Although all patients in the 4CE cohort have a positive test for COVID, the algorithm does not take into account an increasingly common situation: a patient is hospitalized for e.g., a car accident, he/she incidentally has a positive COVID test, and he/she is discharged with no COVID symptoms. Because COVID tests are now routinely performed at hospital admission, some patients will be asymptomatic for COVID disease even after a positive test. Our hypothetical car accident patient would be labeled “severe COVID” by the 4CE severity measure, because of anesthesia medications used for surgeries. The severity phenotype finds any disease severity, including (for example) routine surgical interventions (anesthesia) and procedures performed on sick infants (intubation). Although in early 2020 most admitted patients were being treated for COVID disease, presently hospital systems are opening up, admitting many patients for routine surgeries.

Therefore, here we introduce a two-step approach that can adjust for this complexity. We add a first algorithmic layer to label patients who were admitted with a COVID-positive test but did not develop COVID disease. By filtering these out, our hypothesis is that severity detected by the 4CE phenotype will be COVID-disease severity.

The challenge is that no single data point in the EHR labels patients as “admitted for COVID disease.” Therefore we use a combination of chart review and data analytics to develop this label. To be resilient to site heterogeneity, we use an analytics approach called “hospital dynamics.” In this approach, analysis focuses on the metadata about a hospitalization rather than the treatments themselves. Previous work shows that meta-measures such as total number of lab tests on the day of admission or time of day of lab tests can be highly predictive of disease course.[2]

Methods

We selected a representative sample of our sites to develop this COVID-disease-hospitalization label that can be used at all sites in the 4CE network. These sites include Beth Israel Deaconess Hospital, Mass General Brigham, Northwestern University, and University of Pittsburgh. These sites were chosen because of their high engagement in the network, since this development and validation requires both a local clinical expert (for chart review and manual annotation) and a local data analytics expert (to try various analytic filtering approaches). Our approach is this:

a) Eligible patients for this study are those in the 4CE “COVID-19 cohort”: all hospitalized patients (pediatric and adult) with their first positive test for SARS-CoV-2 seven days before to 14 days after hospitalization.

b) At each development site, clinical experts manually reviewed the charts of admitted COVID+ patients at their institutions and recorded whether they were admitted for COVID-related-reasons and whether they developed (any) severe disease. We developed chart-review criteria requiring only admission and discharge notes, with additional context of laboratory values where possible.

c) Each site utilized this manually annotated gold-standard to develop labeling algorithms that predict which hospitalizations are COVID-related. The 4CE data are in a standard format, and we additionally standardize
the chart review annotation format. This allows us to develop shared analytic programs in R and Python that individual sites within the group can run.

d) The original 4CE severity phenotype algorithm can then be applied on the filtered cohort to more accurately predict severe COVID disease (as opposed to other severe disease unrelated to COVID).

To actualize the chart reviews, we worked with clinical experts and sample charts to define a chart review protocol. Determining whether COVID disease was truly at play can be difficult in certain “edge cases,” where only subtle uncommon symptoms (such as premature labor) are present. Our goal was to balance accuracy of such edge cases with a simplicity that allows high throughput and minimal additional training for clinical experts.

We focused our filter development on “hospital dynamics” measures, such as frequency, presence, and time of day of laboratory orders and categories.[2] Also, previous work has shown that abnormal values for certain labs are highly-associated with severe COVID, so we hypothesized that looking at the hospital dynamics of those lab groups might help us filter COVID admissions.[3]

**Results**

*Chart review.* As COVID is a complex disease that affects many organ systems, we chose a three-category annotation scheme. A pooled breakdown across our four sites is shown alongside the categories.

- **COVID disease [69%]:** e.g., respiratory insufficiency, hemodynamic changes, shortness of breath, anosmia
- **Possibly COVID-related [5%]:** e.g., early labor, liver dysfunction, immune dysfunction
- **Not COVID-related [26%]:** e.g., trauma, scheduled surgery, full term labor, drug toxicity

We separately identified indicators of severe disease: respiratory failure, hemodynamic instability, and ICU care.

We have completed 400 chart reviews at BI, 290 at MGB, 50 at NWU, and 70 at Pitt. We are presently working to complete 400 chart reviews at all four sites, covering the full 18 months since the beginning of the pandemic.

*Labeling COVID-related hospitalizations.* We applied established hospital dynamics measures (related to presence, frequency, and time-of-day/week) to 4CE data points to detect COVID-related hospitalization across our four development sites, with the goal of implementing a label and filtering network-wide as a first-pass to improve the fidelity of the severity phenotype.

We have developed visualization and analysis tools which can be used to select labels. Because hospital system dynamics regarding COVID change dramatically over time, many of our visualizations include the temporal aspect. One visualization shows COVID hospitalizations over time (according to the 4CE cohort) and then plots correct and incorrect classifications of COVID hospitalizations according to a variety of filters. Another uses association rule mining to find filter sets that optimize sensitivity or specificity of detecting true COVID-related admissions. As an example of our results, we have found that an order for LDH, CRP, or cardiac troponin predicts a hospitalization is COVID-related with 83% sensitivity and 80% specificity. Therefore applying a filter like this before data analysis will increase the percent of hospitalizations where COVID was being actively treated, thereby also improving the “severe” label’s ability to label only severe COVID disease. We are in the process of completing chart reviews, finalizing our tools, and choosing the best filters. We will present our final results in March at the Summit.

**Discussion**

As hospitals return to normal care patterns, the detection of “severe COVID disease” requires that EHR severity algorithms be adapted to select only hospitalizations related to COVID disease. We propose a two-pass approach, in which a cohort of presumed COVID hospitalizations is first filtered to eliminate COVID-unrelated hospitalizations with incidental COVID-positive tests. We combine a thoughtfully-designed chart review protocol with a powerful technique, “hospital dynamics,” or the use of metadata about the hospitalization (such as number of laboratory tests ordered) rather than specific treatments or laboratory values. This approach may be more resilient across heterogeneous sites, because it is not dependent on consistent normal ranges or specific treatment codes. We are applying this approach to a subgroup of sites in an international COVID-research consortium.

**References**

Evaluating Multifaceted Ontology Systems: 
A Preliminary Analysis of the GSSO in LGBTQIA+ and Medical Spaces

Clair Kronk, PhD¹, Cynthia Brandt, MD, MPH¹, Judith W. Dexheimer, PhD²
1. Yale School of Medicine, New Haven, CT;  
2. University of Cincinnati School of Medicine, Cincinnati, OH

Introduction
The Gender, Sex, and Sexual Orientation (GSSO) ontology, is a controlled vocabulary with over 10,000 
LGBTQIA+-related terms and an associated website, but while the ontology has been evaluated for use in natural 
language processing and semi-automatic literature review construction in relationship to MeSH and ICD-9/10, it has 
not been validated by users.[1,2] Validation is crucial to understanding usability of the platform, and its perception 
between groups. We surveyed users in the LGBTQIA+ and medical spaces to evaluate the GSSO using the 
Ontology Usability Scale (OUS) [3] and its website using the Software Usability Scale (SUS) [4] to determine if 
there were differences in scores between groups.

Methods
We created a 32-item survey using REDCap, which was approved by the CCHMC institutional review board.[5] 
The survey included three portions: (a) demographics (12 questions), (b) OUS (10 questions), and (c) the SUS (10 
questions). The OUS aimed to ask questions about the GSSO as an ontology, while the SUS asked questions about 
the GSSO website. For the OUS, a satisfactory score is between 38 and 44. We considered the OUS baseline as the 
scores provided for the VSTO [Virtual Solar Terrestrial Observatory ontology] (26), the SIO [SemanticScience 
Integrated Ontology] (36), the DCO [Data Collection Ontology] (29-42), and the GCIS [Global Change Information 
System ontology] (38-44) as our satisfactory range (26-44).[3] For the SUS, a satisfactory score is 68. We scored the 
OUS and SUS using an unweighted methodology (averaging the final scores) and a weighted methodology 
(averaging each individual question’s scores and then computing a final score). We then performed unpaired, two-
sample, two-sided, heteroscedastic t-tests (α = 0.05) between the medical/non- medical groups, and the 
LGBTQIA+/non-LGBTQIA+. We distributed a survey link to LGBTQIA+-related medical and non-medical interest 
groups from September 2020 to July 2021 and posted the link on the GSSO website (https://gsso.info/) in September 
2020. We included the GLBT Museum and Archives mailing list, the Homosaurus mailing list, the American 
Medical Informatics Association (AMIA) Mental Health Working Group (no LGBTQ+-related group existed at the 
time), the subreddits /r/AskTransgender and /r/AskLGBT, the SNOMED Mental and Behavioral Health Clinical 
Reference Group (as SNOMED has no LGBTQIA+-related groups), the Trans PhD Network on Facebook, the 2020 
LD4 Conference on Linked Data in Libraries Slack, the Queer PhD Network on Facebook, and the Trans Peer 
Network on Discord. Surveys were posted with descriptive text, abiding by any community rules established by 
individual groups. Posts were first made in September 2020 and a second time in March 2021.

Table 1. Demographic characteristics of respondents who completed the survey.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race/Ethnicity</td>
<td>35 white; 1 Indigenous America; 1 multiracial</td>
</tr>
<tr>
<td>Education</td>
<td>4 high school or equivalent; 7 some college or university; 9 associate’s or bachelor’s degree; 7 master’s degree; 11 doctoral or professional degree</td>
</tr>
<tr>
<td>Gender Identity</td>
<td>19 female; 9 male; 14 nonbinary; 1 genderless; 5 genderfluid; 2 questioning</td>
</tr>
<tr>
<td>Pronouns</td>
<td>12 he/him; 21 she/her; 15 they/them; 2 xe/xem; 2 e/em; 1 ze/hir</td>
</tr>
<tr>
<td>Sexual Orientation</td>
<td>6 heterosexual; 6 gay; 15 bisexual; 5 asexual individuals</td>
</tr>
<tr>
<td>Transgender Status</td>
<td>26 transgender individuals</td>
</tr>
<tr>
<td>Intersex Status</td>
<td>0 intersex individuals</td>
</tr>
<tr>
<td>Neurodiverse Status</td>
<td>20 neurodiverse individuals</td>
</tr>
<tr>
<td>Healthcare Professionals</td>
<td>11 healthcare professionals</td>
</tr>
</tbody>
</table>

586
Results
Of the 148 responses started, 44 (29.7%) were completed. 32 (73%) participants fully completed the OUS portion, and 29 (66%) completed the SUS portion. Of the 44 responses, 6 (14%) came from the GLBT Museum and Archives, 1 (2%) was from the Homosaurus arm, 1 (2%) was from the AMIA workgroup, 6 (14%) were from the Trans PhD Network, 29 (66%) were from the direct clicks from the website, and 1 (2%) was from the Trans Peer Network. The unweighted OUS score was 38.7 (s = 5.8), while the weighted score was 50.4. The unweighted SUS score was 49.8 (s = 3.8), while the weighted score was 64.2. The average unweighted OUS score in the medical group was 39.9, while the non-medical group score was 38.0. The scores in the SUS group were 50.6 and 49.5, respectively. For the LGBTQ group and non-LGBTQ group, the scores for the OUS were 38.9 and 37.9, and for the SUS were 50.0 and 49.2, respectively. We then performed t-tests between the medical and non-medical groups (OUS: p = 0.32; SUS: p = 0.43) and between the LGBTQIA+ and non-LGBTQIA+ groups (OUS: p = 0.70; SUS: p = 0.68).

Discussion
Scores on the OUS were on the higher end of satisfactory, with the weighted score being higher than the VSTO and the SIO, and in the range of both the DCO and the GCIS. The weighted score was higher than all scored ontologies, showing the GSSO was considered a more than satisfactory ontology in terms of the criteria scored. However, the SUS did not meet acceptable usability with the weighted and unweighted scores being lower than satisfactory. The website’s usability needs to be a major focus, as this may cause negative association with the vocabulary. To keep the ontology usable on multiple platforms, we located it on the NCBO BioPortal, GitHub, and Ontobee, as part of the OBO Foundry. Scores did not vary significantly between the LGBTQIA+ and non-LGBTQIA+ groups, or the medical and non-medical groups, showing that the ontology and website performed consistently between groups in terms of understandability and usability. However, there were not enough responses to further segment the population as LGBTQIA+ medical, LGBTQIA+ non-medical, non-LGBTQIA+ medical, and non-LGBTQIA+ non-medical, so it is unknown how these affinity groups interact on a meta-analytical level.

Response and completion rates of the survey were lower than expected and the length of the survey content) as an issue for such professionals. However, interestingly, the engagement of non-LGBTQ+ medical professionals was similarly low, pointing toward the length or content area (LGBTQIA+ content) as an issue for such professionals. We received a less diverse set of responses than expected, with most responses being from more educated white persons identifying as female or nonbinary. While there was a higher diversity than in comparable surveys in terms of gender, sexual orientation, and pronoun identification, the lack of diversity in terms of ethno-racial and educational backgrounds is crucial given the intersectional violence and discrimination faced by many members of the LGBTQIA+ community. Future work should focus on distributing information in a more equitable manner, and considering how internet-based resources change demographic outlook in terms of ethnic, racial, and educational backgrounds.

References
Identifying Interpretable Clinical Subtypes within Heterogeneous Dementia Clinic Population

Sayantan Kumar, MS;1 Zachary Abrams, PhD;1 Inez Oh, PhD;1 Aditi Gupta, PhD;1 Suzanne E. Schindler, MD, PhD;2 Nupur Ghoshal, MD, PhD;2,3 Albert M. Lai, PhD;1 Philip R.O. Payne, PhD1

1Institute for Informatics, Washington University School of Medicine, St. Louis, MO
2Department of Neurology, Washington University School of Medicine, St. Louis, MO
3Department of Psychiatry, Washington University School of Medicine, St. Louis, MO

Introduction

Dementia is a set of progressive neurodegenerative disorders associated with memory loss, cognitive impairment, and general disability.1 Dementia is highly heterogeneous, and the presence of different brain pathologies and variation in genetic background lead to significant variations in the clinical presentation and disease course. Hence, a heterogeneous group of cognitively impaired patients is composed of different subpopulations, each representing a specific disease course and characteristics.2 In this research, we use unsupervised clustering techniques to identify interpretable subgroups within the dementia population which are related to or predictive of disease progression. Analyzing the cognitive profile of the subgroups can lead to effective clinical decision-making and precision diagnostics tailored to each sub-group.

Methods

Clinical data corresponding to office visits were extracted from the Electronic Health Records (EHR) of patients treated between June 2012 and May 2018 at the Memory Diagnostic Center (MDC) at the Washington University School of Medicine in St. Louis, a large, academic, tertiary-care referral center. Longitudinal data from 1,845 patients with 2,747 visits were eligible for inclusion, where each visit recorded a Global Clinical Dementia Rating (CDR) score. Global CDR is a 5-point scale used to characterize 6 domains of cognitive and functional performance.3 Compared to expensive and/or invasive procedures like neuroimaging biomarkers, the CDR score is a standard metric in dementia research and is recorded for all cognitively impaired patients in the MDC.

The six components of Global CDR: Memory, Orientation, Judgment and Problem Solving, Community Affairs, Home and Hobbies, and Personal Care were used as input for the unsupervised K-Means clustering algorithm to generate the subtypes (clusters).4 Clustering analysis was performed on the 2,747 visits. This approach – using visits as opposed to patients - was taken to enable downstream longitudinal analysis and track the symptom progression rate among patients. The Gap-statistics algorithm was used to decide the optimal number of clusters.5

To gain an insight into which patients have a higher probability of progressing to a more severe stage in the CDR spectrum, patient transitions between different subtypes were analyzed across multiple visits. Patients with only a single visit were censored from our analysis. The differences in progression rate, both within and between Global CDR score categories were measured. The distribution of the six CDR components within each subtype provide validation and interpretability regarding the cognitive characteristics of the identified subtypes.

Results

Figure 1 shows the t-distributed stochastic neighbor embedding (T-SNE) representations of the CDR components across all visits distributed across the Global CDR score categories and the same representations distributed into the 16 subtypes (clusters). Figure 2 shows the number of visits and CDR composition of each subtype ordered by increasing CDR score (more severe dementia). Subtypes can either be homogenous (having a unique Global CDR) or include two Global CDR scores.

Figure 1: T-SNE plot (left) showing the 2D representations of the CDR components across all visits. The different colors show the Global CDR category of each data point. The clustering results (right) show the same representations distributed into subtypes (clusters) with the cluster centroids marked in cyan.

Figure 2: Stacked bar plot showing the CDR composition of each subtype. The x-axis represents the subtypes arranged in increasing order of Global CDR. The y-axis represents the number of visits in each subtype for each CDR category present.
The association between the six CDR components and the subtypes provides an intuitive interpretation of the six cognitive characteristics of each subtype (Figure 3). There is greater variability in early dementia (CDR of 0.5 or 1) leading to more subtypes with a lower CDR score. Patients with more severe dementia (CDR of 2 or 3) have less cognitive variability. Figure 4 shows the transitions from CDR 0.5 to CDR 1 between different subtypes. Subtypes 7, 11, and 15 are more likely to progress to subtypes with CDR=1 than subtypes 4, 14, and 16. We hypothesize these differences are related to the various subtypes having different underlying etiologies of dementia. The distribution of the CDR components of the two categories (subtypes 7, 11 and 15 versus subtypes 4, 14, and 16) suggest that the components of the CDR differentially predict dementia progression.

**Conclusion:** Cluster analysis using CDR component scores in a dementia clinic cohort, even within the same Global CDR, identified subtypes with different risk of dementia progression. Using CDR component scores enables straightforward interpretation of subgroup cognitive characteristics. Future steps include developing a machine learning model based on the subtypes to predict personalized rate of progression of dementia patients and further analysis on how the subtypes are associated with dementia biomarkers, neuroimaging features and cognitive disorders.

**References**


**Funding support:** Centene Corporation contract (P19-00559) for the Washington University-Centene ARCH Personalized Medicine initiative.
Extending Tree-Based Automated Machine Learning to Biomedical Image and Text Data Using Custom Feature Extractors

Rachit Kumar, B.S.¹, Joseph D. Romano, Ph.D.¹, Jason H. Moore, Ph.D.¹
¹University of Pennsylvania, Philadelphia, PA

Introduction

In recent years, there has been an explosion in the availability of biomedical data for predictive algorithms and clinical decision support systems. To learn from these data effectively, machine learning pipelines rely on feature engineering, model selection, and hyperparameter optimization, all of which are highly challenging and traditionally require significant amounts of manual tuning. Automated machine learning (AutoML) automates this process, providing a more structured approach to model design and hyperparameter optimization. However, current AutoML methods offer little to no support for non-tabular data types that are common in biomedicine, including image and free-text data. Current approaches for analyzing these data—including deep neural networks—suffer from interpretability issues and therefore face significant barriers to widespread implementation.

Here, we present a method to support the analysis of these complex data types by implementing feature extraction operators in the Tree-Based Pipeline Optimization Tool (TPOT), a tool that constructs machine learning pipelines using tree-based AutoML. This implementation allows TPOT to determine the best feature extraction strategy for the dataset and task at hand, enabling support for images and text data with the capacity to support other complex data types in the future as well.

Methods

TPOT uses strongly-typed genetic programming (GP) to construct and mutate ML pipelines, while employing Pareto optimization to maximize accuracy while minimizing pipeline complexity. Pipelines are instantiated by generating a tree consisting of individual operators, which include feature preprocessors, selectors, and classifiers, among others. To implement feature extraction in TPOT, new operators were developed that accept images or text as input and return tabular data in the form of feature vectors that can then be analyzed by subsequent operators supported by TPOT. A visual explanation of the new feature extraction operators is shown below (Figure 1).

![Figure 1](image)

**Figure 1.** Overview of TPOT’s new feature extractors for image and text data. Names of the operators are shown in red.

We evaluated the new feature extractors by running TPOT on a variety of image classification problems and text classification problems, comprising a mixture of biomedical and non-biomedical datasets. Each of these datasets has predefined training and sets, and sets that were used for training and accuracy determinations, respectively. Characteristics of these datasets are shown below (Table 1), including the number of classes, total number of instances, and the type of data type each dataset represents. To assess consistency across runs and general accuracy, we ran TPOT 5 times on each dataset, and optimized the pipelines based on Jaccard accuracy score and 5-fold cross validation. We computed intermediate accuracies of the candidate pipelines using training data, and reported overall accuracy based on the testing data.
Table 1. Details of the datasets used for evaluation of TPOT-FE and results of TPOT-FE optimization.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Type</th>
<th>Description</th>
<th>Training / Test Size</th>
<th>Classes</th>
<th>Mean Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>Image</td>
<td>8x8 images of digits</td>
<td>1347 / 450</td>
<td>10</td>
<td>0.981</td>
</tr>
<tr>
<td>FMNIST</td>
<td>Image</td>
<td>28x28 images of clothing</td>
<td>60000 / 10000</td>
<td>10</td>
<td>0.885</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>Image</td>
<td>32x32x3 images of various, distinct objects</td>
<td>50000 / 10000</td>
<td>10</td>
<td>0.519</td>
</tr>
<tr>
<td>STL-10</td>
<td>Image</td>
<td>96x96x3 images of various, distinct objects</td>
<td>5000 / 8000</td>
<td>10</td>
<td>0.477</td>
</tr>
<tr>
<td>pathMNIST</td>
<td>Image</td>
<td>28x28x3 images of pathology patches</td>
<td>89996 / 7180</td>
<td>9</td>
<td>0.686</td>
</tr>
<tr>
<td>dermaMNIST</td>
<td>Image</td>
<td>28x28x3 images of dermatoscope patches</td>
<td>7007 / 2005</td>
<td>7</td>
<td>0.714</td>
</tr>
<tr>
<td>Spam</td>
<td>Text</td>
<td>Raw SMS messages, ham and spam</td>
<td>4736 / 836</td>
<td>2</td>
<td>0.979</td>
</tr>
<tr>
<td>Real and Fake News</td>
<td>Text</td>
<td>News articles from various sources, real and fake</td>
<td>5384 / 951</td>
<td>2</td>
<td>0.888</td>
</tr>
<tr>
<td>Movie Reviews</td>
<td>Text</td>
<td>Movie review dataset for binary sentiment analysis</td>
<td>42500 / 7500</td>
<td>2</td>
<td>0.731</td>
</tr>
<tr>
<td>MT Samples</td>
<td>Text</td>
<td>Deidentified clinical records from various specialties</td>
<td>3949 / 698</td>
<td>22</td>
<td>0.364</td>
</tr>
</tbody>
</table>

Results

TPOT with feature extractors (TPOT-FE) performs well on many of the datasets, achieving performance considerably better than random guessing on every dataset. The average test accuracy for each dataset is given above (Table 1). Future work will include comparisons of TPOT-FE to other text and image classification algorithms.

On smaller image datasets (like MNIST and FMNIST) the vast majority of TPOT pipelines included the FlattenerImageExtractor operator, while more complex or larger data sets (like CIFAR10 or some medMNIST subsets) heavily used the DeepImageFeatureExtractor operator. For text datasets, use of the CountVectorizerTextExtractor or TfidfVectorizerTextExtractor operators was more variable, though TPOT-FE obtained consistent performance with each dataset. Generally, TPOT-FE is consistent within the same dataset as to which feature extractor it tends to select, and obtains similar accuracy across multiple runs, indicating that TPOT-FE is able to effectively explore the operator space of the feature extractors, and selects the best feature extractor for the current dataset in a reproducible fashion.

Discussion

TPOT-FE is a novel method for supporting non-tabular data in AutoML, allowing for automated exploration of the best possible feature extraction strategies for a given classification or regression problem. However, this approach does have some drawbacks. First, feature extraction can lead to significant memory usage as the number of features extracted from complex input data may require more memory than the original input data itself. Second, some feature extraction operators can take a long time to evaluate, though this is alleviated by caching the results of these operators for quicker retrieval. Finally, the implementation of feature extractors by strongly-typed genetic programming requires a custom implementation of operators to define their expected input and output types.

Future work on TPOT-FE will aim to address these issues and implement additional feature extractors; for example, implementing more complex word embedding models (such as word2vec) for text inputs, as well as supporting other complex inputs like time series data (leveraging existing APIs like sktime)\(^3\)\(^4\). Additionally, future updates will allow users to use multiple complex inputs in tandem (such as image and text data simultaneously) to better improve accuracy and to provide support for more complex biomedical workflows involving multimodal data.

References

Advancing Artificial Intelligence and Machine Learning Through Improved Public Competitions

Ezekiel J. Maier, PhD

1Booz Allen Hamilton, Bethesda, MD, United States of America

Abstract

Public artificial intelligence (AI) and machine learning (ML) crowdsourcing challenges can benefit the public and private sector by encouraging innovation and increasing advanced analytics knowledge and awareness. However, interest in crowdsourcing challenges has waned in recent years. Herein we analyze 1,337 completed crowdsourcing challenges to identify common issues and provide recommended approaches to increase the effectiveness of crowdsourcing for advancing AI and ML adoption, innovation, and training.

Introduction

The public and private sector, driven by the benefits of artificial intelligence and machine learning enabling machines to perform intelligent tasks, are taking swift action to (1) modernize infrastructure, including greater utilization of cloud computing, to support AI and ML development and operations; (2) adopt AI and ML solutions for improving business processes; and (3) upskill staff. While these actions are advancing the use of AI and ML, these steps alone are not sufficient. Experiential learning techniques are needed to truly engage and empower the public.

Experiential learning is employed in a variety of learning settings (e.g., medical school) to engage students, enable creativity, and achieve better real-world understanding. Public crowdsourcing challenges, which solicit voluntary ideas, content, or services, can provide similar benefits, bridging the gap between self-paced training and utilizing AI and ML. Organizations utilize crowdsourcing to increase engagement, spur innovation, and promote open scientific advancement. The U.S. Government engages citizens in crowdsourcing activities through its Challenge.gov platform. Private and non-profit sector crowdsourcing platforms include Kaggle, InnoCentive, TopCoder, and DREAM.

Despite engagement from public, private, and nonprofit sectors and the significant gains in educational attainment and internet access, the interest and perceived effectiveness of crowdsourcing challenges is decreasing. For example, Google Trends shows that worldwide web search interest in the term "machine learning" has been at or near all-time highs since early 2019. Yet, except for a brief COVID-19 induced increase in early 2020, interest in the term "crowdsourcing" has decreased by more than 50% since peaking in late 2013. Moreover, Citizen Data Science is rated as entering the Trough of Disillusionment in the 2021 Gartner Hype Cycle for Machine Learning and Data Science. This waning interest is driven by crowdsourcing failing to deliver clear innovative, and operationalizable results. Herein, we analyze 1,337 challenges, to identify common issues, and provide recommended approaches to increase the effectiveness of crowdsourcing as a mechanism for advancing AI and ML adoption, innovation, and training.

![Figure 1. (A) Google Trends web search interest in the terms “data science” and “machine learning”. (B) Google Trends web search interest in the terms “crowdsourcing” and “Kaggle”.](image)

Methods

Data on active and completed public challenges was collected from Kaggle and Challenge.gov. Kaggle challenge information, including titles, deadline dates, top performance incentives, and the number of participating teams, was downloaded using the Kaggle API (https://github.com/Kaggle/kaggle-api). Challenge.gov challenge information,
including titles, descriptions, incentives, and start and end dates, was scraped from the Challenge.gov website. Web scraping was performed using the R programming language (version 4.0.3), and rvest package (version 0.3.6). Analysis was completed in an RStudio coding notebook (version 1.2.1335) using R (version 4.0.3).

**Results**

Since 2010, the U.S. Government has invested significantly in crowdsourcing efforts. In addition to the innumerable person-hours spent organizing and running challenges, the government has allocated more than $204 million United States dollars (USD) in prizes for the 878 completed and active challenges hosted on Challenge.gov. The Department of Health and Human Services (HHS), National Aeronautics and Space Administration (NASA), and Department of Defense, have been crowdsourcing leaders, each running more than 60 challenges since 2010. Moreover, since 2018, more than 25 Challenge.gov challenges have focused on AI and ML model development. During the same timeframe, 488 challenges have been run on the Kaggle platform, offering monetary, prize, and swag incentives. Despite significant investment in incentives, challenge participation has stagnated. For example, among Kaggle hosted challenges, the number of participating teams increased significantly (Wilcoxon rank-sum test p < 0.05) from 2013 to 2015 but has not since 2015. Analysis of Kaggle challenges hints at approaches for increasing participation. For example, while only 3.3% of Kaggle hosted challenges used jobs as an incentive, there is an 80% increase in the median number of participating teams as compared to the 70% of Kaggle challenges that utilized monetary incentives.

**Discussion**

There are three major barriers that decrease the effectiveness of public and private sector AI and ML crowdsourcing challenges: (1) a lack of high-quality, high-impact, findable, accessible, interoperable, and reusable (FAIR) data due to issues securing sensitive information; (2) a narrow engagement focus on specialized groups, such as academics, rather than a wider audience with diverse thinking; and (3) insufficient operationalization of challenge results due to limited focus on the post-challenge phase where knowledge is extracted and operationalized. There are four approaches to increase the effectiveness of AI and ML crowdsourcing challenges: model-to-data, automated machine learning (autoML), machine learning operations (MLOps), and greater utilization of novel incentives.

- **Model-to-data** approaches can address data privacy concerns by evaluating models in a secure private computational environment that holds the underlying sensitive data. In this approach, participants develop and train their model on non-sensitive data, then containerize and submit their model for evaluation in the private computational environment[1]. The challenge community should adopt model-to-data approaches, enabling challenges that use high-quality, real, high-impact data and produce generalizable models

- **AutoML** tools automate many of the steps in the machine learning pipeline, including feature engineering, model selection, model training, and model validation[2]. The challenge community should leverage autoML to expand access to challenges and increase efficiency

- **MLOps** is a set of machine learning and DevOps practices for developing, deploying, and maintaining machine learning solutions, which includes model and data versioning, pipeline automation, testing, continuous integration and continuous delivery, and monitoring. The challenge community should adopt MLOps, including utilizing production deployment platforms such as TensorFlow Extended and Kubeflow, to ensure that community developed models can be operationalized.

- **Novel incentives**, such as jobs, fast-tracking pilots, partnerships, and contracts, will boost engagement while ensuring that clear steps are in place to reward top performance and participation. Challenge organizers should also highlight participant knowledge and contributions by providing letters of recommendation and documentation for augmenting data science portfolios, such as participant code developed, and badges won.

Study limitations include: (1) a dearth of data measuring challenge outcomes that can be used to quantify impacts of the three barriers and four approaches for revitalizing challenges, and (2) unmeasured education impacts of challenges.

**References**

Landscapes of Algorithmically Identified Patient Populations from the Electronic Health Record: A Systematic Review

Jennifer R. Malinowski, PhD1, Rebecca T. Levinson, PhD2, Luke V. Rasmussen3, Suzette J. Bielinski, PhD4, Laura K. Wiley, PhD5

1Write InScite LLC, South Salem, NY; 2Heidelberg University Hospital, Heidelberg, Germany; 3Northwestern University, Chicago, IL; 4Mayo Clinic, Rochester, MN; 5University of Colorado, Aurora, CO

Introduction
A common challenge in translational research is identifying patient cohorts from electronic health records (EHRs). Computational phenotyping aims to address this problem using algorithms (e.g., rule-based logic, natural language processing [NLP], and/or machine learning) to identify which patients have a particular clinical condition or characteristic. The EHR computational phenotyping community is diverse and builds algorithms for a variety of applications and clinical systems, which has led to a proliferation of algorithms and methodologies. To date there are no commonly accepted best practices for algorithm development, evaluation, and reporting. This makes it difficult to evaluate algorithm quality, compare existing approaches, and identify algorithms appropriate for reuse.

The first step to resolve this problem is to understand the current landscape of available phenotypes and current practices. Prior systematic, semi-systematic, and scoping reviews have focused on only a single phenotype (e.g., acute myocardial infarction), clinical domain (e.g., critical care), data source (e.g., primary care database), methodology (e.g., NLP), or publication venue (e.g., informatics journals). There is an urgent unmet need for a comprehensive review that considers all EHR-derived cohorts regardless of phenotype, methodology, or publication venue to inform new standards for reporting and evaluation of phenotyping algorithms. Here we present such a review that specifically examines all phenotype algorithms published to date to determine the identified patient population, development and validation methodologies, and performance characteristics.

Methods
We performed a search of MEDLINE using an iteratively developed search string that combines two sets of terms: 1) intended data source (EHRs), and 2) deliberately broad terms for automated methods of cohort identification. Search results were uploaded to Covidence, a software for systematic review management, and were assessed for inclusion in duplicate by blinded reviewers according to a pre-specified protocol. Conflicts were adjudicated through discussion or the input of a third reviewer, as necessary. The full search string and abstract inclusion/exclusion criteria are shown in Box 1.

Two reviewers extracted data from included articles, blinded to others’ responses. We first extracted information about the focus of the article (i.e., algorithm method development, applied study) and all eligible phenotype(s). Discordance between reviewers was resolved through discussion or by LKW or JRM as a third reviewer. Next, we extracted additional information into custom forms to identifying study details (e.g., study data source, type of applied study, etc.), algorithm methodology (e.g., type of data, type of computational method, etc.), algorithm validation (e.g., type of validation, process for performing manual review, etc.), and algorithm generalizability (e.g., demographics reporting, etc.). We extracted algorithm specific information individually for each qualifying algorithm. Neither quality assessment nor risk of bias were assessed for the included studies, as there is no validated instrument for these types of studies.

Results
The initial literature search for the review was performed on January 29, 2019 and will be updated in Fall of 2021. After removing duplicates, a total of 5,942 studies were screened for inclusion with 22.3% continuing to full text

Box 1. Systematic Review Search String & Inclusion/Exclusion Criteria

<table>
<thead>
<tr>
<th>Search String:</th>
</tr>
</thead>
<tbody>
<tr>
<td>“electronic health records” OR EHR OR “electronic medical records” OR “EMR” AND “natural language processing” OR classifier OR algorithm OR “machine learning” OR “deep learning” OR “artificial intelligence” OR phenotyp* OR “phenome” OR ICD OR probabilistic OR (“data-mining” OR “data mining”)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inclusion Criteria:</th>
<th>Uses EHR data to derive a patient population with a specific disease/condition, disease subtype, or disease symptom.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exclusion Criteria:</td>
<td>Performed outside the United States or only uses claims/administrative data (i.e., no EHR data).</td>
</tr>
</tbody>
</table>
review. Full text review excluded 74.1% of the studies with the majority excluded for not identifying a qualifying phenotype (n=194) or using manual review to identify the final patient population (n=167). Figure 1 shows the PRISMA flow diagram providing a full accounting of all studies reviewed, excluded, and included. Figure 1. PRISMA Diagram

The 342 included studies developed 661 algorithms for 312 distinct clinical conditions. The top 10 most frequently identified clinical conditions are shown in Figure 2. The majority of included studies (54%, n=185) had algorithm development as the primary focus, while 152 (44%) were application papers where the algorithm development was a single component of a larger analytic plan. The remaining 5 papers solely looked at the portability of previously developed algorithms.

Discussion

This is the first comprehensive systematic review of computational phenotyping algorithms that includes all clinical conditions, methodologies, and publication venues. Our results confirm the need for such an unbiased approach as over half of the studies identified were application papers that based on their title alone would not obviously describe phenotype algorithm development. Additionally, many papers describe the development of multiple phenotype algorithms with an average of 2 phenotypes per manuscript. Crucially, many algorithms have been developed for the same clinical traits. There were 43 algorithms published for Type 2 Diabetes alone. When considering all published algorithms for diabetes (of any type) there are more than 96 different algorithms described in the literature. Not only does this make algorithm reuse difficult (which algorithm should you choose?), it represents significant duplication of effort and cost to the research enterprise.

Works Cited

Understanding COVID-19 Information Needs from Conversational Logs

Mollie McKillop, PhD, MPH\textsuperscript{1}, Morgan Carroll, MS\textsuperscript{1}, Gretchen Purcell Jackson, MD, PhD\textsuperscript{12}, Brett South, PhD, MS\textsuperscript{1}
\textsuperscript{1}IBM Watson Health, Cambridge, MA, USA;
\textsuperscript{2}Vanderbilt University Medical Center, Nashville, TN, USA

Introduction

The dynamic nature of COVID-19 and need for socially-informed and often mandated preventative measures required new approaches for governments to share information and interact with their citizens. Limited institutional resources, stressed even further due to the pandemic, necessitated automated, scalable approaches. As a result, many agencies deployed conversational agents, which used natural language processing to answer user questions. This trend presented a unique opportunity to evaluate engagement with conversational agents and understand consumer health information needs. The primary objective of this study was to determine whether implementation of a conversational agent by a large city government could address most citizen information needs; a secondary objective was to characterize the consumer health information needs expressed in citizen conversations with this agent.

System Description and Implementation

The city government in Ontario, Canada serving over 350,000 residents implemented a web-based conversational agent in June 2020 to provide up-to-date information and resources to its residents and to deflect calls/inquiries from their main contact telephone number, as they were overwhelmed with questions pertaining to COVID-19. The conversational agent was implemented using Watson Assistant (WA), a platform for building a natural language conversational interfaces into any application, device, or channel such as a website or automated voice system. WA conversational agents can be trained to include customized information created for a geographic location, government agency, or organization. The four main WA components include: (1) understanding content, (2) classifying topics, (3) retrieving information from a knowledge base, and (4) generating natural language responses. WA interprets user entered questions to identify the intent (i.e., target of a user’s query) to answer questions. WA can be used to search, identify, and abstract information from unstructured information sources or trusted websites to leverage evidence-based sources such as guidance from local and national government agencies. WA treats user input (i.e., an utterance) as a search query and identifies information that is relevant external data sources, which can be customized for any subject matter. WA for COVID-19 conversational agents have previously been described in McKillop et al. 2021\textsuperscript{1}.

For this implementation, the application intents covered topics including COVID-19 symptoms and disease information, testing locations, preventative strategies, quarantine guidance, vaccine and treatment information, permitted activities, shutdowns, travel restrictions, reopening information, municipal services, volunteering opportunities, personal protective equipment (PPE) donations, guidance for small businesses and community organizations, and unemployment assistance/social aid. Any COVID-19 specific health information was provided from the Public Health Agency of Canada, Ontario Ministry of Health and the Regional Public Health Agency. Content for the conversational agent was provided in English.

Measuring Usage, Preliminary Performance, and Information Needs

We extracted conversational logs and usage data collected between 5/26/2020 through 1/19/2021. Usage metrics included the number of unique users, unique conversations (defined as when a user opens the tool and submits at least one input message), and total messages (each response the system sends to the user after an input message). Preliminary performance metrics were assessed through coverage, defined as the percent of utterances that are mapped to an intent. We also computed the percent of total messages that were escalated to a dialogue node which referred the user to a customer service line for more information, also called escalation percent. Consumer information needs were assessed through frequency of intents.

Confidence in mapping a user’s utterance to an intent is a performance metric, which determines the likelihood that an intent is an accurate match for an utterance. The method for calculating this confidence is proprietary. In a system implementation, a confidence threshold for providing a response to the user is set for every intent. The higher the confidence threshold, the ‘stricter’ the system is in providing a response to a user – meaning the system is less likely to provide a poorly matching response, but more likely ask for a clarification. For this application, the confidence threshold for all intents was set at above 50%. If multiple intents matched the user’s input above the 50% threshold,
the intent with the highest confidence was selected. During the study period, mean confidence levels were measured for mapped intents. All analysis was done in RStudio version 3.6.3. This study was determined exempt from human subjects review by the Western Institutional Review Board.

Results

At the time of implementation, the mean age of residents was 40 years. Nearly 90% of residents were Canadian citizens, and about 6% of residents were recent immigrants. Minorities made up the majority (55%) of the city’s population: 40% East Asian, 28% White, 20% South Asian, 4% Southeast Asian, 3% West Asian, 3% Black, and 2% mixed race. Average income of residents was approximately 90,000 CAD and 89% of people aged 25 to 64 had a high school diploma or equivalency certificate while 44% had a bachelor’s degree or higher.

A total of 10,597 users started conversations with the tool, and 13,006 unique conversations took place. A total of 616,691 messages were sent, and 85.2% of all utterances were mapped to an intent. Across all conversations, only 0.2% of messages were escalated to the dialogue node that referred users to a customer service line where they could talk to a government representative. The top five information needs, as measured by intent frequency and comprising 60% of all needs expressed, are listed in Table 1.

Table 1. Top 5 Information Needs of Users as Measured by Intents.

<table>
<thead>
<tr>
<th>Intent</th>
<th>Number of utterances mapped to intent</th>
<th>Examples</th>
<th>Average Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event shutdowns</td>
<td>38,210</td>
<td>“are the daycare centers reopening?”, “are churches open for funerals?”, “are at home gatherings with 10 people permitted?”, “can you plan a celebration?”</td>
<td>0.78</td>
</tr>
<tr>
<td>Facility closures</td>
<td>24,153</td>
<td>“are the parks opened?”, “what businesses are open”, “is the gym open”, “are indoor pools open?”, “is city hall open?”</td>
<td>0.56</td>
</tr>
<tr>
<td>Testing locations</td>
<td>23,586</td>
<td>“where to get tested?”, “where is the nearest testing center?”, how long after exposure can you get tested?”</td>
<td>0.89</td>
</tr>
<tr>
<td>COVID symptoms</td>
<td>16,065</td>
<td>“what are the symptoms of COVID?”, “what are symptoms?”, “is fever a symptom?”</td>
<td>0.94</td>
</tr>
<tr>
<td>Case counts</td>
<td>12,185</td>
<td>“How many hospitalizations are there in (our city)”</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Discussion and Conclusion

To our knowledge, this study is the first to measure success of a conversational agent in addressing COVID-19 questions in a government-based context and to characterize the resultant consumer health needs. Our findings demonstrate the ability of conversational agents to identify an answer for the majority (over 85%) of residents’ COVID-19 public health questions for a diverse and large population of users. Only a very small proportion of queries required escalation to a human customer service agent, illustrating success in automating this task. The most common information needs of citizen users covered a diverse set of topics including shutdowns, facility closures COVID-19 testing locations, symptoms and case counts. In ongoing work we are examining how information needs changed over the course of the pandemic and identifying topics expressed in utterances which could not be mapped to intents.

References

Addressing and Assessing COVID-19 Information Needs via a Weather App

Mollie McKillop, PhD, MPH1, Courtney VanHouten, MA1, Carol Bales, MS2, Winnie Felix, MS, MPH1, Rezzan Hekmat, MHS1, Gretchen Purcell Jackson, MD, PhD1,2, 3

1IBM Watson Health, Cambridge, MA; 2The Weather Company, Atlanta, Georgia; 3Vanderbilt University Medical Center, Nashville, TN

Introduction
The global COVID-19 pandemic resulted in a wide variety of new consumer health information needs, and people sought local data and information about COVID-19 from online sources. At the start of the pandemic trusted information was hard to find and constantly changing. The Weather Company (TWC) app for iOS and Android and weather.com website are consumer-facing tools that deliver free weather information to over 300 million users each month. As a result of the pressing need at the start of the pandemic to share basic public health information, TWC provided free COVID-19 information from trusted sources and developed a dashboard to display COVID-19 U.S. county and state-level per capita case counts using data from multiple national, state and local public health authorities1. In this study we hypothesized our commonly used app and website could be leveraged in a novel way to deliver COVID-19 public health information. We also sought to understand health-related information needs through a survey of users.

Methods
The TWC COVID-19 dashboard launched on March 25, 2020 as part of IBM’s corporate social responsibility initiatives in response to the pandemic and daily access was measured. Case count data was scraped daily from 50+ unique data sources in the US including free text PDFs, APIs and interactive dashboards and went through a verification pipeline. In August 2020, we conducted a cross-sectional survey among TWC app and website users U.S.-based and 18+ years of age. COVID-19 knowledge level, clarity of information, changes in preventative behaviors, and information preferences were assessed (see Table 1 for questions). Respondent sociodemographic characteristics and survey responses were summarized with descriptive statistics, and Chi-square tests of independence to compare distributions (differences considered significant at p<.05) were performed. All analyses used R version 1.4.1103. The study was reviewed and determined exempt by the Western Institutional Review Board.

Results
Since deployment, the dashboard (Figure 1) averaged 1.97 million daily users with an average interaction time of 1.63 minutes per user. Survey respondents (N=6,972) were White (83.2%), female (52.6%), with a mean age of 39 years (31.6%). Most had a bachelor’s degree or higher (51.5%) and lived in rural (30.5%), suburban (48.7%), and urban (15.9%) settings. 28.6% were essential workers.

Table 1. Survey questions and results.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Count (%)</th>
<th>Variable</th>
<th>Count (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge about COVID-19 prevention</td>
<td>Very good</td>
<td>3985 (57.2%)</td>
<td>Very good</td>
</tr>
<tr>
<td></td>
<td>Good</td>
<td>2446 (35.1%)</td>
<td>Local case counts</td>
</tr>
<tr>
<td></td>
<td>Fair</td>
<td>480 (6.9%)</td>
<td>TWC COVID-19 information ‘Very much’ or ‘Somewhat’ helpful</td>
</tr>
<tr>
<td></td>
<td>Poor</td>
<td>34 (0.5%)</td>
<td>Respondents interested in COVID-19 information topics</td>
</tr>
<tr>
<td></td>
<td>Very poor</td>
<td>27 (0.4%)</td>
<td>Case severity</td>
</tr>
<tr>
<td>Perceived clarity of how to prevent COVID-19</td>
<td>Very clear</td>
<td>3977 (56.5%)</td>
<td>Trends</td>
</tr>
<tr>
<td></td>
<td>Clear</td>
<td>2395 (34.4%)</td>
<td>End of pandemic</td>
</tr>
<tr>
<td></td>
<td>Neutral</td>
<td>320 (4.6%)</td>
<td>Prevention</td>
</tr>
<tr>
<td></td>
<td>Poor</td>
<td>124 (1.8%)</td>
<td>Testing</td>
</tr>
<tr>
<td></td>
<td>Very poor</td>
<td>196 (2.8%)</td>
<td>Increased preventative behaviors due to COVID-19</td>
</tr>
<tr>
<td>COVID-19 information sources used ‘Very often’ or ‘Often’</td>
<td>Websites or online news</td>
<td>4824 (69.2%)</td>
<td>Wearing a mask</td>
</tr>
<tr>
<td></td>
<td>Television</td>
<td>3824 (54.9%)</td>
<td>Social distancing</td>
</tr>
<tr>
<td></td>
<td>Gov health agency</td>
<td>3493 (50.1%)</td>
<td>Self-isolation</td>
</tr>
<tr>
<td></td>
<td>Conversations</td>
<td>3382 (48.5%)</td>
<td>Washing hands</td>
</tr>
<tr>
<td></td>
<td>TWC app or website</td>
<td>2916 (41.8%)</td>
<td>Disinfect surfaces</td>
</tr>
<tr>
<td></td>
<td>Newspapers</td>
<td>1976 (28.3%)</td>
<td>Not touching face</td>
</tr>
<tr>
<td></td>
<td>Social media</td>
<td>1709 (24.5%)</td>
<td>Supplements/herbs</td>
</tr>
<tr>
<td></td>
<td>Radio</td>
<td>1278 (18.3%)</td>
<td>Antibiotics</td>
</tr>
</tbody>
</table>
Those 18-29 years or 80+ years (p<2.2e-16), with a high school degree or less (p<2.2e-16), or male (p<2.2e-16) were less likely to have ‘Very good’ knowledge of preventative information. Less than a high school degree education (p=4.4e-06) or Black (p=0.009) were more likely to say COVID-19 information was ‘Very unclear’.

Local data and statistics about COVID-19 from TWC were more frequently reported as very helpful for those 18 to 29 years (p=0.0002), female (p<2.2e-16), Black or African American (p=3.6e-08), while males were less likely to say this information was very helpful (p<2.2e-16). Individuals with a college degree or higher (p=5.8e-13) or male (p<2.2e-16) were less likely to say news articles and videos about COVID-19 was very helpful to them.

Changes in preventative behaviors as a result of information about COVID-19 such as wearing a face mask were less often reported among respondents 30-49 years (p=1.6e-15), with some college or higher (p<2.2e-16), identifying as White (p=1.9e-10), male (p=7.7e-11), living in the suburbs (p=1.5e-10) or acting as essential workers (p=4.8e-05).

Discussion and Conclusion
Consumers use a wide variety of applications providing information about local conditions such as traffic, pollen levels, and weather to make informed decisions about daily activities. The COVID-19 pandemic affected nearly every aspect of daily life, and this study demonstrated avid engagement with a COVID-19 data and information hub delivered through a weather application. Almost 2 million users each day accessed the tool for a total of over 300 million visits, with each visit typically spanning almost two minutes. This illustrates how consumers sought local COVID-19 information from non-traditional sources during the pandemic to inform daily activities and behaviors. Similar case count dashboards to what was assessed in this study have been developed since the start of the pandemic. Yet, this study is the first to report on adoption and assess COVID-19 knowledge and information needs.

At approximately six months after the pandemic onset and deployment of the tool, most users said they had good knowledge about how to prevent the spread of the disease and that it was clear what safety precautions they should take to prevent COVID-19 spread. At the same time, over one-third reported not practicing preventative behaviors important for stopping the spread. Although the majority of respondents practiced preventative behaviors, some reported practices that may actually be harmful, such as increasing their use of antibiotics. Therefore, there exists a continued need to provide trusted information about COVID-19 that prompts appropriate actions.

Our findings may indicate areas of needed education, but with the plethora of COVID-19 information now available current needs center on providing the right information to the right individuals. Significant differences among sociodemographic groups with respect to practice of preventative behaviors and information preferences reveals that needs related to the COVID-19 pandemic are heterogenous and might require tailoring to specific user groups. At the same time, although consumer health information preferences are highly personal, they are not always predicted by sociodemographics. Further research should be done to explore this relationship.

Our survey received over 6,000 responses, but participants reflect only a small percentage of TWC COVID-19 hub users, and may not represent the entire US, with over 83% of our respondents identifying as White and over half (51.5%) having a Bachelor’s or Graduate degree. Finally, this study only explores needs around information topics - additional dimensions such as attitudes and motivations should be explored. Yet, this study is exploratory in that we aim to identify general correlation trends for future investigation. Overall, we find consumers want and need trusted public health data and information from a widely used weather application.

References
Neighborhood deprivation increases the risk of post-induction cesarean delivery
Jessica R. Meeker, PhD, MPH, Heather H. Burris, MD, MPH, Ray Bai, PhD,
Lisa D. Levine, MD, MSCE, Mary Regina Boland, PhD, FAMIA
1Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania
2Department of Obstetrics and Gynecology, University of Pennsylvania
3Department of Pediatrics, University of Pennsylvania
4Department of Statistics, University of South Carolina

Abstract. This study measures the association between neighborhood deprivation and cesarean delivery following labor induction among full term deliveries at Penn Medicine. After adjustment for confounders, we found that people living in the most deprived neighborhoods were at a 29% increased risk of post-induction cesarean delivery (aOR=1.29, 95% CI 1.05-1.57) compared to the least deprived.

Introduction. Among the over 3.7 million pregnant people who give birth in the United States annually, more than 20% of them will experience a labor induction, making induction one of the most common procedures done during pregnancy(1-3). Of these inductions, about one third will end in a cesarean delivery(4, 5). While the definition of a “failed induction” is not as simple as a cesarean delivery after labor induction, a vaginal delivery is often the preferred outcome by delivering women. There are many identifiable risk factors for cesarean delivery such as hypertension, obesity, parity, and gestational age, however one plausible risk factor with limited evaluation to date is neighborhood deprivation. Neighborhood deprivation has been associated with adverse pregnancy outcomes including pregnancy-induced hypertension and preterm birth(6). We sought to evaluate the link between neighborhood deprivation and post-induction cesarean delivery. People of color disproportionately undergo cesarean delivery in the United States. Even when controlling for sociodemographic factors and medical comorbidities, Black peoples have a 50% increased odds of cesarean delivery when compared to White patients(7, 8). Given the interaction of environmental stressors with hormonal pathways, it is biologically plausible that people from areas of neighborhood deprivation may respond more or less favorably to labor induction. Because differences in cesarean delivery outcomes cannot be attributed to sociodemographic factors and patient comorbidities alone, we must evaluate novel systemic risk factors(9) for increased cesarean risk, such as neighborhood deprivation. The aim of this study is to evaluate the contribution of neighborhood deprivation on risk of cesarean delivery after labor induction.

Methods. Our study population included people who had a pregnancy-related delivery diagnosis and procedure codes in their University of Pennsylvania Health System (UPHS) EPIC Electronic Health Record (EHR) system from 2010 to 2017 as well as an International Classification of Diseases versions 9 and 10 codes (ICD-9 and ICD-10) for labor induction validated by the American College of Obstetrics and Gynecologists. We then linked our data with detailed birth logs obtained from two hospitals within UPHS, the Hospital of the University of Pennsylvania (Philadelphia, PA) and Pennsylvania Hospital (Philadelphia, PA). We included all people who delivered at term (≥37 weeks) with a live, singleton gestation. We excluded people with a prior cesarean captured in the EHR and people lacking address information precluding geocoding. The primary outcome for this study was post-induction cesarean delivery for any indication. The primary exposure of interest was neighborhood deprivation. We chose to utilize the University of Wisconsin’s Neighborhood Atlas Area Deprivation Index (ADI), which ranges from 1-100, with a score of 100 being the highest level of deprivation in the US and a score of one being the lowest. We assigned an ADI score for each of the geocoded, block group geoids based on the latitudes and longitude of address at delivery. For each delivery, we binned the change in deprivation score into four levels: lowest deprivation (ADI score of 0-24), moderate deprivation (ADI score of 25-49), high deprivation (an ADI score of 50-74), and highest deprivation (an ADI score of 75-100) using evenly spaced deprivation score categories. We utilized a generalized linear mixed model for univariable and multivariable modeling.

Results. We derived a cohort of 63,334 pregnant people from the UPHS health system. We linked this with a birth log cohort obtained from the Hospital of the University of Pennsylvania and Pennsylvania Hospital from 2010-2017 resulting in a cohort of 35,787 people. After applying our inclusion and exclusion criteria, 24% of these people remained in our final cohort of 8,672 inductions. The post-induction delivery outcomes included 2,027 cesarean deliveries (23%) and 6,645 vaginal deliveries (77%). The average patient age at time of delivery was 28.4 ± 6.2 years. The predominant race self-designations were Black or African American, comprising 58% of people, and White, 30% of people. The majority of people reported their marital status as single (64%). We found that living in neighborhoods with moderate, high and highest levels of neighborhood deprivation resulted in elevated adjusted odds ratios for post-induction cesarean delivery compared to the lowest level of neighborhood deprivation. The odds of post-induction
cesarean delivery were elevated by 29% for the highest-level of deprivation (95% CI 1.05-1.57), 28% for the second highest-level (95% 1.04-1.57), and 20% for the third highest or moderate-level (1.00-1.44) (Table 1).

**Discussion.** Our study assesses the role of neighborhood deprivation on post-induction cesarean delivery as an adverse outcome of induction. We found that people from more deprived neighborhoods were at greater risk of post-induction cesarean delivery after adjusting for a multitude of confounders already known to increase risk. This study assesses the role of structural neighborhood deprivation on labor induction outcomes. In finding that neighborhood deprivation is associated with post-induction cesarean delivery, we can illustrate that neighborhood context may be important to the health of those delivering. Given that labor inductions are one of the most commonly performed procedures during pregnancy, and that cesarean deliveries are associated with increased morbidity, it is important that research continues to better identify individual and neighborhood-level risk factors of post-induction cesarean delivery. Importantly, the finding of a clear association with neighborhood deprivation and increased post-induction cesarean risk can inform public health practitioners and policy makers about the importance of evaluating risks among those from less-advantaged neighborhoods. Our work can help by shedding light on inequities that exist among those from less-advantaged neighborhoods. Future improvement of neighborhood conditions through the remediation of antiquated and inequitable policy is long overdue.

<table>
<thead>
<tr>
<th>Table 1. Associations between neighborhood deprivation and cesarean delivery following labor induction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Covariate</strong></td>
</tr>
<tr>
<td>Neighborhood Deprivation</td>
</tr>
<tr>
<td>Highest (75-100)</td>
</tr>
<tr>
<td>High (50-74)</td>
</tr>
<tr>
<td>Moderate (25-49)</td>
</tr>
<tr>
<td>Lowest (0-24)</td>
</tr>
<tr>
<td>Comorbidities</td>
</tr>
<tr>
<td>Diabetes (versus no diabetes)</td>
</tr>
<tr>
<td>Pregnancy-related hypertensions (versus not)</td>
</tr>
<tr>
<td>Obesity (versus not obese)</td>
</tr>
</tbody>
</table>

*Additionally adjusted for maternal age (continuous), race/ethnicity, parity, gestational age, and marital status*

**References**

Discovering drug combinations impacting cancer incidence

Rachel D. Melamed, PhD

1Department of Biological Sciences; University of Massachusetts, Lowell; Lowell, MA, USA

Abstract

In this work we seek to mine health claims data to find combinations of drugs that may alter onset of cancer. This work has an ultimate goal of preventing cancer related to medical treatment, and of suggesting new treatments for the disease. Because drug combinations impacting cancer are unlikely to be discovered using randomized trials, we develop new methods using observational data to discover these effects.

Introduction

Cancers occur more frequently in older populations due to accumulated cellular damage, but also due to accumulated environmental exposures. Because of this complex history contributing to their development, cancers are heterogeneous, and many are difficult to treat. This makes prevention of cancer and discovery of new treatments an important priority. Older people not only have more time to accumulate mutations in their cellular tissue, but they are also often under treatment for multiple other health conditions. Over one-quarter of the reports in the FDA’s Adverse Event Reporting System cite usage of more than one drug simultaneously at the time of adverse event. Due to the large number of possible drug combinations, the beneficial uses, and harmful side effects of all drug combinations have not been experimentally tested. Therefore, new methods that can discover effects of drug combinations on disease through analysis of observational data are needed. Some drug combinations are known to impact cancer development. Combination hormone replacement therapy increases risk of breast cancer, while combinations of statins with other common drugs (such as celecoxib and metformin) are in investigation for cancer treatment. In this study we use large observational health claims data to discover new drug effects on cancer onset. We develop a novel method based on the marginal structural model to scan for drug combinations impacting cancer risk.

Methods

Our method makes use of the IBM MarketScan claims data, containing histories of 150 million people, including coded prescriptions, diagnoses, and procedures. We take the approach of emulating a target randomized trial using observational data. The ideal randomized trial to discover effects of drug combinations on cancer might first enroll people taking a drug A, and randomly assign them to additionally take drug B (Fig 1A, Arm 2), or to continue to take drug A alone (Fig 1A, Arm 1). Then, our goal is to emulate this randomized trial for all common drug combinations present in the claims data. In the randomized trial setting, all patients who discontinue drug A would be considered to have dropped out of the study and censored from follow up. People may discontinue drug A because, for example, they develop some other health condition. Therefore, analysis of the randomized trial data requires adjusting for possible causes of censoring; similarly, we must adjust for medical history before censoring.

Additionally, we must adjust for medical history leading to prescription of drug B, since our data is observational rather than experimental, and people were not randomly assigned to take drug B. Confounding by indication can be one major source of bias. Figure 1B shows an example: menopause is a risk factor for breast cancer. If drug B is given as a menopause treatment, we would expect higher rates of breast cancer in people who initiate drug B, due only to increased rates of breast cancer after menopause.

To implement the drug-combination study, we build marginal structural models to adjust for time-varying confounding. We build our cohort as all new users of drug A with at least a year of history showing no use of drug A. Patient-time is censored when the person discontinues drug A. To account for time-varying confounding influencing the discontinuation of drug A, we build a logistic regression model to predict discontinuation, given patient history. This model yields propensity weights, which we use for inverse probability weighting. Second, we must account for time-varying confounding influencing initiation of drug B. Using the same method as for censoring weights, we obtain weights to abrogate this source of confounding. The two weights are multiplied to obtain the final time-varying weights. Finally,
using a weighted discrete-time logistic regression we implement the weighted Cox regression to model the effect of drug B initiation on time to onset of cancer.

In result, we can obtain an estimate of the effect of initiating drug B on each type of cancer. However, these estimates are highly dependent on the accuracy of our time-varying weights. Estimating these weights is challenging and prone to model misspecification, and the results are vulnerable to unmeasured confounding. For example, we expect that people who are in good health might be more likely to discontinue many drugs A; if our model is not able to perfectly capture good health, our estimates will remain biased. To address this issue, we develop a novel strategy. We collect a set of "negative control" outcomes, comprising diverse health conditions that occur at a similar incidence to common cancers in our population. Ideal negative controls are those that we do not particularly expect to be, as a whole, associated with the drug combination under investigation. In order to develop a broadly applicable method we collect a set of diverse health conditions with no obvious common underlying cause. We estimate the effect of the drug combination on all common cancers as well as these control outcomes. We expect any effect of general good health will be shared between all outcomes, while cancer-specific effects will not be shared across non-cancer outcomes. Further, many drugs that affect cancer affect multiple cancer sites. If we find a similar drug effect across multiple types of cancer, or similar effects across drugs with the same active ingredient, we hypothesize that that effect is not likely due to confounding but rather is a true drug effect. We implement a Bayesian meta-analysis to share information across many effect estimates, similar to Shahn, et al. This allows us to integrate multiple estimates of the effect of the drug combination on cancer to obtain a final estimate.

**Results**

Our results demonstrate both the difficulty in obtaining unbiased effect estimates, and the promise of our meta-analysis approach. The combination of metformin/sitagliptin with ciprofloxacin shows inflated hazard ratios, but the effect is similar for cancers versus non-cancer control outcomes (Fig. 2). This indicates no effect of the combination on cancer. We estimate the effect of each drug combination, as compared to taking each drug alone. While it would appear to be prohibitive to estimate all drug combinations, most drug combinations are not observed frequently, and only approximately 20,000 combinations are tested.

**Conclusions**

We have developed a method that enables a systematic estimation of the effect of all common drugs on all-cancer risk. We are applying this to all drug pairs and will extend to other non-drug exposures.

**References**

Fusion of Chest Radiographs and Electronic Medical Records using Deep Learning to Predict Intubation among Hospitalized Patients with COVID-19

Kim-Anh-Nhi Nguyen, MSc¹, Sahar Ghanavati, PhD¹, Prathamesh Parchure, MSc¹, Mukunth Rajendran, MSc¹, Arash Kia, MD, MSc¹
¹Clinical Data Science, Clinical Innovation Department, Mount Sinai Health System, New York, NY, United States

Abstract

Both chest radiographs (CXRs) and electronic medical records (EMR), typically obtained early in patients admitted with COVID-19, are key to decide whether patients need intubation. This study evaluates the use of a machine learning model to predict the need for intubation using a combination of CXR and EMR data. The model was trained and validated on historical data. Such a model can prioritize patients with high risk of intubation and result in improving care outcome.

Keywords: COVID-19, Machine learning and predictive modeling, Medical Imaging

Introduction

Delivering timely treatment for COVID-19 patients who need mechanical ventilation has a significant impact on clinical effectiveness and quality of care. We aimed to develop and validate a fusion machine learning (ML) model to predict the need for intubation among hospitalized COVID-19 patients, using a combination of inpatient medical records data and chest radiographs (CXRs).

Methods

Patient selection: we included adult patients (≥18 years of age) admitted in Mount Sinai Hospital between March 8, 2020, and January 29th, 2021, with a confirmed COVID-19 diagnosis by RT-PCR. Patients who were intubated within 24 hours of their admission were excluded.

Data sources: The study was approved by the Institutional Review Board of the Icahn School of Medicine at Mount Sinai, which waived the need for informed consent. EMR data (Epic, Cerner, SCC) at the Mount Sinai Health System included admission-discharge-transfer events, administrative data, time-series data of clinical and laboratory assessments, and electrocardiogram (ECG) results. Chest radiographs were in the format of DICOM images, modalities being CR and DX and orientations being AP and PA.

Modeling: Different fusion models were implemented[1]. First, we used sampled CXRs (last CXR 24 hours before intubation) to build an image classifier by using a pre-trained DenseNet[2] model (transfer learning[3]). Then, in the final selected model (as shown in Figure 1), we used a Random Forest (RF)[4] algorithm of which the input variables were 41 EMR variables and the probability score from the imaging classifier. Training was performed with ten-fold cross-validation with the AUROC as the evaluation metric. Then, a grid-search algorithm was used to tune hyperparameters based on the AUROC. Variable selection for EMR variables was done using recursive feature elimination[5] and pairwise multicollinearity removal[6].
**Evaluation methods:** In this retrospective cohort study, 1094 adults were hospitalized for COVID-19 at a large acute care healthcare system from March 8, 2020, to January 29th, 2021. We observed patients either until intubation or patient discharge. We trained both DenseNet and RF models on a training dataset (N=347) and validated them on a retrospective validation dataset (N=432) and a prospective validation dataset (N=315).

**Results**

At a prediction probability threshold of 0.5, the fusion model provided 84.8% (95% CI: 63.6%-100%) sensitivity, 82.7% (95% CI: 77.3%-87.8%) specificity, 82.9% (95% CI: 77.3 %-88.0 %) accuracy, and 0.87 (95% CI: 0.75-0.96) area under the receiver operating characteristics curve (AUROC) on the retrospective validation set. The model yielded an AUROC of 0.87(95% CI: 0.76-0.96) when applied to the prospective validation dataset. Compared to the image classifier alone that had an AUROC of 0.68(95% CI: 0.44-0.71) and 0.58(95% CI: 0.45-0.70) on the retrospective validation set and the prospective validation set respectfully, the fusion model showed a significant improvement.

**Discussion**

Using a combination of CXR and clinical variables for an ML-based model to rank the level of patient progress toward intubation may assist risk assessment and optimize clinical decision making in choosing the best care plan during the critical stages of COVID-19 illness.

**References**

Harmonization of Disease Pathways from Scientific Literature Using Graph Databases

Mihail Popescu, PhD, Mikhail Kovalenko, MS, Dong Xu, PhD, Richard Hammer, MD
University of Missouri, Columbia, MO

Abstract

We employed a graph database, Neo4j, to represent the disease pathways found in 45 articles and a non-small cell lung cancer (NSCLC) related KEGG pathway. We found that the graph representation was able to display pathways between two given genes even if they were not mentioned in any article in their entirety and it was able to find contradictory paths. Contradictory findings such as “A activates B” and “A inhibits B”, might reflect sample or methodology variations, which might require further research effort for clarification. We conclude that the graph representation of disease pathways might allow for promising developments in personalized medicine.

Introduction

The volume of the scientific literature grows exponentially every year, which makes it harder for researchers to stay current and to harmonize the content of multiple articles. Harmonization of knowledge across scientific literature is necessary due to differences in various factors such as population sample, methodology and terminology. We previously used a Resource Description Framework (RDF) framework for knowledge harmonization1. However, an RDF database framework requires a large amount of algorithmic development. This paper demonstrates a methodology for harmonizing knowledge related to disease biochemical pathways using a graph database, Neo4j. While Neo4j has been previously used in biomedical knowledge representation2,3,4,5, its potential for scientific knowledge harmonization has not been fully explored. As a test case, we analyzed 45 articles published in the last ten years related to the non-small cell lung cancer (NSCLC), a disease that about 200,000 Americans are diagnosed with each year. We have also inserted in our database genes from the NSCLC KEGG6 pathway, hsa05223. Aside from summarizing the existing literature of NSCLC, the developed methodologies could have an important impact in personalized medicine where a specific patient pathway might not fully match the related published pathways.

Methods

This work is part of an NIH-NLM project related to the bio-curation of the scientific literature of the disease pathways. In the first part of this project, pathway knowledge is extracted from both text and figures of the articles. After we tune our knowledge extraction methodology based on deep learning (see http://pathwaydeep.top/ for details), we intend to extract all NSLSC pathway information published in the last ten years (about 4000 review articles). However, for this pilot work, we used an existing curated 45 article dataset (C45A) created by our group for machine learning purposes (retrieved from PubMed using “PI3K AKT Pathway” terms) and the hsa05223 NSCLC KEGG pathway. Each pathway is represented by five nodes: Drug, Gene, Process, Substance, and Unknown, as well as five relations: Activates, Inhibits, Binds/Dissociates, Dissociates, and Alias. Each relation has two properties: source (the PMCID of the article where it was documented, or other IDs such as KEGG or HUGO) and the related pathway. The name of nodes and their relations are part of the KEGG2 ontology. The gene dictionary and their aliases were obtained from HUGO (https://www.genenames.org/). If a node is not found in the gene dictionary (and it is not a Process, Substance or Drug), it will be labeled “Unknown”. The query language for Neo4j is called Cypher. While the nodes are unique in our representation (there are created with the “Match” Cypher statement, which means that if the object exists, do not create a new one), the relations are not, since each relation depends on its source (hence they are created with the “Create” Cypher statement). One of the strengths of Neo4j is that it can create a new object (node or relationship) with the same name but with a different internal ID, such that the two can be distinguished (say the “Inhibits” relation from multiple papers), which enables answers to questions like: is there a path between gene A and gene B, what are all the paths between gene A and B, and how does a specific pathway (say from a patient) align with pathways from the scientific literature. So far, we mainly used Gene nodes and their relations in our work.

Results

Figure 1 shows all paths between EGFR and ERK, and the source of the relations (arc labels). The missing relation labels are for the HUGO “ALIAS” relation. EGFR is a known drug target in the NSCLC therapy and ERK is an activator of cell proliferation. We see that most paths except those from KEGG, are not contained in a single paper.
Moreover, only two articles (3260814 and 4593375) contain NSCLC as a MESH term. While 6 of the remaining 7 articles shown have some connection with cancer (having MESH terms such as “angiogenesis”), 3260814 is mostly related to autism which could suggest some link between the two7. So, this graphical representation allows for a global view of all the paths between two genes and it provides a more comprehensive visualization of the relations between two genes.

Figure 1. All paths found between 2 genes (ERK and EGFR), extracted from 9 articles and KEGG hsa05223.

Figure 2 shows all relations found in our C45A dataset between two genes, PTCH and SMO (part of the hedgehog pathway-HP). We see that two sources (2659383, 5149124) show activation while one shows inhibition (5149124). The reason for this is that we have two pictures in 5149124: one that shows that PRCH inhibits SMO, and the second shows that given ligands can induce its activation. Other reasons for discrepancies could be an algorithmic failure, dataset variation and differences in scientific opinions. Ways to deal with the contradictory data could be voting (trust the most frequent relation, ACTIVATES here) or giving a confidence to each relation (0.67 and 0.33) that could be used in further pathway matching algorithms.

Figure 2. Contradictory relations found between two genes, PCTH and SMO, in our curated dataset (47 articles)

Conclusions

Harmonizing existing biochemical pathway literature might give researchers and clinicians a way of understanding a large body of evidence. Moreover, the graph representation of disease pathways might allow for promising developments in personalized medicine where a pathway inferred from a patient’s data could be compared to the one existing in the literature.

Acknowledgements

This work was supported by the National Library of Medicine of the National Institute of Health (NIH) award number 5R01LM013392. The content is solely the responsibility of the authors and may not represent the official NIH views.

References

Consistency Check of the FHIR W5 Classification System in FHIR Modeling: Towards Improved Semantic Representation of FHIR RDF

Eric Prud'hommeaux1,2, Deepark K. Sharma3, David Booth4, Daniel J. Stone3, Nansu Zong3, Yue Yu3, Guohui Xiao5, Harold R. Solbrig6, Christopher G. Chute6, Kathryn J. Ruddy3, Emily Pfaff7, Guoqian Jiang3

Introduction

The Semantic Web's Resource Description Framework (RDF) is one of the standard exchange formats for FHIR data. FHIR RDF provides a standard, machine-processable semantic foundation for clinical data to be linked with other data using ontologies. It has been used by healthcare and research organizations to build semantic web and AI applications. For example, FHIR RDF has been used by the United Kingdom (UK) National Health Services (NHS) to enable Social Linked Data (SOLID) technology to provide patient-centric data storage and management. In another example, FHIR RDF has been used in a study in building predictive models to analyze primary cancers. Although a robust FHIR RDF specification based on the Shape Expressions (ShEx) has been implemented for standardizing and validating the FHIR RDF graphs and data, there is an increasing need to optimize this representation for better utility to attract users in the Semantic Web, SOLID, and FHIR communities.

Recently, the potential role of the FHIR W5 classification system (W5 for short) in improving FHIR RDF semantic representations has been of increasing interest to the FHIR RDF community. The initial FHIR model development process did not include a process for harmonizing property names across core resources. Consequently, FHIR RDF uses properties that include resource names (e.g., "Observation.subject") to ensure that no property gets re-used with inconsistent semantics. W5 was published in later release to establish consistent W5 metadata definitions across FHIR Resources. The W5 mappings captured in the logical model serve two purposes: 1) offering guidance to work groups designing resources and helping ensure consistency of content created by different work groups; 2) providing a standard view for implementers for processing and manipulating all resources that adhere to the same pattern. However, no study has been done on investigating whether the W5 patterns have been consistently applied in FHIR modeling. The objective of this study is to create methods and tools to check the consistency and utility of existing W5 mappings and enable more coverage of the mappings in FHIR resources.

Materials and Methods

Materials: 1) FHIR Resource Definitions: a master set of FHIR R5 resource definitions in JSON, including all the value sets, profiles, etc. defined as part of the FHIR specification, and the included implementation guides. W5 mappings are expressed in these resource definitions. 2) FHIR Model Ontology: an OWL representation of the StructureDefinitions of all FHIR resources. It formally enumerates the classes, predicates, domains, ranges and specific datatypes that are used in describing the FHIR instance data in RDF. Methods: We developed a JavaScript program to process the resource definition file (profiles-resources.json) in the FHIR JSON definitions. This definition file is a Bundle with FHIR operations and structure definitions in the entry array. Specifically, we first filtered for structure definitions, and removed meta structures such as CapabilityStatement, CompartmentDefinition, Bundle, StructureDefinition, Resource, DomainResource. Second, we extracted W5 mappings (indexed by the FHIR resource property names) from each bundle entry, i.e., Resource StructureDefinition. Third, we identified those FHIR properties with/without a W5 mapping by traversing all FHIR resources. There are three scenarios for the properties with/without W5 mappings: 1) All-mapped (meaning that any FHIR resource with this property has a W5 mapping); 2) None-mapped (meaning that this property without a W5 mapping for any FHIR resource); and 3) Some-mapped (meaning that this property in some FHIR resources has W5 mappings but the same property in other FHIR resources does not have W5 mappings). Fourth, we summarized the outcomes of identified mappings with property index and unmapped properties in a YAML file (https://github.com/ericprud/fhir-w5/blob/main/summary.yaml). We also developed a Java program that processed the summary YAML file and loaded the identified W5 mappings and unmapped properties into the FHIR Model Ontology. The purpose is to provide an ontological representation of the W5 patterns.
Results

We identified 4 properties belonging to the All-mapped category; 395 properties belonging to the None-mapped category; and 134 properties belonging to the Some-mapped category. Figure 1 shows distribution of mapped and unmapped FHIR properties with W5 patterns. Left panel shows 20 W5 properties with count for mapped properties (no-context). Middle Panel shows 17 no-context properties mapped to the W5 property FiveWs.subject[x] with count for mapped and unmapped properties (with context). Right panel shows a portion of the FHIR W5 properties represented in FHIR Model Ontology, indicating 17 mapped properties and 3 unmapped properties for the patient property that is mapped to FiveWs.subject[x]. These three unmapped properties are MolecularSequence.patient, RelatedPerson.patient, and SupplyDelivery.patient. Two co-authors (EP, GJ) reviewed the three unmapped properties and agreed that they should have the W5 mapping with FiveWs.subject[x]. The review result indicated that unmapped properties we identified could potentially reflect the opportunities to advance the W5 patterns in FHIR modeling.

![Figure 1. Distribution of mapped and unmapped properties with W5 patterns.](image)

Discussion

The idea of W5 mappings is crucial to FHIR scaling to a broadening set of resources composed from consistent terms. The style of W5 mappings has evolved over time. The mappings aimed at the five interrogatives leave no distinction between e.g. different kinds of "who" (participant, auditing agent, primary source). Later mappings expand the emergent vocabulary to include more nuance with W5 properties like "author" "source". These survey tools provide a way to methodically revisit W5 mappings to better exploit the emergent vocabulary. In this pilot study, we identified mapped and unmapped FHIR properties with W5 patterns, illustrating the need to have a community-based review for clarifying potential ambiguity of existing mapped properties (For instance, Observation.subject and Observation.focus are both mapped to FiveWs.subject[x]; these two properties could be stated to be subProperties in RDFS, but it is unclear what use cases would benefit from such subsumption) and establishing W5 mappings for unmapped properties. We have loaded all the mapped and unmapped properties with W5 patterns in FHIR Model Ontology which will serve as a semantic foundation for future curation and harmonization.

Conclusion

The W5 mappings do not, at present, have sufficient coverage to completely guide the development of a meaningful RDF vocabulary. However, analysis with these tools offer an opportunity to harmonize W5 mappings and create a FHIR RDF vocabulary both precise enough to convey clinical data and abstract enough to allow users to exploit commonalities between different resources.

Acknowledgements: This study was supported in part by funding from the NIH FHIRCat (R01 EB030529) and the NCATS CD2H (U24 TR02306).
Defining Venous Thromboembolism Using the Electronic Health Record: A Data Driven Approach
Avery Pullman, BS1, Erin Coghlan1, Ania Syrowatka, PhD1,2, Michael Sainlaire, MS1, Wenyu Song, PhD1,2, Patricia C. Dykes, PhD, RN1,2
1Department of Medicine, Brigham & Women’s Hospital, Boston, MA; 2Harvard Medical School, Boston, MA

Introduction:

Venous thromboembolism (VTE), consisting of pulmonary embolism (PE) and deep vein thrombosis (DVT), is a common, preventable public health problem affecting approximately 300,000 – 600,000 individuals in the United States each year, and requires timely and adequate treatment. However, symptoms for VTE tend to be non-specific, making timely diagnosis a challenge. To improve VTE detection and diagnosis, informaticians and researchers leverage healthcare databases to identify instances of VTE for research and quality reporting purposes. Traditionally, identifying cases of VTE has relied mainly on the use of ICD codes. However, previous studies have demonstrated that the use of ICD codes alone are often subject to error due to limitations in available clinical data, diagnostic errors, and coding errors made by human operations. These studies indicate that the use of ICD codes alone for defining a VTE provides a poor Positive Predictive Value (PPV) of approximately 64%. Other studies have proposed methods to improve this PPV by defining a VTE as the combination of an ICD code and an imaging code for VTE within a given encounter. This method results in a notable improvement, raising the PPV from 64% to 75%. The team at Brigham and Women’s Hospital (BWH) attempted to increase this PPV further through adding an additional Electronic Health Record (EHR) derived feature, medication data, via the presence of an anticoagulant order 6 hours prior to or following the imaging scan (table 1). This podium abstract, therefore, reviews a novel algorithm for defining VTE using a data driven perspective through combining ICD, imaging, and medication codes. These three pieces of information were selected because, in combination, they indicate that the patient was billed for a VTE, had an imaging scan done to identify a VTE, and was treated for a VTE. Therefore, our team hypothesized that if all these data elements were present within a given encounter, the patient likely had a true VTE.

Methods:

Study population:
Records for our target population of patients, that is patients, aged 18 and older who had a primary care visit from December 2016 – January 2020, defined as an office visit with an internal medicine, general medicine, or family medicine provider, were extracted from the EHR within the Mass General Brigham system. From this cohort, we selected patients who had an ICD code for VTE (Table 1) within the 30 days following that index visit. We then examined the patients who had an imaging code linked to the same encounter as the ICD code and had an anticoagulant order 6 hour prior to or following their imaging scan (Table 1). The ICD, Imaging, and RxNorm codes that were selected resulted from an extensive literature with the Harvard Countway Librarian and were validated with a Technical Expert Panel.

Chart review and Statistical Analysis:
To test the accuracy of this novel VTE identification algorithm, we randomly selected 550 patients from this cohort and a trained chart abstracter examined each patient’s imaging results from the identified encounter to determine the presence or absence of a VTE as noted by the “imaging indication”. We then compared the results of this gold standard review to that of that determined by the algorithm. Using this method, if the abstracter found that the patient had a VTE during their hospital encounter, aligning with the same hospital encounter indicated by the algorithm, it was considered a true positive, if not, it was
marked as a false positive. The number of true positives and false positives were recorded in order to calculate the novel algorithm’s PPV which was then compared to the alternative methods proposed by previous studies.

**Results:**

From our randomly selected 550 patients that our VTE identification algorithm determined had a true VTE, it was found that upon gold standard review, 524 actually had a true VTE. This provided a PPV of 94%. Of those incorrect cases, which there were 26, the majority were instances where the provider suspected a PE, conducted imaging to confirm the presence of a PE, but instead found a pleural effusion, which was treated with anticoagulants. Therefore, the event was billed as a VTE, imaged as if it were a VTE, and was treated like it were a VTE. Therefore, our algorithm incorrectly noted them as a VTE.

**Discussion:**

This podium abstract describes a novel data-driven method for accurately defining a VTE based on the use of billing, imaging, and medication information. This method for defining a VTE provides a notable increase in PPV, up to 94%, when compared to previous methods for defining a VTE, such as though using the combination of ICD and imaging codes, or ICD codes alone (75% and 64% respectively) allowing researchers and informaticians to accurately study and report on instances of VTE. The team at BWH is currently working on conducting further reviews to quantify this method’s sensitivity, specificity, and Negative Predictive Value. The team at BWH is exploring opportunities for further improving this diagnosis pipeline by utilizing natural language processing on the imaging scan’s “imaging indication”. Additionally, the BWH team looks forward to testing this algorithm at an additional site on a separate EHR system in the coming months.

**Figures and Tables:**

<table>
<thead>
<tr>
<th>VTE Codes</th>
<th>I26.9, I80.2, I80.3, I82.4, I82.5, I82.6, I82.7, I80.1, I82.8, I80.9, I82.9, I26.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imaging Procedure Codes</td>
<td>173016, 146694, 146692, 142167, 147498, 142161, 142163, 144330, 142182, 173016, 146694, 146692, 142167, 147498, 142161, 142163, 144330, 142184, 352026, 142294, 147036, 173052, 173050</td>
</tr>
<tr>
<td>Anticoagulant Rx</td>
<td>Warfarin, Heparin, Enoxaparin, Dalteparin, Fondaparinux, Rivaroxaban, Apixaban,</td>
</tr>
</tbody>
</table>

**Citations:**


Using PEDSnet, a National Clinical Research Network, To Track the Evolution of COVID-19 in Children Within the United States

Hanieh Razzaghi, MPH1, H. Timothy Bunnell, PhD2, Sarah Deakyne Davies, MPH3, Richard Hoyt, BS4, Brianna Magnusen, MD5, Nathan Pajor, MD6, Daksha Ranade, MPH, MBA7, and L. Charles Bailey, MD, PhD1 for the PEDSnet COVID-19 Team

1Children’s Hospital of Philadelphia, 2Nemours Children’s Health System, 3Children’s Hospital of Colorado, 4Nationwide Children’s Hospital, 5St. Louis Children’s Hospital, 6Cincinnati Children’s Hospital, 7Seattle Children’s Hospital

Introduction
The SARS-CoV-2 pandemic has been a significant worldwide public health problem since November 2019, with over 58 million cases and 833,000 deaths in the US since March 20201. While overall disease burden in children has been milder than adults, risk is higher for children with chronic conditions 2, and child-specific manifestations such as MIS-C3 have been identified as well. Importantly, as the pandemic has evolved with changes in predominant viral variant, there has been widespread speculation about changes in the risks to children, fueled by small case reports or summary reporting. Clinical research networks comprising multiple health care settings provide a valuable resource for rapid evaluation of large and diverse populations, with greater clinical detail than public health reporting, and larger scale and lower latency than CRF-based registries. PEDSnet is a pediatric-focused, geographically and demographically diverse network that provides services to 3% of the nation’s children (~2.5 million patients annually)5. Given the rapidly evolving pandemic, we used PEDSnet to ask whether the characteristics of severe disease in children have changed with successive phases of the pandemic in ways that would require different approaches to prevention and treatment.

Methods
Seven PEDSnet health systems were included in this study (see above). Methods for data collection, standardization, and quality assessment have been described 4, 5, 6. Data for patients with SARS-CoV-2 test results were refreshed biweekly, standardized to the PEDSnet Common Data Model. Viral PCR or antigen testing was used as the criterion for confirmed infection. Assessment of chronic conditions used the body system taxonomy of the Pediatric Medical Complexity Algorithm (PMCA) 7, considering >1 diagnoses in the 3 years prior to their cohort entry date (time of test). Potentially severe disease was defined as hospital admission within 3 days before or 14 days after a prior test, with the presence of a diagnosis associated with medical manifestations of COVID-19 8. Respiratory severity included use of mechanical ventilation or diagnosis of ARDS or respiratory failure. Non-respiratory severity included use of pressors, or diagnosis of shock, sepsis, thrombosis, SIRS, cardiac inflammation, encephalitis, or MIS-C. Multivariable logistic regression was performed using the glm( ) implementation in R 4.1, using demographic characteristics, PMCA body systems, and PEDSnet institution as categorical input variables and potentially severe disease as a categorical outcome. Results are shown as odds ratios with 95% confidence intervals.

Results
Between March 2020 and October 2021, 671,309 patients were tested for SARS-CoV-2, of whom 85% had at least 2 pre-test visits to PEDSnet; 76% of tested patients had one result; 96% had 3 or fewer tests. 51% of tested patients were male; 4% identified as Asian, 16% Black, and 16% Hispanic; 9% were <1, 29% 1-4, 33% 5-11, and 29% 12-21 years old; 25% were tested in an ED or inpatient setting; 44% had one or more chronic medical conditions. 53,600 (8.4%) had test-confirmed infection, of whom 4% met the criteria for potentially severe COVID-19. Rates of test positivity tracked the overall dynamics of the pandemic in the United States. We did not observe significant change in the frequency of severe manifestations, whether respiratory failure/ventilation or hypotension/vascular events, from the early pandemic to the time of alpha-variant predominance in late 2020-early 2021 or the emergence of delta variant in mid-2021 (Figure 1), though hospitalization rates rose transiently with peaks in prevalence. Mortality was ~0.1% of test-positive patients, and was lower during the delta-prevalent era than the alpha-prevalent era (data not shown).

Figure 1. Trends in severe illness
Overall, Black/AA and Hispanic children were overrepresented in the test-positive cohort, as were children with cancer, metabolic and endocrine conditions, but not respiratory conditions. Because previous data have established strong relationships between disease severity and both chronic medical conditions and ethnicity, we examined the relationship between the two. Multivariate modeling demonstrated increased odds of severe disease in young infants and adolescents, and in Asian and Black/AA patients (Figure 2). Similar trends were seen for many chronic disease groups, with cardiac, oncologic, neurologic, and pulmonary disorders reaching statistical significance (Figure 3). Progressive conditions carried the highest odds.

**Discussion**

Using PEDSnet, a national child-focused clinical and data network, we were able evaluate over time a range of potential risk factors for severe COVID-19. We did not observe increases in overall rates of severe disease or mortality between alpha and delta variants of SARS-CoV-2, nor were there major shifts in which children were at higher risk for severe disease. However, racial disparities in severity appeared to decrease slightly over time, particularly with children of Hispanic ethnicity. Our data suggest that while overall case and admission counts may vary with $R_0$ of different viral strains, risk of life-threatening outcomes does not increase, and changes in targeting of mitigation measures has not been indicated. We will continue to evaluate risk profiles as new variants emerge, with attention to differential risks in vaccinated vs unvaccinated and previously infected vs uninfected children.

**References**

Computerized clinical decision support for fall prevention: Defining end-user requirements for primary care staff and patients

Hannah Rice, BA1, Pamela N. Garabedian, MS2, Kristen Shear, MAJ, AN3, Zoe Burns, MPH1, Ragnhildur Bjarnadottir, PhD, MPH2, Nancy Latham, PhD1, Robert Lucero, PhD, MPH3, RN, Denise Schentrup, APRN, DNP3, Patricia C. Dykes, PhD, MA, RN1
1Brigham & Women’s Hospital, Boston, MA, USA; 2Mass General Brigham, Somerville, MA, USA; 3University of Florida College of Nursing, Gainesville, FL, USA

Introduction: Falls in community-dwelling older adults are common, and there is a lack of computerized clinical decision support (CCDS) designed to provide healthcare providers with effective, individualized fall prevention recommendations. The process of user-centered design (UCD) can address the limitations of CCDS and enhance overall system usability. UCD helps prevent design errors and usability issues by gathering input from potential end-users early and throughout the design and development process. The goal of this research is to identify end-user (primary care staff and patients) requirements through a user-centered design process for a computerized tool that will generate actionable CCDS to protect older adults from preventable falls and injuries.

Methods: Primary care staff and community-dwelling patients aged 65 or older associated with Brigham & Women’s (BWH) affiliated primary care clinics and the University of Florida (UF) Health Archer Family Health Care clinic were eligible to participate in this study. Based on the literature and previous experience, our team designed a guide for semi-structured and exploratory interviews with primary care staff and patients to elicit perspectives on the following: 1) what is needed for effective fall prevention, 2) development and use of personalized fall prevention plans in primary care, and 3) current practices for addressing pre-identified fall injury risk factors. Our team also designed an exploratory interview guide for primary care providers to facilitate a virtual workflow observation in which the provider participant could demonstrate the activities, steps, and cognitive processes involved in fall risk assessment and prevention planning using their EHR during a patient encounter. Team members conducted virtual video and in-person interviews.

Our research team identified end-user requirements for the fall prevention CCDS through content analysis. The first author independently reviewed transcripts to develop a preliminary coding framework for user requirement categories. The user experience expert reviewed the coding framework to confirm and added codes. Once this initial review was completed, the first author grouped and sorted the common requirements. The research team reviewed and validated the codes and categories. Finally, the first author grouped the categories into major themes, which were validated by the entire research team.

Results: A total of 24 primary care staff members and 18 patients participated in semi-structured and exploratory interviews. We identified six fall prevention user requirement themes for primary care staff and four fall prevention user requirement themes for patients. Themes for primary care staff included; 1) does not add significant time to office visits, 2) includes standardized resources to support recommendations, 3) helps staff work with patients to resolve ambivalence toward their fall risk, 4) facilitates systematic communication between primary care staff, 5) is based on evidence-based fall prevention recommendations, and 6) allows for in-person assessment of patient symptoms and diagnoses. Patient requirements included; 1) raise awareness of preventative strategies, 2) a need for personal support networks, 3) intrinsic and extrinsic motivation to engage with behavior change, and 4) provide expert guidance to successfully adhere to personal fall prevention practices.

Discussion/Conclusions: Identifying primary care staff and patient user requirements is a critical step in informing the design of CCDS for fall prevention management. These findings suggest that there are many care gaps in fall prevention management in primary care and that personalized, actionable, and evidence based CCDS has the potential to address these gaps. Interviews with primary care staff and patients revealed that both intrinsic and extrinsic forms of motivation are necessary to encourage adherence to fall prevention practices. While overcoming barriers to the use of CCDS is part of the solution, patients must also follow their healthcare provider’s recommendations. Our findings demonstrate the importance of designing tools centered around both direct and indirect users. In this case, the primary care staff who engage with the decision support, and the patients who are recipients of recommendations. Identifying and more fully understanding barriers to planning, implementing, and adhering to tailored fall prevention activities, will allow our research team to design a CCDS tool that meets the needs of primary care teams and patients.
<table>
<thead>
<tr>
<th>User type</th>
<th>User requirement theme</th>
<th>Sample Quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary care</td>
<td>CDS must not add time to pre-existing workflows</td>
<td>“Not that I shouldn’t [address fall prevention], but the visit is only 35 minutes, there are probably 5 prescriptions that came up to be refilled, and 2 other questions. It gets buried among a lot of other stuff.” – Provider 1</td>
</tr>
<tr>
<td>Primary care</td>
<td>Standardized resources for primary care team members to share with patients when creating fall prevention plans</td>
<td>“I would love to have a smartface or something where I could think through when someone comes in with a fall, these are the things I could do for them and this is what’s right for them.” – Provider 3</td>
</tr>
<tr>
<td>Primary care</td>
<td>Work with patient to resolve ambivalence around fall risk and to highlight reason(s) for change</td>
<td>“I find it’s very difficult, because in the population that I see, which is primarily older, people are very resistant to accepting that they have a risk for falls.” – Provider 5</td>
</tr>
<tr>
<td>Primary care</td>
<td>Systematic communication between and among care team members, patients and family</td>
<td>“Most of [fall prevention] has been communication, talking with families, and getting other services involved to help with that.” – iCMP Nurse 1</td>
</tr>
<tr>
<td>Primary care</td>
<td>Evidence based, safe exercise recommendations</td>
<td>“I really don’t want [patients] trying to do [exercise prescriptions] on their own because I’m concerned they’re going to hurt themselves.” – PA 2</td>
</tr>
<tr>
<td>Primary care</td>
<td>In-person assessment of patient symptoms and diagnoses</td>
<td>“Traditionally, before the pandemic, I would go out to the waiting room to greet my patient and escort them into my office, which represented a significant part of the examination. I got a lot of sense of how stable they were.” – Provider 5</td>
</tr>
<tr>
<td>Patient</td>
<td>Understanding of personal fall risk and awareness of personal preventative strategies</td>
<td>“I'm probably a big denier when it comes to physical stuff because I think I’m pretty strong and very active. How could somebody really assess the truth for me…it’s self-realization of [fall-risk] and how do you get someone to really realize that?” – Patient 3</td>
</tr>
<tr>
<td>Patient</td>
<td>Patient support network to encourage adherence to fall prevention plans</td>
<td>“I have a really good spouse. Since COVID, we are walking 3 to 5 miles a day. Before that, he never walked much, but I got him into walking.” – Patient 3</td>
</tr>
<tr>
<td>Patient</td>
<td>Intrinsic and extrinsic motivation to engage in and maintain behavior change</td>
<td>“I do exercise every day, but I know me, and I wouldn’t do anything that takes longer than 15 minutes…I know that if I were supposed to do 20 minutes, I probably wouldn’t do it.” – Patient 6</td>
</tr>
<tr>
<td>Patient</td>
<td>Expert guidance to trust and feel confident in fall prevention recommendations</td>
<td>“Because for that kind of advice, which I get from my physical therapist, I'm totally compliant. I do the exercises that I do religiously. I'm careful about walking, but I just follow her directions. I don't think the primary care doctor has the knowledge to do that or the time that I told you of.” – Patient 7</td>
</tr>
</tbody>
</table>

References

Enabling High-Validity Real-World Evidence in NASH

Dan Riskin, MD1,2,3; Shannon L. Rhodes, PhD4; Eric Zollars, MD, PhD4; Arun J. Sanyal, MBBS, MD5; Ricardo Dent, MD4

1Verantos Inc., Menlo Park, California, USA; 2Department of Surgery, Stanford University, Stanford, California, USA; 3Department of Surgery, Stanford University, Stanford, California, USA; 4Amgen, Inc. Thousand Oaks, California, USA 5Professor of Medicine, Virginia Commonwealth University, Richmond, Virginia, USA

Objective
In the rapidly evolving field of real-world evidence (RWE), researchers use data routinely collected via clinical encounters and/or used for reimbursement to understand and improve care. However, missingness within data collected from electronic health records (EHR) has hindered the field for certain clinical conditions where therapies are absent or where care is not reimbursed. This study examines the interplay between EHR data sources and applied technologies in non-alcoholic steatohepatitis (NASH), a disease for which there is currently no approved therapeutic in the US.

Design
Retrospective observational study using de-identified EHR data collected between 2016 and 2020.

Methods
A predefined set of clinical concepts relevant to NASH was identified by a panel of clinicians. These were extracted from structured EHR data using SQL queries (traditional approach) and from unstructured EHR data using artificial intelligence (AI) (advanced approach). Performance was evaluated against chart abstraction using standard metrics. A post-hoc analysis was performed to assess unstructured data processing with different forms of AI, specifically NLP alone versus NLP plus pattern-based inference.

Results
The dataset included 6,087 primary care clinical encounters (3,137 unique patients) from a U.S. tertiary care academic medical center. For the traditional approach, average recall, precision, and F1-score were 37.2%, 98.9%, and 54.1%, respectively. For the advanced approach, average recall, precision, and F1-score were 96.0%, 97.4%, and 96.7% respectively, in the post-hoc analysis. In the primary analysis, it was noted that NLP alone resulted in lower recall for two important concepts, liver fibrosis (68.8%) and alcohol use (63.3%) compared to other concepts (e.g. fatty liver 98.1%, jaundice 89.0%). Augmenting NLP extraction with pattern-based inference resulted in improved recall for both liver fibrosis (93.6%) and alcohol use (86.1%).

Conclusions
RWE requires an understanding of the research question being asked and domain-related data quality. In NASH, use of traditional approaches to data resulted in missingness for certain symptoms and disease diagnoses. Artificial intelligence (AI) techniques extracting information from unstructured data decreased that missingness. However, NLP alone was found to be insufficient for two key concepts. We conclude that in an underdiagnosed and undercoded condition such as NASH, AI techniques including NLP and pattern-based inference represent important tools to enable high-validity real-world evidence.
A Pilot Study on Extrapolating Level of Confidence and Correctness Between Nurses And Physicians In Interpreting Ambiguity Around Phrases With Adverse Drug Event Mentions

^Rubina Rizvi MD, PhD1; ^Brett R. South MS, PhD1; Bedda L. Rosario, PhD1; Robert Medl MA1; Karlis Draulis BS1, Elisabeth L. Scheufele, MD1; Henry J. Feldman, MD1

1IBM Watson Health, Cambridge, MA, USA

^Equal contributor first authors

Introduction Clinical notes, entered as unstructured data, are often written as telegraphic, short sentences, having acronyms, abbreviations, and inconsistencies in punctuation and grammar.1-2 This results in challenges around correct interpretation of intended meaning of a phrase and could lead to uncertainty and variable interpretations, hampering communication and potentially resulting in undesirable patient outcomes. One common problem involves ambiguity around drug mentions and a specific health condition and/or symptoms that entails lack of clarity in discerning the true relation between the mentioned drug and the condition i.e., whether it is a treatment indication, a treatment failure, or an ensuing adverse drug event (ADE).3,4 Also, medical domain knowledge could vary widely among consumers of clinical note based upon several factors such as educational background, training, specialty. As a result, there could be differences in how such ambiguous phrases are decoded by various groups of people e.g., physicians, nurses, medical student, patients. The purpose of this study is to understand how two group of clinicians i.e., physician and nurses may decode ambiguous phrases in different ways and with varied confidence levels. This study helps us discern disparities in human interpretations arising from differences in education/training and provides insights for development of NLP applications.5,6

Methods For this study, we utilized the MIMIC III v.1.4 database that represents approximately 1.3 million unstructured clinical texts. We focused on admission notes written by resident physicians for intensive care unit (ICU) patients that contained a probable ADE. We initially filtered the MIMIC III database using the NoteEvent.category selecting those notes containing “resident” and “admission”. The admission information linked to the note provided the source of the patient excluding notes where originating service was not the emergency department. To identify probably ADEs, we used the Annotator for Clinical Data Medication Adverse Event engine to process the notes, excluding those ADEs with a system probability of < 60%. We then removed any duplicate phrases, with the final system output including medication/agent ascribed to the ADE with surrounding text limit of +/- 100 characters and preceded by an “on” clause e.g., GI bleed on Coumadin. Two rounds of manual, human review were conducted to identify candidate ambiguous phrases. In the first-round reviewers applied several heuristics (i.e., complete, and meaningful sentence with common drug mentions, and removal of duplicate phrases) to identify an initial set of phrases.7 In the second round, two physicians (ELS, HJF) selected those phrases they considered most concerning having potentially critical outcomes.

Once candidate phrases were identified, we used Alchemer to construct a survey that represented 39 different ambiguous phrases along with potential interpretations of each phrase. The final version of the survey was 7 pages long with first and last page providing instructions and thank you note, respectively. There were 6 demographics related question displayed on the second page, where three questions were required i.e., familiarity with domain knowledge, highest education, and role/experience), and then 39 ambiguous phrase related questions appearing on the next 4 pages (10 questions/page except the last page.) Between 06/14/21-07/31/21 survey respondents were obtained using variety of communication channels including LinkedIn, various informatics working groups and subcommittees, as well as practicing clinicians. For each question, survey respondents were asked to choose the best interpretation of the ambiguous phrase along with their confidence measured on a 5-point Likert scale.

We examined correctness and respondents’ confidence in their interpretation of the ambiguous phrase. A ground truth created by the senior physician on our team (HJF) was used to assess correctness of respondents’ answers. Descriptive statistics were summarized as frequency and percentage for categorical data or as median and interquartile range for ordinal data, as appropriate. For each ambiguous phrase, difference in binomial proportions was performed to compare proportions of correct response between groups (physicians versus nurses). For each, ambiguous phrase, Wilcoxon-Mann-Whitney test was performed to assess differences in confidence scores between groups (physicians versus nurses). Unadjusted and adjusted (Bonferroni correction) p-values are reported to account for multiple testing. All tests were 2-sided, and the significance level was set to 0.05. Analyses were conducted using R version 4.0.1 (R Foundation for Statistical Computing 2020). Negative binomial regression was employed to investigate the association between factors of interest and the total number of correct responses. Factors examined included age (as categorical variable), gender, education, and group (physicians versus nurses).

Results There was a total of 44 completed responses (out of total 118 people accessing the survey), with survey completion rate of 37.3%. These responses were received from non-clinicians (n=7, 15.9%) and clinicians [n=37, physicians (n=30, 68.1%), nurses (n=7, 15.9%)]. Only responses from clinicians were included in the study. Most clinicians were female (n=25, 58.1%), between
55-64 years (n=13, 29.5%), and described themselves as healthcare professionals (n=39, 88.6%), with majority having advanced degrees (n=40, 90.9%).

Table 1. Number of responses aligning with the Gold Standard for ten random ambiguous phrases

<table>
<thead>
<tr>
<th>Response ID</th>
<th>Gold Standard Response (GS)</th>
<th>Responses in alignment with the GS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hx of DVTs on Coumadin</td>
<td>I have no idea what this phrase means</td>
<td>2 out of 44 (4.7%)</td>
</tr>
<tr>
<td>Atrial fibrillation on disopyramide</td>
<td>The patient has atrial fibrillation and is being treated with Disopyramide</td>
<td>31 out of 44 (70.5%)</td>
</tr>
<tr>
<td>h/o erosive gastritis on omeprazole</td>
<td>I have no idea what this phrase means</td>
<td>1 out of 44 (2.3%)</td>
</tr>
<tr>
<td>h/o GI bleed and thrombocytopenia on Plavix</td>
<td>The patient has a history of gastrointestinal bleed and thrombocytopenia, resulting from taking Plavix. Symptoms/condition likely an adverse drug reaction</td>
<td>27 out of 44 (61.4%)</td>
</tr>
<tr>
<td>Severe pulmonary hypertension on Viagra</td>
<td>The patient has pulmonary hypertension and is being treated with Viagra</td>
<td>33 out of 44 (75%)</td>
</tr>
<tr>
<td>Hx of hemorrhagic CVA on Coumadin</td>
<td>I have no idea what this phrase means</td>
<td>4 out of 44 (9%)</td>
</tr>
<tr>
<td>h/o GI/rectus sheath bleeds on coumadin</td>
<td>The patient has a history of gastrointestinal/rectus bleeds, resulting from taking Coumadin. Symptoms/condition likely an adverse drug reaction</td>
<td>31 out of 44 (70.5%)</td>
</tr>
<tr>
<td>DVT/PE s/p IVC filter on Lovenox</td>
<td>The patient has deep venous thrombosis/pulmonary Embolism and is status post Inferior vena cava filter and is being treated with Lovenox</td>
<td>35 out of 44 (79.5%)</td>
</tr>
<tr>
<td>Myelodysplastic Syndrome on decitabine</td>
<td>The patient has Myelodysplastic Syndrome and is being treated with Decitabine</td>
<td>35 out of 44 (79.5%)</td>
</tr>
</tbody>
</table>

For all questions, overall and stratified by group, the proportion of correct responses ranged from 0% to 100% (Table 1). Results showed no statistically significant difference in proportions of correct response between physicians and nurses (both unadjusted and adjusted p-values > 0.05). For all questions, the overall median (first quartile, third quartile) confidence score was 4 (3, 5) [nurse 4 (3, 5); physician 4 (3, 5)]. Results showed no statistically significant difference in confidence scores between physicians and nurses (both unadjusted and adjusted p-values > 0.05). In a negative binomial regression model, multivariable model, group was independently associated with total number of correct responses (p=.027). The total number of correct responses was greater for physicians than nurses, and doctors performed significantly better than nurses.

**Conclusion** Addressing ambiguity in unstructured clinical text is an unsolved problem for NLP systems and also represents a challenge for human interpretation. Results from our study demonstrates no observable difference in confidence scores, however number of correct responses identified (in reference to the ground truth) were greater among physicians than nurses groups. We acknowledge limitations introduced by the sample size, potential order bias in selecting a response because of partial random order of questions and having one physician expert annotating the ground truth interpretation of phrases. Future work is warranted to examine the way in which physicians and nurses apply heuristics or fall back on medical training to disambiguate these types of phrases. Understanding how clinicians address ambiguity is the first step in improving performance of NLP systems for this task.

Canary in the Care Coordination Coalmine: Specialist Utilization as a Signal of Care Coordination Among Members Before and After Direct Primary Care Exposure

Bedda L. Rosario, PhD,1 Issa Sylla, BA,2 Kristina Steinberg,3 MD, George Kim, MD,4 Yoonyoung Park, ScD,2 Ching-Hua Chen, RN, PhD,5 Sasha E. Ballen, MS,3 Judy George, PhD.1

1IBM Watson Health, Cambridge, MA; 2IBM Research, Cambridge, MA; 3Information Technology, R-Health Inc., Elkins Park, PA; 4Johns Hopkins School of Medicine, Baltimore, MD, 5IBM Research, Yorktown, NY, USA

Introduction

Direct Primary Care (DPC) is a growing value-based care delivery model intended to promote primary care that is comprehensive, patient-centered, and coordinated (i.e., high-touch).1–2 Specifically, DPC is a retainer-based delivery model that aims to improve quality while curtailing avoidable healthcare costs through better care coordination and patient engagement. Supporting the aims of DPC models, patients in such models are granted 24x7 physician office access at no additional patient out-of-pocket costs.1

Previous research in a Medicare Advantage population has shown that healthcare utilization, as measured by total healthcare costs and number of hospital admissions, was lower after one year of a high-touch care model when compared to traditional practice model, where lower costs and inpatient utilization reflected enhanced care coordination afforded by the model.2 Moreover, a recent Medicare value-based care model study found that a certain level of specialist involvement for primary care services was necessary (40–45%) to lower costs-per-patient; too little (<35%) or too much (>60%) of specialists use as a proportion of all office visits was associated with higher costs.3 The study concluded that an adequate level of specialist use is necessary to manage per-patient medical costs and implies non-specialist (primary care provider) care coordination facilitates achieving the appropriate balance of specialist/non-specialist delivery of care. This study examines differences in specialist utilization before and after DPC exposure for adults in a preferred provider organization (PPO) commercial insurance plan.

Methods

We conducted a retrospective pre-post study for members ages 18 years and older enrolled in a regional DPC program (R-Health) that offers patients 24x7 multi-modal communications (audio, text, video) in addition to office-based visits with R-Health’s primary care providers. Eligible PPO members (i.e., enrollees) of employers contracted with R-Health self-select to access R-Health primary care providers while retaining access to services from any PPO provider in their PPO network.

Study inclusion criteria included minimum 12 months of claims history prior to DPC enrollment (pre-DPC), minimum 12 months of DPC enrollment (post-DPC) and one or more DPC office-based physician visits. DPC member specialist utilization was captured from a combination of enrollment, health plan claims data, and DPC electronic health records combined at the member level from October 1, 2016 to October 29, 2020 (Figure 1). The primary outcome of interest was specialist utilization defined as the proportion of all primary care services (PCS) attended by specialists. This proportion was defined as PCS attended by specialists divided by PCS attended by specialists and non-specialists.3 Primary care services were identified by evaluation and management physician office billing codes delivered by phone, video, or in-person for new or established patients (99421-99423, 99441-99443, 99201-99215, G0402, G0438-G0439).3

We computed the specialist utilization at two time points (pre-DPC [T0], post-DPC [T1]) and the difference between time periods (T1-T0). Since the member’s first in-person office visit with a DPC provider includes a comprehensive
orientation to the DPC practice and patient intake, we defined the date of the first DPC office-based physician visit as the index date (i.e., exposure to DPC intervention).

Descriptive statistics were summarized as mean and standard deviation or median, first quartile, and third quartile for quantitative data or as frequency (%) for categorical data for member demographics and specialist utilization, as appropriate. Wilcoxon signed rank test was conducted with member as the unit of analysis to compare differences in specialist utilization as well as visit counts to specialists and all primary care services providers (i.e., specialists + non-specialists) before and after DPC engagement. All tests were 2-sided and the significance level was set to 0.05. Risk adjustment was not required since each member served as their own control (pre/post DPC, repeat measures).

**Results**

From a sample of 8,492 DPC patients, 1,472 individuals met study inclusion criteria. Table 1 summarizes descriptive statistics for the study variables. The cohort was primarily female (69.7%), mean age of (mean±std) 48±13 years, enrolled in DPC for 29.9±9.1 months, a median of 4.4 (first quartile=2.5, third quartile=8.6) miles to the physician office, and a median of 2.8 (0.9, 6.9) months from DPC enrollment to index date. The proportion of PCS visits delivered by specialist declined by 40% from 50.5% pre-DPC to 30.2% post-DPC (p<.0001) and most members decreased or had no change following DPC exposure as illustrated in Figure 2. The count of PCS visits delivered by specialists increased though not statistically significant (2.7 pre-DPC to 2.9 post-DPC, p=0.094) whereas the count of total PCS visits delivered by both specialists and non-specialists increased significantly (4.1 pre-DPC to 7.3 post-DPC, p<.0001).

**Conclusion**

In a sample of privately insured adults in a PPO health plan who self-selected into a DPC value-based care delivery model, the proportion of specialist visits declined by 40% (from 50.5% to 30.2%) in the first 12-months following R-Health engagement relative to the 12-months prior. The proportional decline is largely observed in an increase of visits to non-specialists. There was no statistically significant change in the count of specialist visits following DPC exposure. Prior research in a Medicare value-based care model found that a certain level of specialist use is necessary to curtail per member costs. More research is required to determine the proportional use of specialists for primary care office visits in a non-Medicare, PPO population and the association with per member per year costs and quality outcomes. Using the literature-based metric of specialist utilization offers the foundation for future research to examine whether and how variation in specialist utilization is associated with quality outcomes and costs for patients. This study examines an innovative indicator derived from claims and structured electronic health record data to potentially signal care coordination at the member level.

**References**


A Machine Learning Approach to Predict Decreased Opioid Prescribing Among a Cohort of 85K Providers

Jaime Y. Smith, PhD, MAE¹,²
¹Surescripts, Arlington, VA; ²George Mason University, Fairfax, VA

Abstract

This analysis examines innovative techniques, including four data mining classifiers to predict prescriber providing behavior and identify the top features for decreased opioid e-prescribing between 2017 and 2018. Prior opioid prescribing, provider location and specialty were predictive. Results from this study have the potential to identify primary drivers of provider prescribing behavior and the opioid epidemic.

Introduction

The opioid epidemic is at the forefront of the national public health crisis in the United States (U.S.)¹,². Largely spurred by abuse and misuse of prescriptions opioids, this public health emergency has motivated state legislatures to intervene with electronic prescribing (e-prescribing) related policies (as well as other regulations) aimed to reduce the impact of the opioid epidemic on populations across the country². Specifically, the states of New York (NY), Connecticut (CT), and Maine (ME) have all implemented e-prescribing mandates to address the opioid epidemic between the years of 2016 and 2018. To date, very little is known about the impact of e-prescribing laws on the opioid crisis. Therefore, this analysis examines innovative techniques, including, four data mining classifiers (Support Vector Machines, Logistic Regression, Random Forest, and Naïve Bayes) to predict prescriber providing behavior and identify the top features for decreased opioid e-prescribing between 2017 and 2018.

Methods

Prescriber-level data from a large health information network was obtained, including prescriber demographics such as prescriber state and specialty³. Prescribers must have electronically transmitted at least one opioid e-prescription in January 2017, and any e-prescription (controlled or non-controlled) at any time in 2018, in order to remain in the study. A cohort of approximately 85K providers was identified that e-prescribed nearly 46 million opioids between 2017 and 2018³. Using an 80/20 test and train data set on the following classifiers: Support Vector Machines, Logistic Regression, Random Forest, and Naïve Bayes, this analysis identifies features that predict decreased prescribing of opioid e-prescriptions for the calendar year 2018. The following features were used in the model: provider location (state), provider specialty, the number of opioids and opioid-product containing e-prescriptions (Alfentanil, Benzhydrocodone, Buprenorphine, Butorphanol, Codeine, Dihydrocodeine, Eluxadoline, Fentanyl, Hydrocodone, Hydromorphone, Levorphanol, Meperidine, Methadone, Morphine, Nalbuphine, Opium, Oxycodone, Oxymorphone, Paregoric, Remifentanil, Sufentanil, Tapentadol, and Tramadol) in 2017, and the prescribers’ total number of e-prescriptions from 2017. The research question analyzed is: Do opioid e-prescribing in the prior year and prescriber demographics predict decreased opioid e-prescriptions (defined as a greater than average decrease + 1 standard deviation) for the calendar year 2018 based on data mining techniques?

Results

Within the cohort, average annual e-prescriptions per prescriber increased by 158 e-prescriptions between 2017 and 2018. Average opioid e-prescription rates decreased from 9% of total e-prescriptions in 2017 to 8% in 2018. The Random Forest Classifier was the best performing model and it identified several features of importance, including prior year (2017) e-prescribing rates of oxycodone, hydrocodone, tramadol, codeine, hydromorphone, morphine, fentanyl, methadone, tapentadon, and buprenorphine as strong predictors of decreased opioid e-prescribing in the following year (2018) (Figure 1; Figure 2). Provider location in the state of NY was identified as an important feature to predict decreased opioid e-prescribing among the cohort, as well as the following specialties: family practice, surgery, internal medicine, physician assistants and nurse practitioners (Figure 2).
Conclusion

Results from this study will have policy implications for states that are interested in implementing e-prescribing mandates in the future. Additionally, this analysis has the potential to identify primary drivers of provider prescribing behavior and the opioid epidemic. Lastly, this analysis seeks to highlight the importance of e-prescribing and related mandates on the evolution of this technology during the opioid crisis.

References

Predicting Persistent Respiratory Sequelae in COVID-19 Patients

Zachary H. Strasser, MD, MBI1,2, Hossein Estiri, PhD1,2, Sina Rashidian, PhD1,2, Shawn Murphy, MD, PHD1,2,3
1Massachusetts General Hospital, Boston, MA; 2Harvard Medical School, Boston, MA, 3Mass General Brigham, Boston, MA

Introduction

Accurately identifying patients with temporal sequences of clinical events will assist researchers in finding specific cohorts and generating hypotheses for precision medicine research.1 This is of particular value for the coronavirus disease 2019 (COVID-19) pandemic. While COVID-19 has a defined acute period, in a subset of individuals it is followed by ongoing symptoms.2,3 In this study we develop a heuristic for identifying patients across the Mass General Brigham network who have ongoing respiratory symptoms after COVID-19, and then apply various machine learning algorithms to the pre- and acute COVID periods to predict who is most likely to have these ongoing respiratory symptoms.

Methods

First, we developed a methodology for accurately identifying the cohort of patients with ongoing respiratory symptoms after a COVID-19 infection. Then two Artificial Intelligence methods were applied to predict who among COVID-19 patients were most likely to develop ongoing symptoms. The AI methods include MLHO4 and a deep learning method consisting of Recurrent Neural Networks (RNNs) with Long short-term memory (LSTM) architecture.

We extracted patient charts with a COVID-19 diagnosis (defined by a positive COVID PCR) from the Mass General Brigham (MGB) network. We considered various surrogates as features for positive screens to identify COVID-19 patients with new or worsened respiratory symptoms. 200 charts were randomly selected and set aside, for testing the final algorithm. We then performed chart reviews with a clinician on a random sample of 100 patients from the study cohort to create gold-standard labels to confirm the sequence of ongoing respiratory symptoms after COVID-19. Then we created an UpSet plot5 to develop a heuristic for identifying the structured data elements that identify this group.

The chosen heuristic was applied across the electronic health records to label patients with ongoing symptoms. The MHLO (performs dimensionality reduction using the MSMR algorithm and trains a gradient boosting machine), and RNN models were trained on this labeled set with all structured features available to identify patients with new and ongoing respiratory disease. Then the heuristic, MHLO, and RNN were tested and compared on the test set of 200 chart reviewed patients to determine which method works the best for identifying patients with ongoing respiratory symptoms. All patient charts were then labeled as having or not having ongoing respiratory disease after COVID with the best algorithm.

Then using demographic features, as well as diagnoses, procedures, and medication records from the pre- and acute phase of COVID-19, we applied the two AI methods: MHLO and LSTM to predict the likelihood of developing ongoing respiratory symptoms after a COVID-19 infection. With these models we identified the important features that predict the likelihood of developing ongoing respiratory symptoms.

Results

99,369 patients were diagnosed with COVID-19 in the Mass General Brigham healthcare network between March 1st, 2020 and August 1st, 2021. Of these, 35,842 met our inclusion and exclusion criteria. Criteria included a positive COVID PCR test, and a new diagnosis of “Respiratory Signs and Symptoms” from the Clinical Classification Refined Software at least 21 days after a COVID-19 diagnosis was used as an initial positive screen. ~8,000 COVID-19 patients were identified with this screen. 100 charts were randomly selected for chart review, of which 46 were confirmed to have new or worsened respiratory symptoms based on clinical notes.

The 100 gold labeled charts were then organized into an Upset plot to determine which commonly accessible features (chest x-ray, CAT scan of the chest, pulmonary function test, exercise oxygen test, and/or an interstitial lung disease diagnosis) were present more than 21 days after a COVID-19 diagnosis and likely associated with ongoing symptoms (Figure 1). Based on the positive predictive values of each of these features, different heuristics were created for labeling all of the charts across the network.
Figure 1. An upset plot used to create heuristics for labels throughout the EHR with CXR (chest X-ray), CT (CAT scan of the chest), PFT (pulmonary function test), O2 (oxygen exercise test), and ILD (interstitial lung disease phenotype). Data with small patient numbers have been censored and removed.

We then used this heuristic across all the positively screened patients (excluding the test cases). The MHLO and the LSTM model were trained for identification of patients with ongoing respiratory symptom. The three models are then compared on the 200-person test set. Then using demographics, diagnoses, procedures, and medications, we used MHLO patients were likely to have the outcome of ongoing respiratory symptoms. Then using demographics, diagnoses, procedures, and medications before COVID and at the time of COVID, we predicted which patients were likely to have the outcome of ongoing respiratory symptoms. The most important features for prediction were identified for each of the heuristics (see Table 1 for an example of one of the heuristics).

Table 1. Important features for predicting ongoing respiratory symptoms post COVID-19

<table>
<thead>
<tr>
<th>Heuristic for representing ongoing respiratory disease</th>
<th>Pre-COVID Features</th>
<th>Acute COVID Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any CXR, CT, PFT, Exercise O2 procedure ordered, or ILD diagnosis made, after a COVID-19 in those with their first “Respiratory Sign and Symptom” after COVID</td>
<td>Abnormal lung field finding</td>
<td>- CXR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusion

The results of this study have the potential to improve our understanding of post COVID-19. The majority of features identified in each of the models come from the acute period, suggesting that the events during the acute period of COVID-19 are better predictors for ongoing respiratory symptoms than pre-existing features. Additionally, this provides a methodology that will be used by the Electronic Medical Records and Genomics (eMERGE) Network across academic centers nationally to identify patients who have COVID-19 and then develop ongoing respiratory symptoms. This project is actively being developed and a more refined project will be presented in March, 2022.

References

An Electronic Health Record-based Approach to Identify and Characterize Patients with Immune Checkpoint Inhibitor-associated Arthritis

Steven Tran, BS1, Luke Rasmussen, MS1, Jennifer Pacheco, MS1, Carlos Galvez, MD1, Kyle Tegtmeyer, MD1, Yuan Luo, PhD1, Jeffrey Sosman, MD1, Abel Kho, MD1, Theresa Walunas, PhD, MS1

1Northwestern University Feinberg School of Medicine, Chicago, IL

Introduction

Immune checkpoint inhibitors (ICIs) are a pillar of cancer therapy with demonstrated efficacy in a variety of malignancies. However, they are associated with immune-related adverse events (irAEs) that affect many organ systems with varying severity, inhibiting patient quality of life and in some cases the ability to continue immunotherapy.1 Research into irAEs is nascent, and without designated irAE diagnosis codes, identifying patients with adverse events poses a critical challenge for future research efforts and patient care. Among these irAEs, ICI-associated arthritis (checkpoint arthritis) is a recently identified phenomenon. This study's objective was to develop an electronic health record (EHR)-based model to identify and characterize patients with ICI-associated arthritis.

Methods

Forty-two patients with checkpoint arthritis were identified through chart abstraction from a cohort of all patients who received checkpoint therapy for cancer (n=2,612) in a single-center retrospective study. All EHR clinical codes (N=32,198) spanning the patients' lifetime to the present were extracted including International Classification of Diseases (ICD)-9 and ICD-10 for diagnoses, Logical Observation Identifiers Names and Codes (LOINC) for labs, RxNorm for medications, and Current Procedural Terminology (CPT) for procedures. Logistic regression, random forest, gradient boosting, support vector machine, K-nearest neighbors, and neural network machine learning models were trained to identify checkpoint arthritis patients using these clinical codes using a 50/50 test/training set split (stratified to include similar case/non-case ratios) and 5-fold cross-validation. Models were evaluated using receiver operating characteristic and precision-recall curve area under the curve (ROC-AUC, PRC-AUC) as well as sensitivity, specificity, and positive and negative predictive value (PPV, NPV) optimizing Youden’s J. Bootstrapped 95% confidence intervals were calculated for all metrics. The most important variables were determined from the logistic regression coefficients and random forest relative feature importance. Models were retrained on smaller fractions of the important variables from logistic regression to determine the minimum variable set necessary to achieve accurate identification of checkpoint arthritis.

Table 1. Performance metrics for models trained on the top 62 clinical codes. AUC was calculated from the ROC curve and PRC Sensitivity, specificity, PPV, and NPV were determined at the threshold maximizing Youden’s J. Bootstrapped 95% confidence intervals were calculated for performance metrics. AUC = area under the curve, ROC = receiver operating characteristic, PRC = precision-recall curve, PPV = positive predictive value, NPV = negative predictive value

<table>
<thead>
<tr>
<th>Model</th>
<th>ROC-AUC</th>
<th>PRC-AUC</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.928</td>
<td>0.626</td>
<td>0.846</td>
<td>0.510</td>
<td>0.144</td>
<td>0.997</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.953</td>
<td>0.671</td>
<td>0.955</td>
<td>0.919</td>
<td>0.171</td>
<td>0.999</td>
</tr>
<tr>
<td>Gradient Boosting</td>
<td>0.931</td>
<td>0.665</td>
<td>0.804</td>
<td>0.649</td>
<td>0.219</td>
<td>0.996</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>0.948</td>
<td>0.681</td>
<td>0.891</td>
<td>0.512</td>
<td>0.170</td>
<td>0.998</td>
</tr>
<tr>
<td>K-Nearest Neighbor</td>
<td>0.800</td>
<td>0.618</td>
<td>0.670</td>
<td>0.984</td>
<td>0.406</td>
<td>0.995</td>
</tr>
<tr>
<td>Neural Network</td>
<td>0.903</td>
<td>0.588</td>
<td>0.783</td>
<td>0.917</td>
<td>0.145</td>
<td>0.996</td>
</tr>
<tr>
<td>SVM Anomaly Detection</td>
<td>0.943</td>
<td>0.449</td>
<td>0.891</td>
<td>0.893</td>
<td>0.130</td>
<td>0.998</td>
</tr>
</tbody>
</table>

Results

Logistic regression and random forest were the highest performing models on the full variable set of 32,198 clinical codes (AUCs: 0.911, 0.894, respectively). Retraining the models on smaller fractions of the most important variables demonstrated peak performance using the top 31-62 clinical codes, or 0.1-0.2% of the total variables. AUCs for logistic regression and random forest using the top 62 codes were 0.942 and 0.950, respectively (Table 1). The most important
features for both machine learning algorithms, all positively associated with checkpoint arthritis, included: presence of erythrocyte sedimentation rate (ESR), C-reactive protein (CRP), rheumatoid factor lab, creatine kinase lab, thyroid labs, joint pain, prednisone, and immunization (Figure 1).

Discussion and Conclusion

Our study demonstrates that a machine learning approach based on structured EHR data can be used to identify checkpoint arthritis patients. The high performance of the models using only the 0.1-0.2% most important variables suggests that only a small number of clinical attributes are needed to identify these patients. The variables most important for identifying checkpoint arthritis included several unexpected clinical features, such as thyroid labs which are generally associated with endocrine complications of immunotherapy and immunizations, indicating potential underlying irAE associations that warrant further exploration. Features in the top 10% of importance tended to rank highly in both logistic regression and random forest models, while those ranking below the 90th percentile had scattered rankings between the two models. The flexibility of this approach and its effectiveness for identification of checkpoint arthritis suggests that it could be applied to identify and characterize other irAEs. Limitations of this work include the large imbalance between checkpoint arthritis cases and non-cases as well as the imbalance between predictors and samples. Unfortunately, the case/non-case imbalance is a property of checkpoint arthritis and explains the high ROC-AUC and modest PRC-AUC and PPV. The small sample size is due to the relatively few number of patients who have cancer and receive checkpoint therapy. Another limitation is the synonymy, and therefore collinearity, among the predictors used. This was a side effect of using terminal clinical codes, many of which refer to the same or similar medical concepts. Future steps of this work are to 1) test this approach at an external site to address the small sample size, determine model generalizability, 2) further explore the association between the unexpected clinical features, thyroid labs and immunizations, and development of checkpoint arthritis, and 3) investigate data-driven methods of combining synonymous clinical codes (across diagnoses, labs, meds, and procedures) into meaningful groups.

Figure 1. The 31 most important variables determined by the logistic regression (A, coefficients) and random forest (B, relative importance) models.

References

Domain Adaptation of a Deep Learning Symptom Extractor for Different Patient Populations and Clinical Settings

Grace Turner, Kevin Lybarger, PhD, Alison Brennan, Elizabeth J. Chang, MD, Naniella Dorvall, Jagjeet Gill, Vy Huynh, Kylie Kerker, Jolie Shen, Erica Qiao, Matthew Thompson, MD, Meliha Yetisgen, PhD
University of Washington, Seattle, WA, USA

Introduction

Symptoms and their characteristics are vital for appropriately diagnosing illnesses. Although diagnosis codes are regularly stored as discrete data in the Electronic Health Record (EHR), symptom information is primarily found in free-text clinical notes. Utilizing this text-encoded symptom information in secondary use applications, like large-scale studies and clinical decision support (CDS) systems, requires extraction of the symptom information from notes. In our prior work, we developed a deep learning symptom extraction model that was trained on the COVID-19 Annotated Clinical Text (CACT) Corpus. This corpus includes clinical notes with detailed symptom annotations. It is biased to include high rates of COVID-19 infection, so the corpus frequently includes COVID-19 symptoms (e.g. shortness of breath and fever). In this abstract, we investigate the generalizability of this model to lung and ovarian cancer patient populations. Patients with lung and ovarian cancers experience complicated symptomatic profiles over time and may benefit from a CDS system that utilizes automatically extracted symptom information.

Methods

CACT has 1,472 annotated notes - 1,028 train ($C_{train}$) and 444 test ($C_{test}$). This corpus was sampled from clinical notes of 230,000 patients who received treatment at the University of Washington Medical Center (UWMC) between May-June 2020. Of those patients, 28,000 had at least one COVID-19 test result. CACT includes telephone encounters notes, progress notes, and emergency department notes. CACT uses an event-based annotation scheme, where each symptom event includes a “trigger” that specifies the specific symptom (e.g. pain or cough) and arguments that characterize the symptom (e.g. assertion, duration, or anatomy). Four medical students annotated the corpus. The trigger inter-rater agreement was 0.85 F1.

To adapt the symptom extractor to the lung and ovarian cancer domains, we created two additional annotated data sets: Lung Cancer Annotated Clinical Text (LACT) and Ovarian Cancer Annotated Clinical Text (OACT). These new data sets were annotated using the same event-based annotation scheme as CACT. LACT utilized notes from an existing dataset of 4,673 lung cancer patients who were diagnosed with lung cancer at UWMC and Seattle Cancer Care Alliance (SCCA) between 2012-2020. From this data set, we randomly sampled and annotated 270 notes from the 24 months prior to cancer diagnosis. The sampled note types include admission, discharge, progress, and emergency room notes. LACT was randomly split into 100 test notes ($L_{test}$) and 170 training notes ($L_{train}$). OACT utilized notes from an existing dataset of 173 ovarian cancer patients who were diagnosed at UWMC and SCCA between 2012-2021. From this data set, we randomly sampled and annotated 220 notes from the 12 months of notes prior to cancer diagnosis. The sampled note types include progress, admission, discharge, gynecology, and emergency room notes. OACT was randomly split into 100 test notes ($O_{test}$) and 110 training notes ($O_{train}$). LACT was annotated by five medical students, four of which also annotated OACT. The inter-rater trigger annotation agreement is 0.83 F1 for LACT and 0.82 F1 for OACT.

The symptomology of patients varies across the three explored domains: COVID-19, lung cancer, and ovarian cancer. The symptom extractor uses a neural, span-based extraction architecture that jointly (i) extracts trigger and argument spans and (ii) predicts the linkages between triggers and arguments. In our annotation scheme, the trigger identifies the specific symptom experienced by the patient (e.g. pain) and is the most important of the annotated phenomena. We explore the domain adaptability of the symptom extraction architecture and data sets by experimenting with different combinations of out-domain and in-domain training data, evaluating performance on the domain-specific test sets. Triggers were defined as a true positive if there was an exact match between gold and predicted trigger spans. Performance is evaluated using precision (P), recall (R), and F1. Changes in performance were further evaluated on a unique trigger level.
Results

Table 1 presents the performance on test datasets for different combinations of training data. When only $C_{train}$ is used, the recall is 0.13-0.14 lower for $L_{test}$ and $O_{test}$ than for $C_{test}$. The inclusion of the in-domain training data improves the recall by 0.11-0.12, resulting in similar recall as $C_{test}$. This demonstrates that a relatively small amount of in-domain training data can significantly improve performance.

Table 2 presents the top 15 symptoms with the largest impact on overall recall. Several key symptoms for lung cancer (e.g., coughing, adenopathy) and ovarian cancer (e.g., bleeding, distended) are listed among those top ranked symptoms. Incorporating the in-domain data improved recall for previously absent symptoms (e.g. rubs), as well as for frequent symptoms in the out-domain data (e.g., pain).

Table 2: Top 15 symptoms with largest recall impact. TP represents true positives. Key symptoms for lung cancer and ovarian cancer are bolded.

<table>
<thead>
<tr>
<th>Most Improved Triggers in $L_{test}$</th>
<th>Most Improved Triggers in $O_{test}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trigger Span</td>
<td>Δ TP</td>
</tr>
<tr>
<td>pain</td>
<td>18 (261 → 279)</td>
</tr>
<tr>
<td>constipation</td>
<td>9 (14 → 23)</td>
</tr>
<tr>
<td>gallops</td>
<td>8 (1 → 9)</td>
</tr>
<tr>
<td>lesions</td>
<td>7 (2 → 9)</td>
</tr>
<tr>
<td>masses</td>
<td>6 (0 → 6)</td>
</tr>
<tr>
<td>problems</td>
<td>6 (1 → 7)</td>
</tr>
<tr>
<td>cyanosis</td>
<td>5 (1 → 11)</td>
</tr>
<tr>
<td>edema</td>
<td>5 (40 → 45)</td>
</tr>
<tr>
<td>rubs</td>
<td>5 (0 → 5)</td>
</tr>
<tr>
<td>coughing</td>
<td>5 (3 → 8)</td>
</tr>
<tr>
<td>weight loss</td>
<td>5 (22 → 27)</td>
</tr>
<tr>
<td>adenopathy</td>
<td>5 (0 → 5)</td>
</tr>
<tr>
<td>murmurs</td>
<td>5 (0 → 5)</td>
</tr>
<tr>
<td>suicidal ideation</td>
<td>4 (1 → 5)</td>
</tr>
<tr>
<td>rash</td>
<td>4 (37 → 41)</td>
</tr>
</tbody>
</table>

Discussion

Generalizability is one of the fundamental challenges of clinical information extraction and is attributable to the heterogeneity of clinical text across patient populations and clinical settings. In our experimentation, symptom extractor recall drops markedly when transitioning from the COVID-19 domain to lung and ovarian cancer domains. We adapted the COVID-19-focused model to the lung and ovarian cancer settings by augmenting the training set with a relatively small amount of in-domain data. The inclusion of the in-domain data returns performance to comparable levels.

References


Acknowledgements

This work was supported by the CanTest Collaborative, which is funded by Cancer Research UK [C8640/A23385], the Gordon and Betty Moore Foundation (Development and pilot testing of a measurement tool for diagnosis of lung cancer), NIH/NCI (1R21CA258242), and NIH/NLM (Biomedical and Health Informatics Training Program - T15LM007442).

Table 1: Domain adaptation performance for symptom trigger extraction.
Introduction
Real time healthcare service delivery through remote technology, known as telehealth or telemedicine, is a cost-effective alternative to the traditional face-to-face consultations between provider and patient. Benefits include greater access to care due to the flexibility to provide services in the evening or weekends, timely diagnosis of conditions, and the potential for closing gaps in care. At the same time, telemedicine can reduce the need for patients to seek care in potentially unnecessary care settings, such as the emergency department, or foregoing care due to lack of access. Telemedicine can be especially beneficial to individuals who live in rural and provider shortage areas because they face more barriers to seek and receive timely face to face treatment by primary care physicians and specialists. However, little is known about telemedicine adoption and take-up among individuals with and without mental health conditions.

Objective
The goal was to examine the growth in the number of Medicaid enrollees receiving care at providers who deliver telemedicine, enrollees receiving telemedicine, the growth in the number of enrollees with a mental health visit, and the growth in mental health visits delivered through telemedicine in urban, suburban, and rural counties.

Methods
Medicaid claims data for Medicaid Fee for Service and Managed Care Organization (MCO) members were obtained from Texas Health and Human Services Commission. The claims include all inpatient and outpatient care interactions, and are complemented with enrollees’ demographic, plan enrollment, and eligibility information. Our study sample included de-identified Medicaid claims data from September 1, 2012 to August 31, 2018 (reflecting the collection of data during each state fiscal year) for Medicaid members enrolled due to a disability.

We examined growth in total enrollees, enrollees with a mental health condition, enrollees with a telemedicine encounter, mental health enrollees with a telemedicine encounter, volume of mental health visits, mental health telemedicine visits, and number of telemedicine providers. Enrollees were identified based on whether they had a telemedicine visit or saw a provider who delivered telemedicine. All data was stratified by urban, suburban, and rural counties to examine differential growth rates by geography. We describe our sample by first presenting enrollee characteristics for the full sample and then separately by telemedicine and non-telemedicine take-up. We used T-tests to test for differences in socio demographic characteristics across enrollees. We then perform Analysis of Variance (ANOVA) to evaluate whether the growth rate in enrollees from state fiscal year 2013 to 2018 were different based on geography. Specifically, we estimate whether the average yearly growth rate was different by geography, telemedicine vs. non telemedicine, and mental health vs. non-mental health. We then use the ANOVA Tukey test to evaluate difference in growth rates between these groups. We evaluate growth rates from state fiscal year 2014 to 2018 where the year 2013 is the first baseline year from which we calculate growth rates. Lastly, to limit the impact of outliers in growth rates across years within the sometimes small groups, the Tukey test makes comparisons between the log of the growth rates.

Results
There were 519,454 enrollees in our sample and 9% (48,024 enrollees) received at some point care delivered through telemedicine. The average age of the sample was 37, 50% were female, and 28% were dual eligible (Medicare and Medicaid coverage). Little difference in demographic characteristics emerged between enrollees who received telemedicine and those that did not. About 80% of the tele and non-telemedicine enrollees lived in urban and suburban counties, though a larger share of telemedicine enrollees lived in rural counties compared to non-telemedicine enrollees (13% vs 10%, p<0.01). On average, an enrollee had about one visit with a primary diagnosis of a mental health condition. Overall, 18,978 telemedicine visits were delivered per year, with the majority of telemedicine visits being related to mental health as about 72% of visits had a primary diagnosis for a mental health
condition. Table 1 displays the number of total enrollees, enrollees with a mental health diagnosis, enrollees with a telemedicine visit, and enrollees with telemedicine visit with a mental health diagnosis in each state fiscal year. The table also displays the growth rates in enrollees across all outcomes from 2013 to 2018. The number of enrollees treated by a provider who delivered telemedicine grew by 132%, with the largest growth rate in urban and suburban enrollees (255% and 107%, $p<0.05$), and rural enrollee growth being the slowest with 32%. Around one third of all enrollees were diagnosed with mental health conditions across all years. In 2013, 46% out of all enrollees had at least one mental health visit and 38% of enrollees had at least one mental health visit in 2018. Growth in mental health enrollees was smaller than total enrollee growth (90% relative to 255%), and urban and suburban growth rates of total enrollees (152% and 72%) with a mental health visit surpassed the rural growth rate of 25%. Comparing growth rates of mental health enrollees to non-mental enrollees suggests statistically different growth rates ($p<0.06$) though not different by geography ($p<0.51$).

<table>
<thead>
<tr>
<th>All Clients</th>
<th>Clients with a Mental Health Visit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>Urban</td>
</tr>
<tr>
<td>2013</td>
<td>54,455</td>
</tr>
<tr>
<td>2014</td>
<td>65,844</td>
</tr>
<tr>
<td>2015</td>
<td>78,909</td>
</tr>
<tr>
<td>2018</td>
<td>126,580</td>
</tr>
<tr>
<td>Growth</td>
<td>132.4%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clients with a Telemedicine Visit</th>
<th>Clients with a Telemedicine Clients with a Mental Health Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>Urban</td>
</tr>
<tr>
<td>2013</td>
<td>5,596</td>
</tr>
<tr>
<td>2014</td>
<td>6,244</td>
</tr>
<tr>
<td>2015</td>
<td>7,741</td>
</tr>
<tr>
<td>2016</td>
<td>8,818</td>
</tr>
<tr>
<td>2017</td>
<td>9,480</td>
</tr>
<tr>
<td>2018</td>
<td>10,145</td>
</tr>
<tr>
<td>Growth</td>
<td>81.3%</td>
</tr>
</tbody>
</table>

Table 1. All Clients, Clients with a Mental Health Visit, Telemedicine Clients, and Telemedicine Clients with a Mental Health Diagnosis

Conclusion

Growth in telemedicine delivery was strong in Texas from 2012 to 2018. We observe higher telemedicine growth for enrollees in urban and rural areas and those who have been diagnosed with a mental condition. Of importance is that those enrollees with a mental health condition and who receive at least some care through telemedicine have more mental healthcare visits compared to enrollees who seek mental health care only in the face to face setting. Higher rates of mental health visits for enrollees that use telemedicine suggests that telemedicine visits may not be substituted for face-to-face visits or that these enrollees may be generally less healthy and require additional care.

References

Mondo Disease Ontology: Building a Community-Based Disease Resource
Nicole A. Vasilevsky, PhD1, Sabrina Toro, PhD1, Nico A. Matentzoglu, PhD2, Dazhi Jiao, MS3, Nomi L. Harris, MS4, Melissa A. Haendel, PhD1, Peter N. Robinson, MD5, Chris Mungall, PhD4
1University of Colorado, Anschutz Medical Campus, Aurora, CO; 2Semanticly, Athens, Greece; 3Johns Hopkins University, Baltimore, MD, 4Lawrence Berkeley National Laboratory, Berkeley, CA; 5The Jackson Laboratory, Farmington, CT

Introduction

The Mondo Disease Ontology (Mondo) integrates and harmonizes multiple disease terminologies into a coherent logic-based ontology that provides precise semantic mappings between terms (1). Mondo is used by many groups for their research and clinical ontologies, databases, and portals (mondo.monarchinitiative.org/pages/users/). For example, ClinGen (clinicalgenome.org/) curators use Mondo terms in their curation of genes and variants of clinical relevance to aid in precision medicine. The INCLUDE Data Coordinating Center (DCC) (includedcc.org/) uses Mondo for the standardization and harmonization of Down Syndrome patient data. We work closely with the research and clinical communities to keep Mondo current, and to provide a resource that is relevant and optimized for a variety of user needs and preferences regarding terms labels, terms classification, and ontology representation.

We previously reported on the creation of Mondo (2). Here we describe recent efforts to address specific users’ needs: 1) the revision of the top-level classification of the disease entities to create a better representation for clinical users, 2) the creation of an application-specific synonym subset for use in ClinGen Gene Curation Interface, and 3) the revision and enhanced logical classification of chromosomal anomalies for the INCLUDE DCC.

Methods

Clinically oriented-based view of Mondo: Per request by medical experts, we reviewed and revised the classification of Mondo to correspond to the “Harrison's Principles of Internal Medicine” (Harrison) textbook (3) organization. We identified the Mondo high-level terms to be removed from the high-level classification, new high-level grouping terms to be created, and high-level terms to be excluded from this clinically oriented-based view.

User-specific synonym tags: ClinGen users have specific preferences for disease naming that sometimes differ from the primary names or labels in Mondo. ClinGen preferred labels were added to Mondo as exact synonyms which were annotated with a new synonym tag called ‘clingen_preferred’.

Axiomatization of chromosomal anomalies: We leveraged the use of Dead Simple Ontology Design Patterns (DOSDP) (4) to consistently and automatically apply equivalence and subclass axioms to the ‘chromosomal anomaly’ branch of the ontology. Terms were defined using axioms representing the type of chromosomal anomaly and the chromosome/chromosome region at the root of the disease. The use of chromosome/chromosome region in the axioms required conversion of chromosomes and chromosome bands data from UCSC Genome Browser into an OWL classification called Monochrom (github.com/monarch-initiative/monochrom). All of the changes described above were implemented using Protege ontology editing software.

Results and Discussion

Mondo provides multiple ways to classify a disease based on a variety of criteria (such as the etiology, the origin, the anatomical structure affected, etc., of the disease), and its hierarchical classification allows for multi-inheritance. While this classification is useful for computational purposes, a simplified and more practical representation may be preferred for some users (5,6). We aimed to create a representation of Mondo that would be more practical and easier for clinicians to browse based on their approach to disease classification. The “Harrison's Principles of Internal Medicine” (Harrison) textbook being a landmark in medicine (7), we created a clinically oriented view of Mondo mirroring the way diseases are organized in this textbook. This new Harrison view will be offered in a Monarch instance of the Ontology Lookup Service (OLS) (ols.monarchinitiative.org/index) and for download in upcoming releases, available at github.com/monarch-initiative/mondo/releases (8).

Disease nomenclature varies greatly between communities, and there are so far no official guidelines. It would be optimal and more practical for a group to use a version of Mondo in which term labels use the group’s preferred
nomenclature. Here, we report the ongoing work with the ClinGen team. The nomenclature used by ClinGen is added to Mondo terms as an exact synonym, which is annotated with the new synonym tag called ‘clingen_preferred’. These tags will allow the creation of a Mondo representation in which the displayed disease term names are the preferred ClinGen nomenclature. This will enable optimal search and retrieval of Mondo terms by the ClinGen curation team. A similar process could be used to enable other groups to access Mondo via their preferred nomenclature.

The INCLUDE DCC provides access to data, analysis tools, and resources for the Down syndrome community, and standardizes and harmonizes Down Syndrome patient data using Mondo. This group needed a more precise representation of Down Syndrome subtypes, which motivated the revision of the ‘chromosomal anomaly’ branch of Mondo. We leveraged the use of patterns that we created using axioms representing the type of chromosomal anomaly and the chromosome/chromosome region at the root of the disease. The use of patterns and axioms will ensure consistency and automatic classification of the chromosomal anomaly terms, which will be available in an upcoming release.

We constantly work with the community to assess their needs. We gather community input and feedback via GitHub (https://github.com/monarch-initiative/mondo/issues), and during workshops which we organize on a regular basis. Requests are prioritized for future development depending on user needs. We invite the community to get involved and contribute to Mondo; visit github.com/monarch-initiative/mondo for details, including upcoming workshops and how to join our mailing list.

References:


Deep Clinical Phenotyping of Race and Ethnicity-Stratified Alzheimer’s Disease Patients Leveraging Electronic Medical Records Data

Sarah Woldemariam, PhD1, Alice Tang, BS1, Tomiko Oskotsky, MD1, Kristine Yaffe, PhD2, Marina Sirota, PhD1

1 Bakar Computational Health Sciences Institute, University of California San Francisco, San Francisco, California, USA; 2 Department of Psychiatry and Behavioral Sciences, University of California, San Francisco, San Francisco, USA

Introduction

Alzheimer’s disease (AD) comprises 60-80% of dementia cases1, with tremendous patient, family, and societal cost that is expected to increase dramatically with greater prevalence2. Moreover, AD disproportionately impacts Black Americans, who have at least 40% increased risk of developing late-onset AD relative to Whites and Asians3. With an abundance of clinical data available, it is essential to deeply investigate the Race and Ethnicity (R&E)-stratified risk factors and pathogenesis of AD to aid in preventing disease and developing a cure. Electronic medical records (EMR) are a rich source of data that can be used to characterize AD-associated clinical phenotypes. While EMR have been used to study sex-specific differences in AD, R&E-specific differences have not been explored4. Here we leverage EMR data from UCSF, a quaternary care hospital5, to provide insight into R&E-stratified AD comorbidities.

Methods

Patient cohorts were identified from >5 million patients in the UCSF OMOP EMR database, which contains information on patient encounters since 1982. Due to deidentification, dates are shifted by at most a year and all patients over 90 years old are represented as 90 years old. AD patients 65 years or older were identified by ICD-10-CM codes G30.1, G30.8, or G30.9. Control cohorts were identified from all other patients by propensity score (PS) matching (matchit R package) on sex, estimated age, R&E extracted from UCSF EMR, and death status at a 1:2 AD:control ratio. To evaluate comorbidities, ICD-based diagnoses recorded for our patients were identified with the earliest entry of every diagnosis. Diagnoses were converted into phecodes6, compared using the phecode-corresponding phenotype name, and grouped by phecode categories. To compare R&E differences, a sub-cohort with equal numbers of Black, Asian, and White patients and matched controls (maintaining a 1:2 AD:control ratio) were identified after additional rounds of PS matching that matched R&E-stratified patients on sex, estimated age, and death status. For each phenotype, the number of Black and White patients and the number of Asian and White patients were separately compared with matched R&E controls using Fisher Exact (if <5 patients in any category) or Chi Square tests to identify significant phenotypes among each R&E and visualized with log-log odds ratio plots and Miami plots.

Results

![Figure 1: Black vs. White AD comparison. (A) Phenotypes were compared between AD and controls within each R&E, using Fisher Exact and Chi Square tests. The log values of the odds ratios are plotted on the axes, and the points colored by significance (Bonferroni corrected, p-value cutoff < 4e-5). (B) Percentage of phenotypes in each phecode category that is significant for Black (top) and White (bottom) patients. (C) Miami plot of phenotypes grouped by R&E and phecode categories, with log_{10} of p-values from Fisher Exact and Chi Square tests.](image)

633
We identified 7,409 AD and 14,818 PS-matched control patients. When comparing Black and White patients (422 AD and 844 control patients for each R&E) (Figure 1), we found 124 Black-specific and 9 White-specific significant comorbidities. Top Black-specific AD comorbidities include kidney and ureter disorders and hypertensive complications. We also found that Black AD patients are more likely to have hypercholesterolemia, hyperlipidemia, and hypertension relative to White patients. When comparing Asian and White patients (422 AD and 844 control patients for each R&E) (Figure 2), we found 48 Asian-specific and 25 White-specific significant comorbidities. Top Asian-specific AD comorbidities include multiple diabetes phenotypes. Asian AD patients are also more likely to have vascular dementia relative to White AD patients. Interestingly, late effects of cerebrovascular disease is both an Asian-specific and Black-specific significant comorbidity when separately analyzed with White AD patients. Also, White patients specifically were more likely to have bipolar disorder and teeth-related disorders for both comparisons.

Figure 2. Asian/White AD comparison. (A) Phenotypes were compared between AD and controls within each R&E, using Fisher Exact and Chi Square tests. The log values of the odds ratios are plotted on the axes, and the points colored by significance (Bonferroni corrected, p-value cutoff < 4e-5). (B) Percentage of phenotypes in each phecode category that is significant for Asian (top) and White (bottom) patients. (C) Miami plot of phenotypes grouped by R&E and phecode categories, with log_{10} of p-values from Fisher Exact and Chi Square tests.

Discussion

Many of the R&E-specific significant comorbidities we found have been previously suggested to disproportionately impact Black Americans and Asians. For example, studies suggest that Black AD patients have higher hypertension rates relative to White AD patients\(^1\), while higher vascular dementia rates are found in Asia\(^6\). Our analysis also shows higher rates of cardiovascular risk factor-related comorbidities in Black AD patients beyond hypertension, such hypercholesterolemia and hyperlipidemia. Although there are limitations in ways R&E is captured in the EMR, taken together, our findings demonstrate how EMR can be used to comprehensively explore R&E disparities in AD comorbidities\(^1\). We are currently analyzing R&E-stratified comorbidity networks, comparing significant comorbidities for Latinx patients, and expanding our analyses UC-wide, which has over 19,000 AD patients beyond UCSF. This work will allow other researchers to deeply explore the relationships between AD and R&E-specific comorbidities, with the goal of targeted preventative interventions to reduce R&E health disparities in AD prevalence.

References

Reversal of infected host gene expression identifies repurposed drug candidate for SARS-CoV-2 and its new variants

Jing Xing1, Rama Shankar1, Meehyun Ko1, Aleksandra Drellich2, Shreya Paiithankar1, Eugene Chekalin1, Thomas Dexheimer3, Mei-Sze Chua4, Surender Rajasekaran1,5, Chien-Te Kent Tseng1,6, Seungtaek Kim7, Bin Chen1,3,*

1Department of Pediatrics and Human Development, Michigan State University, Grand Rapids, Michigan, USA; 2Departments of Microbiology and Immunology, University of Texas Medical Branch, Galveston, Texas, USA; 3Department of Pharmacology and Toxicology, Michigan State University, Grand Rapids, Michigan, USA; 4Department of Surgery, Stanford University School of Medicine, Palo Alto, California, USA; 5Helen Devos Children Hospital, Grand Rapids, Michigan, USA; 6Center of Biodefense and Emerging Disease, University of Texas Medical Branch, Galveston, Texas, USA; 7 Zoonotic Virus Laboratory, Institut Pasteur Korea, Korea

#These authors contributed equally. *Correspondence to Bin Chen: chenbi12@msu.edu

Abstract

Emerging SARS-CoV-2 variants impose increasing health care burdens due to high contagiousness. Using the robust host gene expression signature derived from SARS, MERS, and SARS-CoV-2, we discovered a repurposing candidate for SARS-CoV-2 and likely its variants and then validated its efficacy in preclinical models. Following analyses of drug-treated RNA-Seq showed that this drug targets multiple genes and pathways important for viral infection, highlighting the significance of our computational approach.

Introduction

Since early December 2019, the newly emerged SARS-CoV-2 has infected almost 214 million people globally and claimed more than 4 million deaths. Several vaccines are available for SARS-CoV-2, but different emerging variants impose increasing threats due to higher infections and severity. Thus, there is utmost need to discover novel repurposing drugs to inhibit viral infection irrespective of new variants. Viral infection involves multiple biological processes inside the host cell, which could be targeted at a system level (Figure 1A). Here, we propose a computational approach to discover new drugs against viral infection using SARS, MERS, and SARS-CoV-2-induced host gene expression signatures. We hypothesize that the drug targeting the robust host signature of multiple coronaviruses is able to target the host signature of new variants.

Methods

Given the assumption that compounds showing a reversed transcriptomic profile of the disease dysregulated genes could alleviate the disease symptoms1, we collected gene expression profiles of the samples infected by SARS, MERS or SARS-CoV-2 and computed differentially expressed genes (namely disease signature) for all these three viruses in individual datasets. The drug repurposing library with their gene expression change is collected from the LINCS L1000 project. Drugs with known inhibition on the three viruses were employed to select valid signatures, resulting in a meta signature of three coronaviruses that was later used further for drug repurposing. Then, we tested the antiviral effect of the candidates in Vero E6 cells and Calu-3 cells and identified one promising candidate. Further, RNA-seq of the Calu-3 cells with the challenge of SARS-CoV-2 prior to the treatment of our final candidate was performed to evaluate their transcriptome reversal and elucidate the drug mechanism of action.

Results

In total, 12 RNA-Seq datasets of SARS or MERS infected host response were processed, resulting in 215 infection status signatures (Figure 1B.1), among which 13 were selected based on their recall of known anti-coronavirus compounds (Figure 1B.2). To propose new drug repurposing candidates for SARS-CoV-2, we applied a consensus ranking of the 13 disease signature models when evaluating all 1740 repurposing candidates in the library (Figure 1B.3). We incorporated 49 RNA-Seq datasets of SARS-CoV-2 infection to generate a robust coronavirus signature set of 23 SARS/MERS/SARS-CoV-2 infection models. On an external positive drug set, the robust coronavirus signature scores significantly correlated with drug EC50 values (Spearman R = 0.46, P = 0.005, Figure 1B.5). Based on these disease signatures, we predicted drug candidates and selected 11 drugs to test their anti-SARS-CoV-2 activity in vitro (Figure 1B.4). The most potent drug among all proposed drugs was identified as C11, which inhibited viral replication with an IC50 of 87 nM, and its CC50 was 21 μM in Vero cells (Figure 1C). The IC50 of this drug was 520
nM with no observable cytotoxicity in Calu-3 cells (Figure 1C). RNA-Seq results of drug-treated samples showed that 0.5 µM of C11 could reverse the SARS-CoV-2 induced host transcriptomic change (Figure 1D). To gain insights into the antiviral mechanism of C11, we built a protein-protein interaction (PPI) network including C11-binding and virus-binding host proteins. A few hub proteins (NF-κB, JUN, TMPRSS2, BCAT2) in this network were suggested as potential targets of the inhibitor (Figure 1E). These targets were also identified as “pro-viral” genes in different CRISPR screenings2-4 (Figure 1F). These findings suggest common biological changes exist among SARS, MERS, and SARS-CoV-2 and could be extrapolated to new variants and future coronaviruses.


Conclusion

The robust coronavirus infection signatures created by merging the disease signatures of SARS, MERS, and SARS-CoV-2 led us to discover a potent drug candidate with anti-SARS-CoV-2 activity. This drug targets multiple pathways involved in the viral life cycle, suggesting the power of this system-based approach for drug discovery. This approach provides an advantage over the other drug discovery methods, where a single pathway or gene is targeted. This method provides an additional option to explore the drugs for new variants and future pandemics.

References

Observability and Its Impact on Differential Bias for Clinical Prediction Models

Mengying Yan, MS1, Michael J. Pencina, PhD1, Benjamin A. Goldstein, PhD1
1Department of Biostatistics & Bioinformatics, Duke University, Durham, NC, USA

Introduction

Attention has been paid to algorithmic fairness in prediction decision making1,2. When bias exists along a spectrum of demographics factors (e.g., race, sex, age), herein called sensitive variables, it can create and perpetuate health inequities and algorithmic unfairness. Using electronic health records (EHR) data with incomplete capture of patient outcomes (i.e., limited observability) to develop clinical prediction models (CPMs) can induce such biases. Here we focus on differential observability, that the limited observability is different across a sensitive variable, and how differential bias raises. There has been question of whether one should include sensitive variable in CPM3,4. We use simulations to demonstrate the possible impact of including a sensitive variable in a CPM.

Methods

Observing an outcome of interest (e.g. readmissions) within an EHR involves a two-stage process (Figure 1). Stage 1 is the process by which the outcome is generated; stage 2 is the process by which the outcome is observed. One mechanism for differential bias is when the observability of such an event differs across a sensitive variable. For example, people who live further away from the hospital under study are less likely to readmit to the same hospital, that information will not be observed within the EHR. Therefore, within the data, people who live further away will have lower readmission rates. Suppose $X_s$ is the sensitive variable, $X_1$ is some other risk factor, $E$ is the true outcome and $Y$ is the observed outcome. Differential bias is defined as:

$$
\Delta \text{Bias} = \text{mean}[P(E = 1|X_s = 1) - \hat{P}(Y = 1|X_s = 1)] - \text{mean}[P(E = 1|X_s = 0) - \hat{P}(Y = 1|X_s = 0)]
$$

Figure 1: Diagram illustrating the differential observing process by sensitive variable $X_s$

We sought to investigate how the inclusion of a sensitive variable into a CPM could affect such bias. We performed a series of simulations (1000 iterations) based on the two-stage process:

$$\logit(P(E = 1)) = \beta_0 + \beta_1 X_1 + \beta_s X_s$$

$$\logit(P(Y|E = 1)) = \gamma_0 + \gamma_s X_s$$

Here, $X_s$ potentially affects the true risk through $\beta_s$, and potentially affects probability of observing the outcome through $\gamma_s$. There is no differential observability issues when $\gamma_s = 0$. After simulating the true data, we learned a CPM, where both included and did not include $X_s$ as a predictor. We then compared learned risk to the true risk.

Results

We first consider the two simple cases where the sensitive variable $X_s$ only impacts either of the two stages. In (a) and (b), sensitive variable only impacts observing the outcome (stage 2). Here including $X_s$ in the model does induce more differential bias (b). Conversely, in (c) and (d), there is no differential observability issues, that is, $X_s$ only impacts the true outcome (stage 1). Here ignoring it in the CPM (c) leads to larger differential bias than adding it in the model (d).

Although differential bias can be smaller when $X_s$ is excluded under certain observability assumptions, metrics based on observed data are always better when including $X_s$ in the model. Table 1 shows the value of two different commonly used metrics AUC and calibration-in-the-large in different scenarios. Note that cases with smaller differential bias
Figure 2: True versus predicted risk by sensitive variable. $X_s$ impacts stage 1/2, including/excluding $X_s$ in model

Table 1: $\Delta$AUC and $\Delta$Calibration under different scenarios with 95% confidence interval.

(bold) are not necessarily corresponding to smaller $\Delta$AUC and $\Delta$Calibration. This result suggests that metrics that are only based on observed data may not be able to detect the true differential bias induced by observability differences.

When $X_s$ has effects on both the true outcome and observing the outcome (Figure 3), there is no simple answer to when to include the variable in the model. The differential bias is determined by the direction and magnitude of the effects of that variable on both stages.

Figure 3: Heatmaps of $|\Delta_{in}Bias| - |\Delta_{ex}Bias|$ when $\beta_s$ and $\gamma_s$ vary, i.e. sensitive variable effects vary.

Discussion

Our study furthers the literature on algorithmic fairness. We propose a formal definition of differential bias which is based on statistical bias. We explore the question of whether one should include a sensitive variable into a CPM and show that the answer depends on whether the sensitive variable truly affects the outcome or just the observability of the outcome. However, we also show that answering this question may not be possible solely on observed data and may instead be an implicit assumption of CPMs.

References

Machine Learning Powered Identification of Cancer Diagnoses and Diagnosis Dates from EHR: an integrated pipeline that outperforms ICD-coded identification

Yunchen Yang, Ph.D. a, *, Timmy O’Connell, Ph.D. a, *, Arielle Redfern a, David Corrigan, Ph.D. a, Youngjun Kim, Ph.D. a, Abbi Kern a, Emily Reed a, Kristen Miller a, Lourdes Soler b, Rene Dempsey a, Jeremie Carlson a, Eric Schadt, Ph.D. a, b, Rong Chen, Ph.D. a, b, Xiang Zhou, Ph.D. a.

a Sema4, Stamford, Connecticut, USA; b Icahn School of Medicine at Mount Sinai, New York, New York, USA
* Contributed equally

Introduction

Cancer is the leading cause of death worldwide, with 10 million deaths recorded in 2020. Given the pathological complexity of cancer, research leveraging large cohorts of data has the potential to improve the therapeutic landscape. As a result, the accurate identification of cancer cases and diagnosis-related temporal information is crucial, which is an essential step of numerous down-streaming analyses. Electronic health records (EHR) capture vast amounts of clinical information of cancer patients, which therefore make them a great source to facilitate various aspects of cancer research, especially the identification of diagnosis (Dx) and diagnosis dates (DDx). However, EHR data is also sparse and noisy, where nuanced information is mostly embedded in the narrative text. Traditional chart review on EHR data is tedious and time-consuming, which cannot adapt to “big data” analysis. To overcome such challenges, many automated methods have been developed to identify cancer Dx and DDx from EHR data. Despite previous advances, challenges remain. Firstly, ICD coded Dx, which may contain up to 50% false positive cases, was widely used as a standard method of capturing Dx information. In addition, the extraction of DDx is still rarely addressed in previous studies. Herein, we developed a new pipeline which maps Dx-related concepts in EHR data and identifies diagnosed cases and associated DDx using machine learning (ML) models. This study encompassed 15 common cancer types and outperformed the traditional ICD code centered method.

Methods

As shown in Figure 1, for each cancer type, 500 patients were selected by random sampling from all the patients with ICD coded visits. A manual chart review was performed to annotate the patients with diagnosed cancer and associated DDx at a patient level. To ensure the quality of the work, the rule for annotating reference standard dataset was evaluated, defined, and refined by human experts. We then applied an open-sourced natural language processing (NLP) module, ConceptMapper, to capture cancer-related concepts and parse the sentences to populate the snippet-level disease-specific features. A gradient boosting classifier (GBC) was then trained to identify cancer Dx. In parallel, the concept-associated temporal information was extracted as features to train another GBC-based model to elect DDx of each identified patient. The ML models were trained by both the structured ICD coded information and the EHR embedded unstructured features. The precision/recall and overall accuracy of the validation set within 5-fold cross-validation were determined to evaluate the performance.

Results

To demonstrate the significance of EHR embedded unstructured information on Dx and DDx identification, a set of ML models were trained by structured features only and the performance scores were compared with those Freeman et al. (2020).}

Figure 1. The study design of the pipeline.
accomplished by the developed pipeline. As shown in Table S1 and Figure 2, the mean accuracies across all cancer types, of the models trained by structured features, are 66.44% on the validation set. With the additional unstructured features, the mean accuracies were boosted to 90.31% on the validation set, respectively. The precision/recall and overall accuracy were significantly improved across all cancer types after implementing the extraction and processing of unstructured information from EHR data.

**Discussion**

In this study, we developed a ML-powered abstraction pipeline for the identification of cancer diagnoses and DDx, which was employed across 15 common cancer types. The pipeline extracted unstructured information from the EHR data via NLP approaches and trained a GBC ML model to predict the Dx information. Compared with the ICD coded identification, the pipeline showed significantly improved overall accuracy (ICD% accuracy vs. ML% accuracy, Dx: 66.44% (±9.04%) vs. 90.31% (±4.68%), DDx: 64.73% (±11.80%) vs. 81.35% (±3.92%)). The method holds good the potential to be applied on other large dataset with more cancer types by simply optimizing current concept dictionaries and regular expression patterns to capture additional features. In the future, we will expand this methodology to additional cancer types and improve the algorithm by validating the performance on larger cohorts of patients.

**References**

Differential gene expression analysis stratified by APOE genotypes across brain regions in Alzheimer’s disease

Grace Yang1,2, Yaqiao Li, BA1, Stella Belonwu, PhD1, Alice Tang, BS1, Tomiko Oskotsky, MD1, Marina Sirota, PhD1

1UCSF Bakar Computational Health Sciences Institute, San Francisco, CA; 2Dougherty Valley High School, San Ramon, CA

Introduction
Alzheimer’s Disease (AD) is a progressive neurodegenerative disease that causes neuronal and brain function loss1. The E4 allele of the apolipoprotein E (APOE) gene is one of the most significant genetic risk factors for AD, with the probability of AD increasing with the number of E4 copies. However, the transcriptomic differences associated with the different APOE isoforms remain unclear. In this study, we completed a differential gene expression analysis stratified by 0, 1, or 2 copies of APOE4 with the neutral APOE3 subtype (APOE3/3, APOE3/4, APOE4/4), APOE3/3 referring to two APOE3 alleles, followed by pathway enrichment analysis for 7 brain regions to identify regional differences in APOE isoform-specific enriched genes/pathways. We also performed an expression quantitative trait loci (eQTL) analysis to identify genes with varying expression levels associated with the different APOE genotypes.

Methods
Bulk RNA-Seq data of 7 brain regions was obtained for patients with an APOE genotype of APOE3/3, APOE3/4, or APOE4/4 from 3 public datasets: MayoRNASeq dataset1 (temporal cortex (TCX) (78 AD & 65 controls) and cerebellum (CBE) (78 AD and 66 controls)), the Religious Orders Study and Memory and Aging Project1 (dorsolateral prefrontal cortex (DLPFC) (144 AD and 70 controls)), and the Mount Sinai Brain Bank dataset1 (anterior prefrontal cortex (Brodmann Area 10 (BA10) (59 AD and 19 controls), posterior superior temporal gyrus/Wernicke’s area (BA22) (52 AD and 12 controls), perirhinal cortex (BA36) (36 AD and 19 controls), and inferior frontal gyrus/Broca’s area (BA44) (47 AD and 10 controls)). The APOE isoform groups were annotated using similar criteria as Wan et al1, if not otherwise annotated by the dataset. The data was batch-corrected using the R package limma2 and outliers (PC1 or PC2 (principal component) Z-scores greater than 3) were removed at a sample level. Differential gene expression analysis was performed using the R package DESeq22, stratified by genotype with a Wald test (AD vs Control), and with the Benjamini Hochberg (BH) p-value correction method. Differentially expressed genes (DEGs) were selected with log2 fold change (LFC) > 0.4 and a false discovery rate (FDR) of < 0.05, with parameters chosen to see observable differences in the data. Using g:Profiler3, the DEG lists were input to generate pathways from the Gene Ontology databases. Significant pathways (p-value < 0.05) were transferred to Cytoscape’s visualization application, EnrichmentMap4. Related pathways were collapsed into biological theme clusters, which were defined by the AutoAnnotate application4. eQTL analysis was performed (ANOVA test and BH correction) to identify APOE4 dosage-dependent GE changes in AD for DLPFC, the only brain region with an APOE4/4 control.

Results
From the pathway enrichment analysis, two pathway clusters, movement motility locomotion and intracellular signal transduction, were upregulated in AD carrying APOE3/3 for 3 brain regions that make up the temporal lobe4 (TCX, BA22, and BA36) (Figure 1a) and downregulated in BA10 (responsible for working memory)5 (Figure 1b). Pathway clusters relating to the coagulation cascade (platelet activation coagulation and wound blood coagulation), collagen, and the extracellular matrix were upregulated in AD carrying APOE3/4 in the TCX (associated with recognizing language and memory acquisition4 and CBE (responsible for voluntary movement)4 (Figure 1d). The morphogenesis neuron projection cluster (neuron development), which was shared between the CBE and BA22 (involved in language comprehension)6, and the inflammatory response woundung cluster (DLPFC), were downregulated in AD carrying APOE 3/4 (Figure1c). Several ion transport clusters were downregulated in BA22 in AD carrying APOE3/4, while similar clusters were downregulated in BA36 in AD carrying APOE3/3. Due to the lack of APOE4/4 control samples in 6 of the regions, and no DEGs in the DLPFC in APOE4/4, we were unable to carry out APOE4/4 pathway analysis. From the eQTL analysis, we found 8 DEGs in AD patients with different APOE isoforms in DLPFC (responsible for working memory)6 compared to controls. Several of the DEGs have known functions that relate to APOE, such as NR5A2 (Figure 2a), which codes for a transcriptional regulator of lipid metabolism, LYZ (Figure 2b), which codes for the lysozyme2, and ZIC1 (Figure 2c), which codes for a transcription factor that activates the APOE gene1.

Discussion
Through pathway network and eQTL analysis, we found regional differences within the 7 brain regions based on APOE genotype, as well as genes that exhibited a significant difference in expression with different APOE4 dosages. AD patients with the APOE3/3 genotype had significantly different expression than controls in regions CBE, TCX,
understanding AD pathogenesis on a transcriptomic level, which can benefit future treatment research. Research is needed to provide more insight into the role that the genes identified play in AD, this study adds to the understanding AD pathogenesis on a transcriptomic level, which can benefit future treatment research.

References


Generalizing an Antibiotic Recommendation Algorithm for Treatment of Urinary Tract Infections to a New York City Hospital System

Garrett Yoon, MS1, Vincent J. Major, PhD1
1NYU Grossman School of Medicine, New York, NY, USA

Abstract

Artificial intelligence systems struggle to generalize to a population external to their creation. A promising system designed to automate antibiotic treatment for urinary tract infection was applied to an external dataset of three thousand records from the New York City area. Model discrimination decreased and, using the published resistance thresholds, the balance of errors shifted, cascading through the system and reducing the potential benefit to augment clinical decision-making to reduce excessive antibiotic treatment.

Introduction

Inappropriate antibiotic prescription exacerbates antibiotic resistance in hospitals and communities. Common low-risk infections such as urinary tract infection (UTI) pose a crucial challenge: recurrence is common but antibiotics prescribed days before confirming susceptibilities. Clinical guidelines and empirical results can assist decision-making but are often trumped by perceptions that an individual’s benefit outweighs community risk. Antibiotic stewardship—encouraging evidence-based prescribing to stem overuse—could be advanced by clinical decision aids.

A machine learning system to recommend treatment for uncomplicated UTI was developed1 and the code and dataset (AMR-UTI) released2. Four antibiotics were considered, in increasing breadth: nitrofurantoin (NIT), trimethoprim-sulfamethoxazole (SXT), ciprofloxacin (CIP), and levofloxacin (LVX).

Risk models can struggle generalizing to different time periods, geographical locations, and data collection processes3. Superficially minute data differences can affect the model’s output, skew distributions, and shift outcome metrics. External validation of predictive models—testing on data beyond the development setting—is crucial to assess performance and clinical utility in practice.

Methods

Uncomplicated UTI specimens collected at NYU Langone Health between 2012–2020 were assembled with all 790 features of AMR-UTI to create the NYU-UTI dataset. Using the authors’ released codea, predictions of antibiotic resistance and treatment recommendations were replicated from AMR-UTI. Replicated models were applied to NYU-UTI to calculate discrimination (area under receiver operating characteristic; AUROC) and population-level metrics of inappropriate treatment (IAT) and second-line usage (SLU) defined in the original study1.

Results

In the nine-year period, 3,021 specimens met the inclusion criteria to form NYU-UTI. Between 12–21% of specimens were resistant to each antibiotic and physicians prescribed 49% NIT, 20% SXT, 28% CIP, and 3% LVX (31.6% SLU, 12.4% IAT).

Application of the models to NYU-UTI resulted in generally lower discrimination (Table 1). Performance is higher within the subset of patients with prior exposure or resistance. If all recommendations were followed, inappropriate antibiotic and second-line rates would be higher for NYU-UTI than AMR-UTI (Table 2).

---

a https://github.com/clinicalml/amr-uti-utm

643
Table 1. Discrimination of each antibiotic agent when replicated and generalizing to NYU data.

<table>
<thead>
<tr>
<th>Drug</th>
<th>AUROC (95% CI)</th>
<th>Published AMR-UTI Full Test Cohort</th>
<th>Replicated AMR-UTI Full Test Cohort</th>
<th>NYU-UTI Full Test Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Prior antibiotic resistance or exposure</td>
<td>Prior antibiotic resistance or exposure</td>
<td>Prior antibiotic resistance or exposure</td>
</tr>
<tr>
<td>Nitrofurantoin (NIT)</td>
<td>0.56 (0.53-0.59)</td>
<td>0.61 (0.55-0.66)</td>
<td>0.60 (0.53-0.56)</td>
<td>0.54 (0.51-0.58)</td>
</tr>
<tr>
<td>TMP-SMX (SXT)</td>
<td>0.59 (0.57-0.62)</td>
<td>0.67 (0.64-0.71)</td>
<td>0.60 (0.57-0.62)</td>
<td>0.68 (0.64-0.72)</td>
</tr>
<tr>
<td>Ciprofloxacin (CIP)</td>
<td>0.64 (0.60-0.68)</td>
<td>0.76 (0.71-0.80)</td>
<td>0.64 (0.60-0.67)</td>
<td>0.77 (0.71-0.83)</td>
</tr>
<tr>
<td>Levofloxacin (LVX)</td>
<td>0.64 (0.60-0.68)</td>
<td>0.77 (0.71-0.82)</td>
<td>0.63 (0.59-0.67)</td>
<td>0.77 (0.71-0.83)</td>
</tr>
</tbody>
</table>

Table 2. Primary outcome metrics if all recommendations are followed.

<table>
<thead>
<tr>
<th></th>
<th>Published AMR-UTI</th>
<th>Replicated AMR-UTI</th>
<th>NYU-UTI</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of Patients Recommended (n)</td>
<td>99.2% (3911/3941)</td>
<td>95.9% (3781/3941)</td>
<td>99.0% (2992/3021)</td>
</tr>
<tr>
<td>Inappropriate Antibiotic Therapy (95% CI)</td>
<td>9.8 (8.9-10.8)</td>
<td>10.0 (9.1-10.9)</td>
<td>12.6 (11.4-13.8)</td>
</tr>
<tr>
<td>Use of Second-Line Therapy (95% CI)</td>
<td>11.0 (10.0-12.0)</td>
<td>11.8 (10.8-12.8)</td>
<td>14.7 (13.5-16.0)</td>
</tr>
</tbody>
</table>

Discussion

The models trained on Boston data generalize moderately well to New York City but distributional differences weaken discrimination and key metrics of antibiotic stewardship. Despite the weaker performance, the system is more empiric than physician prescribers with a potential 54% reduction in second-line therapy (compared to 70% less in AMR-UTI) with no notable increase in IAT. The published system selects resistance cutoffs based on false negative rates within cross-validation to achieve IAT ≤ 10%. These cutoffs were used as-is but failed to reduce IAT on NYU-UTI. To reduce IAT, part of the dataset could be used to re-threshold.

Strict inclusion criteria included confirmed susceptibility for each antibiotic, exactly one being prescribed, and a narrow age range of 18–55 years. Together, these criteria shrink the sample size and may introduce bias from lab ordering and prescribing patterns.

Conclusion

Data from three thousand UTI specimens were compiled to externally validate a machine learning system to recommend antibiotic treatment. Discrimination decreased for each model, impacting the balance of inappropriate and second-line treatments. Recommendations may still help curb excessive antibiotic treatment.

References

Systematic comparison of phenomic profiles between All of Us Research Program and UK Biobank participants

Chenjie Zeng¹, Lisa A. Bastarache², David J. Schlueter¹, Jacob Keaton¹, Joshua C. Denny¹
¹NHGRI/NIH, Bethesda, MD ²Vanderbilt University Medical Center, Nashville, TN

Introduction

Large scale electronic medical records (EHR) linked biobanks have been shown useful in accelerating biomedical research. Owing to different underlying populations and distinct strategies in the recruitment and data collection, even “population scale” cohorts can vary significantly in their represented phenotypes and affect their suitability for certain types of analyses. Here, we compared the phenomic profiles of participants from two leading such resources and identified factors likely contributing to the differences. We showed that association results from both studies could be comparable after properly controlling for these factors.

Methods

The All of Us Research Program (AoU) is an ongoing longitudinal cohort study launched in the U.S. in 2018 aiming to enroll 1 million diverse participants and collect a wide range of data including EHRs. As of August 20, 2021, 203,813 had visit data available in their EHRs (1). We restricted analyses to participants with established EHR medical homes, with at least three outpatient visits in 5 years documented in the EHRs. A total of 172,529 participants were included. The average age at recruitment was 52.8 years, with 63% being female and 47% being non-white.

The UK Biobank (UKB) is a prospective cohort study of 500,217 participants aged 40-69 years at recruitment with 54% being female and 94% being white, who were recruited between 2006-2010 and are continuously followed (2). The average age at recruitment was 56.5 years. For phenotyping of diseases, we included data fields of in-hospital diagnoses, cancer registry and death registry as well as self-reported health outcomes mapped to ICD codes. UKB does not yet have outpatient primary care diagnoses for all participants integrated.

We extracted ICD codes and mapped them to phecodes that aggregate them into groups as previously described (3, 4). Phenotypes were coded with 1 if participants had the phecode at least once (cases) and otherwise with a 0 (controls). For sex-specific phenotypes, participants were limited to the specific sex. The prevalence ratio (PR) for each phenotype was calculated where the proportion of cases in AoU was divided by the proportion of cases in UKB. The score test was used to estimate the statistical significance. We defined phenotypes with a log score test statistics > 4 (corresponding to a P value <1×10⁻¹²) as phenotypes of significantly different prevalence. Lasso logistic regression was used to identify predictors of phenotypes showing significant differences between these two cohorts. As a proof of concept, we conducted a phenome-wide association study (PheWAS) for diabetes in both studies, controlling for the identified predictors when available in addition to age, sex, race, and sites (3).

Results

A total of 2,574 shared phenotypes were identified. The most common phenotype in both cohorts was hypertension (42% and 29% in AoU and UKB, respectively). The PRs between AoU and UKB ranged from 0.02 to 12871(Figure 1). A total of 1,940 phenotypes showed a significant difference in the prevalence, 1,863 of which showed a significant higher prevalence in AoU. These phenotypes spanned all categories, most seen in cardiovascular and endocrine groups, such as diabetes (PR = 2.4, P < 1×10⁻²⁷⁶). On the other hand, 77 phenotypes showed a lower prevalence, most seen in the neoplasms, such as colorectal cancer (PR= 0.65, P = 1.1×10⁻⁷¹). Predictors for a phenotype with a significant difference included a non-neoplastic category, relating to non-inpatient visits, Mendelian diseases, sex-specific phenotypes, a code with subordinate codes, a high prevalence in AoU, a low prevalence in UKB, average age at enrollment and average length of years in EHR in cases of AoU. In the PheWAS analysis, we adjusted for EHR lengths, age at enrollment and racial groups. The correlation of the effect size for phenotypes (r) was 0.6 and 0.3, respectively, for diabetes and cystic fibrosis, suggesting that both studies can be
comparable in the association analyses for common diseases but might not be for rare disorders. However, association results of some of the known clinical correlates were comparable for both diseases (Table 1).

Conclusion
We found substantial heterogeneicities in phenomic profiles of participants in the AoU and UKB cohorts, with most phenotypes found with higher prevalence within AoU. These heterogeneicities may be due to differences in the underlying populations and approaches in data collection, such as current inclusion of both inpatient and outpatient diagnoses within AoU. PheWAS results for diabetes, however, were similar, suggesting that the impacts of the underlying heterogeneity on association analyses in common diseases could be alleviated through appropriate adjustment of these factors. Additional data are needed for rare diseases. Nonetheless, a more refined approach that include multiple resources of data from EHRs is needed for accurate phenotyping. We anticipate that the ongoing efforts that collect the general practice data in UKB will be helpful for this purpose.

![Figure 1](image-url) Comparing prevalence of phenotypes in AoU and UKB. The y-axis represents the absolute value of log prevalence ratio (PR). The x-axis includes all significantly different phenotypes. The upward pointing triangles represent phenotypes with a higher prevalence in AoU while the downward represent those with a lower prevalence.

### Table 1 Association results of known clinical correlates with diabetes and cystic fibrosis in AoU and UKB.

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Diabetes</th>
<th>Cystic fibrosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AoU (OR 95% CI)</td>
<td>UKB (OR 95% CI)</td>
</tr>
<tr>
<td>Chronic kidney diseases</td>
<td>5.2 (5.0-5.3)</td>
<td>4.8 (4.7-5.0)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>8.6 (9.3-8.8)</td>
<td>7.4 (7.3-7.6)</td>
</tr>
<tr>
<td>Obesity</td>
<td>5.2 (5.1-5.4)</td>
<td>5.4 (5.3-5.6)</td>
</tr>
<tr>
<td>Hyperlipidemia</td>
<td>6.7 (6.5-6.9)</td>
<td>4.6 (4.5-4.7)</td>
</tr>
</tbody>
</table>

### References
Creating Classifiers to Distinguish Between Speaker Types in Recorded Nurse-patient Verbal Communication in Home Healthcare: a Feasibility Study

Maryam Zolnoori,1 Sasha Vergez,2 Ali Zolnour,3 Zoran Kostic,1 Maxim Topaz1
1Columbia University, 2Visiting Nurse Service of New York, 3University of Tehran

INTRODUCTION

Patient’s verbal communication provides a window to a wide range of pathological entities, including speech and language features that can help diagnose pulmonary hypertension,1 neurological disorders and mental illness;2 this enables the patient’s spoken language to act as a biomarker for screening patients with these illnesses. Recently, emergent studies have started to record verbal communication between clinicians and patients to develop diagnostic and risk identification algorithms for timely detection of patients with these illnesses. One initial challenge for this line of research is distinguishing between speaker identification. To start taking advantage of signals that can be uncovered from a patient's verbal communication in clinical settings, we conducted a feasibility study at the Visiting Nurse Service of New York (VNSNY). VNSNY provides home healthcare services (HHC) to vulnerable older adults at risk of negative outcomes (e.g., hospitalizations). In this feasibility study, we constructed an analytic pipeline using natural language processing (NLP), speech analysis techniques, and machine learning algorithms to extract linguistic and acoustic features from patient verbal interaction with HHC nurses. One of the first challenges for us was to create some type of an automated recognizer of the speaker type, specifically was the speaker patient or a nurse? The focus of this study is on the development of a critical component of this pipeline for distinguishing patient spoken language from the nurse spoken language using the transcribed audio recorded verbal patient-nurse communication. We investigated the effectiveness of different machine learning classifiers to distinguish the patient language from the nurse language.

METHODS

Audio-recording patient-nurse verbal communication: Patient-nurse verbal communication was recorded using Saramonic Blink500 Pro. This device includes two dual-channel stereo wireless microphones that allow recording in MP4 format. To record encounters using this device, we attached one of the wireless microphones to the patient’s clothes and another one to the nurse’ clothes. We received an IRB approval from the study site.

Sample size of the study: We audio recorded 10 patient-nurse encounters using Saramonic Blink500 Pro.

Transcribing the audio-recorded encounters: The machine learning classifier for differentiating patient spoken language from the nurse spoken language was developed on the transcriptions of the audio recorded encounters. All recorded encounters were transcribed using Amazon Web Services (AWS) general Transcribe. In our preliminary study, we showed that AWS-GT has a higher transcription quality (measured using Word Error Rate[WER])3 compared with AWS-Medical Transcribe and Wav2Vec4 and Kaldi5, two open source automatic speech recognition systems.

Evaluating the quality of speaker identification of AWS - General Transcribe (GT) links each transcribed word to a speaker Id, which is a numerical value. For example, if two persons are involved in the conversation, each word will be linked to spk_1 or spk_0. To measure the quality of speaker identification, a member of our research team, selected three recorded encounters and manually linked each spoken word to the accurate speaker, and compared them with the AWS-GT speaker identification. The average accuracy of AWS-GT was 0.85%.

Generating patient and nurse utterances: Machine learning classifiers were trained on patient and nurse utterances. We defined an utterance as a continuous block of speech (patient or nurse) without interruption. To generate utterances, we used the speaker identification provided by AWS-GT in the transcription file. All transcribed words assigned to one speaker (e.g., spk_0) were grouped together as an utterance before the AWS-GT system assigned a transcribed word to the second speaker (e.g., spk_1). Using this approach, we generated all the utterances spoken by spk_0 and spk_1. Next, for each transcribed, we assigned the identified speaker to the patient or nurse.

Machine learning (ML) models: As discriminative ML models, we used support vector machine (SVM) and XG-boost. SVM was selected due to its high performance for limited sample size with high dementia space. XG-boost was selected due to its scalability and its optimized implementation of gradient boosting machine for learning the nonlinear relationship between input features and classes. We used the implementation of the models in the Scikit-learn library. As a generative model, we used Bi-LSTM neural network. Bi-LSTM is a sequence processing model that consists of two LSTM. One LSTM is responsible for taking the input features in a forward direction, and the other in a backward direction. BiLSTM effectively improves the context available to the algorithm.

Generating features from text: We used three approaches for generating features from transcribed verbal communication. i) Term Frequency-Inverse document Frequency (TF-IDF), ii) Part of speech (POS) tagging, and iii) word embedding methods. TF-IDF quantifies words in the document by computing the weight of each word, signifying the importance of the word. We used the
implementation of TF-IDF in the Scikit-learn library. For POS tagging, we used the TextBlob toolkit of python. Using this toolkit, the words in the document (transcribed communications) were tokenized and a POS tag was assigned to each tokenized word. Next, count of nouns, verbs, adjectives, adverbs, and pronouns were computed, and imported to the model as a list of features. For the word embedding method, we used the word2vec model and its implementation in the Gensim library of python. Performance of machine learning algorithms were evaluated using the cross validation method (CV = 5).

RESULTS
Overall, the data included 618 patient utterances and 623 nurse utterances. Table 1 provides F1-score, AUC and accuracy for SVM, XG-boost, and Bi-LSTM with different feature selection methods. The features generated using Word-2-vec were only used as input features for Bi-LSTM because it provided the best performance for both SVM and XG-boost. SVM with word2vec had a highest F1-Score (75.89) compared to other ML algorithms. Also, SVM with TF-IDF had the highest Area Under Curve (AUC). Our findings suggest that SVM outperformed Bi-LSTM for this classification task. This is mostly due to the study's small sample size with high dimensional features. One of the remarkable properties of SVM is its ability to learn independent of feature space dimensionality. SVM measures the complexity of the training dataset based on the margin with which it separates the data, not the number of features. This means that the trained model might have good generalizability even in the presence of many features. Therefore, the SVM is a particularly suitable model for data classification with small sample size and high dimensional features such as speaker type identification in audio-recorded patient-clinician conversations.

<table>
<thead>
<tr>
<th>Model</th>
<th>Input feature</th>
<th>F1-Score</th>
<th>AUC</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>TF-IDF</td>
<td>75.8</td>
<td>83.06</td>
<td>76.46</td>
</tr>
<tr>
<td></td>
<td>POS</td>
<td>63.56</td>
<td>54.25</td>
<td>54.73</td>
</tr>
<tr>
<td></td>
<td>TF-IDF + POS</td>
<td>72.74</td>
<td>81.23</td>
<td>74.43</td>
</tr>
<tr>
<td></td>
<td>Word2Vec</td>
<td>75.89</td>
<td>82.04</td>
<td>76.46</td>
</tr>
<tr>
<td></td>
<td>Word2Vec + POS</td>
<td>74.98</td>
<td>82.02</td>
<td>75.67</td>
</tr>
<tr>
<td>XG-Boost</td>
<td>TF-IDF</td>
<td>69.72</td>
<td>77.22</td>
<td>70.04</td>
</tr>
<tr>
<td></td>
<td>POS</td>
<td>57.66</td>
<td>59.1</td>
<td>57.09</td>
</tr>
<tr>
<td></td>
<td>TF-IDF + POS</td>
<td>72.23</td>
<td>79.5</td>
<td>72.51</td>
</tr>
<tr>
<td></td>
<td>Word2Vec</td>
<td>73.53</td>
<td>80.02</td>
<td>73.42</td>
</tr>
<tr>
<td></td>
<td>Word2Vec + POS</td>
<td>72.02</td>
<td>79.19</td>
<td>71.95</td>
</tr>
<tr>
<td>Bi-LSTM+NN</td>
<td>Word2Vec</td>
<td>72.43</td>
<td>79.48</td>
<td>72.1</td>
</tr>
</tbody>
</table>

Discussion
This study has an important implication for future studies aimed at audio-recording patient-physician verbal communications in clinical settings and we inform development of an automatic pipeline for analyzing patients’ spoken language and verbal interaction. Speaker type identification is currently a major challenge of automatic speech recognition (ASR) systems. Although existing commercial ASR systems such as AWS-Transcribe have acceptable performance in speaker identification for high quality audio-recorded verbal communication in clinical setting, further steps are needed to distinguish the patient’s spoken language from clinicians or caregivers involved in communications. In this study, we showed that ML algorithms have strong potential to distinguish patient language from the nurse language. One major limitation of this study is the small sample size. Currently, we are actively audio-recording patient-nurse verbal communication at VNSNY. As our feature plan, we aimed at improving the quality of the ML models using larger sample size and including other features, such as acoustic features or medical concepts (e.g., symptoms).

References
Developing a risk identification algorithm for identifying patients at risk of medication non-adherence in patients with depression

Ali Zolnour, Christina Eldredge, Maxim Topaz, Anthony Faiola, Paul Fontelo, Maryam Zolnoori
1University of Tehran, 2University of South Florida, 3Columbia University, 4University of Kentucky, 5National Library of Medicine

INTRODUCTION
The prevalence of antidepressant use among adults in the United States population increased from 7.7% in 1999–2002 to 13.2% in 2015-2018 time periods,1 with an increase in the global market cost estimated at $15.8 million in 2023. The therapeutic benefits of antidepressants depend on adherence to prescribed drug regimens, yet between 30 to 68% of patients discontinued their medication without completing an adequate course of treatment. Antidepressants non-compliance can negatively impact health outcomes leading to depression relapse, withdrawal symptoms, emergency department visits, and reduced quality of life.2 Furthermore, non-adherence can increase burden on the healthcare system, caregivers, and patient families. Although self-report screening tools such as Adherence Barriers and Questionnaire (ABQ) for identifying patients at risk of drug discontinuation are well validated, they are associated with some methodological limitations such as detecting significant factors, inadequate sensitivity, and sampling bias. The premise of this study is that patient self-reporting of their experiences with antidepressant therapy on drug review forums may constitute a reliable source to uncover various dimensions which lead to antidepressant discontinuation that may not be identified by the self-report screening tools. The identified factors can be used for development of sensitive screening tools and algorithms to identify patients at risk of abrupt drug discontinuation. In summary, the objective of this study is to utilize patient self-reported experiences in online healthcare forums to develop a risk-identification algorithm to identify those patients with depression at risk of drug discontinuation.

METHOD
In the first step, we used a novel mixed method approach to annotate the patients self-report experiences for antidepressant related information and drug discontinuation. In the second step, we used the annotated datasets to develop a risk identification algorithm to identify patients at risk of drug discontinuation.

Identifying data source and drug source: The data for this study were collected from a healthcare forum called “askapatient.com” that collects information on patients’ experiences with medication along with age, gender, duration of usage, and patients rating for medications ranged in a Likert scale from 1 (strongly dissatisfied) to 5 (strongly satisfied). A total of 892 patient antidepressants reviews were randomly collected from the forum for the four most commonly prescribed antidepressants, including Sertraline (Zoloft) and Escitalopram (Lexapro) from the SSRI class, and Venlafaxine (Effexor) and duloxetine (Cymbalta) from the SNRI class.

Developing the Analytical Framework: We used the Framework Method3 with deductive-inductive approach to identify themes affecting patients’ discontinuation. In the deductive approach, we identified factors affecting patients’ antidepressants discontinuation toward antidepressants using a thorough literature review. The factors were used as the initial themes for constructing the initial analytical framework for data analysis. In the inductive approach (open coding), 110 drug reviews were randomly selected for initial analysis using the initial analytical framework. Passages of drug reviews that could not be covered by the initial analytical framework were discussed in our regular team meetings for generating new themes. In the final step, themes were refined and those that covered less than 5% of the drug reviews were eliminated or merged with other themes. The identified themes were used for summarizing patients’ experiences with antidepressants and developing highly structured data from narrative patient reports. The final set of themes for annotation of the dataset of this study included five entities, 1. “presence of Adverse Drug Reaction (ADR),” 2. “perceived distress from ADR (ADR-PD),” 3. “Medication Effectiveness (EF),” 4. “the patient’s complaint about the lack of knowledge for medication (KN),” and 5. “patient-physician interaction.” The details of the methodology for developing this dataset were explained in our previous study which focused on identifying factors affecting patient attitudes towards antidepressants.3

Annotating the dataset using the analytical framework: In the first step of this phase, we addressed the grammatical and punctuation errors in the patient comments, and then split the comments into sentences using NLTK (a Python module). Overall, 500 patients post (out of 892 posts) including 4,000 sentences were annotated using the analytical framework at sentence level. The remaining posts, including 400 posts (out of 892 posts), were annotated at the post level for explicit expression of drug discontinuation. Four annotators with health backgrounds participated in the process of data annotation. All 4,000 sentences and 400 posts were double-coded. Disagreement between annotators was resolved by discussion between annotators.

Calculating the Inter-annotator agreement (IAA): The Inter-annotator agreement (IAA) was calculated using Cohen Kappa. The overall IAA for the entire dataset was 0.75 with highest IAA (0.88) for “presence of ADR” and lowest IAA (0.5) for “patient-physician interaction.” The “patient-physician interaction.” was removed from the dataset due to low agreement between annotators. The final annotated data included two datasets, dataset (A) that included annotated sentences for four variables, ADR, ADR-PD, EF, and KN, and dataset (B) that included posted sentences for “expression of drug discontinuation.”

Developing machine learning (ML) algorithms: Development of risk identification algorithm for identifying patients at risk of drug discontinuation composed of three main phases (see figure 1).

Phase 1: identifying sentences with expression of ADR, ADR-perceived distress, effectiveness, and lack of knowledge (KN): We used dataset (A) to build ML classifiers for identifying sentences with expression of ADR, ADR-PD, EF, and KN. To balance labeled sentences for each of these entities, we used the back translation technique4, which generates new sentences close in meaning to original sentences. Term Frequency-Inverse Document Frequency (TF-IDF) was used for feature extraction from the sentences. Several ML algorithms including support vector machine (SVM), Random Forest (RF), XGBoost, and General Linear Model (GLM) were trained and tested on the 4,000 sentences of dataset (A) to build ML algorithms for these sentences. Performance of the ML algorithms measured using cross validation (CV = 5). SVM had the highest precision, recall, and F-score for all entities (Table 1).
Phase 2: identifying the patient's attitude toward antidepressants. Dataset (A) includes patients' rating for antidepressants. Patients rated their attitudes towards the medications using a 5-point Likert Scale from 1 (strongly negative attitudes) to 5 (strongly positive attitude). We summarized the patients' attitudes in three classes -- negative class (rating 1 and 2), neutral class (rating 3), and positive class (rating 4 and 5). To balance the data for these three classes, we used a SMOT resampling method. We used the same ML algorithms and cross validation method in phase 1 to predict patients' attitudes toward antidepressants.

Phase 3: developing risk identification algorithms for identifying patients at risk of drug discontinuation: Dataset (B) was used for developing the risk identification algorithm for predicting drug discontinuation. To create a set of features, we used ML algorithms developed in Phase 1 and Phase 2 of identifying patients’ expression for ADR, ADR-perceived distress, Effectiveness, lack of knowledge about medication (KN), and attitudes towards medications. We also used word2vec, a word embedding method for creating vector features from a dataset to generate feature vectors from the text. Additionally, these antidepressants related features and the word embedding features were combined with patients’ demographic information (age and gender) and duration of antidepressant usage reported by the patients in the forum.

RESULT

Table 1. Precision, Recall, and F-score, and AUC of ML models of Phase 1 and Phase 2

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF</td>
<td>92.40</td>
<td>92.41</td>
<td>92.39</td>
<td>92.41</td>
</tr>
<tr>
<td>ADR</td>
<td>86.72</td>
<td>86.12</td>
<td>86.41</td>
<td>86.46</td>
</tr>
<tr>
<td>ADR-A</td>
<td>89.80</td>
<td>89.82</td>
<td>89.75</td>
<td>89.82</td>
</tr>
<tr>
<td>KN</td>
<td>85.70</td>
<td>85.10</td>
<td>85.31</td>
<td>85.36</td>
</tr>
<tr>
<td>Attitude</td>
<td>80.39</td>
<td>80.28</td>
<td>80.27</td>
<td>80.28</td>
</tr>
</tbody>
</table>

DISCUSSION

This work is a major contribution to identification of patients at risk of drug discontinuation which has significant implications for developing clinical interventions aiming at improving patient adherence towards medications. The algorithm adequately identifies patients who reported antidepressants ADR, determines perceived distress associated with the ADR (low, medium, high), identifies patients who complain about lack of knowledge for the ADRs, and identifies patient perception of antidepressant effectiveness. Clinicians could encourage patients to record their actual experiences with antidepressants and their impacts on daily functioning. This information may help clinicians tailor interventions to improve patient perception about drug effectiveness and the potential ADRs. Moreover, because the full effects of antidepressants are typically not seen for four to six weeks, clinicians should track patient self-reported symptoms and adjust treatment accordingly to encourage completion of an adequate trial of the medication. Several studies have shown that physician support can significantly reduce the incidence of medication non-compliance.

References

Modernizing Data Pipelines and Visualizations of Injury-related Emergency Department Visit Data – Lessons Learned

Kamran Ahmed, MD MS, Nimesh Patel, MA, Keming Yuan, MS, Alen Alic, MS, Nimi Idaikkadar, MPH, Royal Law, PhD MPH, Michael Coletta MPH, Asad Islam, MS

Centers for Disease Control and Prevention, National Center for Injury Prevention and Control; Atlanta, Georgia, USA

Introduction

Injuries and violence are the leading causes of death in the United States for children, adolescents, and adults ages 18-44 years and rank in the top 10 causes of death for people 45 years or older. The Centers for Disease Control and Prevention’s National Center for Injury Prevention and Control (NCIPC) collaborates with the Consumer Product Safety Commission on a surveillance system called the NEISS All Injury Program (NEISS-AIP), which collects data from 66 hospitals to track injury-related emergency department visits. While NCIPC receives case data weekly, data are processed and visualized annually and the lack of timely information on injury trends may delay situational awareness and implementation of public health interventions. We aim to improve the timeliness of using NEISS-AIP data by integrating emerging technologies to more quickly and accurately understand injury trends. This work is in line with the CDC’s strategy to modernize data and technology, where data drives action in real time, efficiently, flexibly, rapidly, and with impact.

Approach

We assessed weekly and annual NEISS-AIP datasets to understand workflows, identify issues, and understand data wrangling needs. We also conducted focus group discussions, interviews with key stakeholders, and documentation review and analysis to validate project goals and elicit dashboard visualization requirements. We then applied natural language processing (NLP) by training word embedding models on the case narrative to capture key injury indicators more accurately. Finally, we used CDC’s cloud-based enterprise data analytics, and visualization (EDAV) platform for data analysis and building the visualizations. We used the EDAV’s Tableau Server for creating the data visualization dashboard and HTML and Java Script approach for our front-end web-based user interface.

Discussion

The requirement gathering process with early engagement of key stakeholders was critical to this project’s success and involved effective management, prioritization and validation of requirements with key stakeholders and users during the early dashboard design phase. To modernize data pipelines using the latest technology, tools, and techniques, we applied modern data automation approaches using CDC’s EDAV cloud-based platform which significantly improved timeliness of data, reduced data management burden, and allowed for better coordination of data activities and systems. Since all data visualization needs as defined during the requirement elicitation process could not be met with one analytics tool, we used a hybrid data visualization approach using different cloud-based analytics tools (Tableau and Azure Synapse Analytics) and embedded charts in a web-based dashboard. It is well known that manual cleaning and unifying messy and complex data sets for easy access and analysis is time intensive, increases workload, and poses significant risk of human error in data compilation. We used cloud-based tools to automate this process that drastically decreased dependency on manual methods and improved workflow efficiency. Another challenging task during the project was dealing with the clinical free-text data due to misspellings. We successfully applied word embedding models for this NLP task together with regular expression in an effective way to extract insights from medical narrative data that will enable faster decision-making.

Conclusion

The lessons from this study offers an opportunity to understand some of the most common challenges experienced during the implementation of data modernization initiatives. These lessons may contribute to developing guidance to support public health jurisdictions on developing similar data modernization plans, where data drives action more efficiently and flexibly.
A Rapid Detection Tool for Cognitive Impairment in Everyday Clinical Settings

Saki Amagai, BA1, Cindy Nowinski, MD, PhD1, Zahra Hosseinian, MA1, Laura Curtis, MS1, Julia Yoshino Benavente, MPH1, Marina Arvanitis, MD1, Mike Wolf, PhD1, Richard Gershon, PhD1. Northwestern University, Chicago, Illinois

Abstract

Early detection of Cognitive Impairment (CI) is imperative to identify potentially treatable underlying causal conditions. To address this, we created a brief, standard set of self-administered CI screening measures applicable for use in diverse settings and populations. It is comprised of two cognitive measures adapted from the NIH Toolbox for Assessment of Neurological and Behavioral Function® implemented as a downloadable app integrated with an EHR system.

Introduction

Cognitive impairment (CI) and dementia are significant public health burdens that can have profound social and emotional effects on older adults. Early detection of CI is imperative to identify potentially treatable underlying causal conditions, and in cases where cure is not possible, to provide supportive services to minimize the effects of CI. While primary care settings seem ideal for identifying CI, almost 90% of CI may be missed.1 Screening is unstandardized, and available tools may be challenging for clinical implementation because of their length, cost, or need for specialized equipment or highly trained administrators. We created a 7-minute self-administered CI screening tool, MyCog, applicable for use in diverse clinical settings, using an iPad app integrated with EPIC.

Methods

MyCog includes two cognitive tests adapted from the NIH Toolbox for Assessment of Neurological and Behavioral Function®: Picture Sequence Memory (PSM) and Dimensional Change Card Sorting (DCCS).2 PSM assesses episodic memory while DCCS measures cognitive flexibility. Both tests have demonstrated sensitivity to CI. Based on end-user feedback, we shortened PSM and created self-administered versions of the tests. Data from a validity study (N=100 adults ages 60+) is being analyzed and will be presented. Clinical validation is in process.

Conclusion

MyCog integration with our electronic health record (EHR) system, Epic, will enable clinicians to easily order the tests and access and interpret patients’ scores. In addition, the cognitive screening results are linked to a standardized clinical protocol with care management recommendations (e.g., additional evaluation to rule out reversible causes of CI, referrals, developing a care plan) thereby enhancing MyCog’s clinical utility. See Figure 1 for paradigm.

Figure 1. Cognitive Impairment (CI) Screening Paradigm.

References

Modeling Optimal Air Ambulance Base Locations: A MEXCLP Approach

Douglas R. Amirault\textsuperscript{2, 3}, Sabin Ahmed, Ph. D.\textsuperscript{1}, Robert D. Lieberthal, Ph. D.\textsuperscript{2}, Susan Haas, M.D., M.Sc.\textsuperscript{1}, David M. Hechtman, B.A.\textsuperscript{1}, Lauren Rayson, M.S.\textsuperscript{1}

\textsuperscript{1,2}The MITRE Corporation, \textsuperscript{1}McLean, VA, \textsuperscript{2}Bedford, MA \textsuperscript{3}; Bates College, Lewiston, ME

Abstract

Air ambulances deliver a critical healthcare service to patients, especially those located far from a hospital. However, allocation decisions may be inconsistent. We describe the application of Maximum Expected Covering Location Problem (MEXCLP) to the air ambulance base location problem to New England. We find that the MEXCLP base locations improve the covered population by 6.5-16.4 percent depending on the busy fraction.

Model and Data

The demand data inputted into the model come from CMS and are the number of Medicare hospitalizations from each zip code in 2016. The supply data for potential base locations comes from the Dartmouth Atlas of Health Care and are the centroids of hospital service areas. Allagash, a Python spatial optimization library, currently supports the Location Set Covering Problem and Maximum Covering Location Problem models. We use Allagash to implement the Maximum Expected Covering Location Problem (MEXCLP)\textsuperscript{1} model, introducing a busy fraction $p$ corresponding to the percentage of time an air ambulance is dispatched. We test our implementation using an example from the original MEXCLP paper and applied this model to the use case of maximizing coverage in New England.

On the maps below, black circles represent existing air ambulance locations, blue circles represent optimal locations, and red circles represent demand. All graphs locate 15 air ambulances, the current number operating in New England, and have a critical distance of 20 nautical miles around which an air ambulance covers demand.

**Figure 1.** MEXCLP optimal air ambulance base locations in New England for busy fractions $p = 0.0$, $0.2$, and $0.4$.

<table>
<thead>
<tr>
<th>$p$</th>
<th>0.00</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.35</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covered population increase (percent)</td>
<td>16.4</td>
<td>16.3</td>
<td>15.6</td>
<td>14.3</td>
<td>13.6</td>
<td>10.1</td>
<td>9.3</td>
<td>6.5</td>
</tr>
</tbody>
</table>

**Table 1.** Covered population increase by busy fraction. Higher fractions prioritize backup coverage of high demand.

Conclusion

Allocation of air ambulances according to population need at the system level will likely improve access in areas currently not served by air ambulance bases\textsuperscript{2}. Extensions of these results using interoperable data sets are likely to improve the utility of our tool for allocation of this critical service that can be used by health care stakeholders.

References

Exploring EHR-related Workload Patterns for Physician Trainees
Laura R. Baratta, Sunny Lou, MD, PhD, Hanyang Liu, Thomas Kannampallil, PhD
Washington University, St. Louis, MO

Introduction
Previous research has reported on the potential association between EHR-based clinical workload and burnout. However, few have explored the workload pattern of trainees at risk of burnout. We conducted an exploratory analysis to evaluate the characteristics of workload and assessed the specific workload patterns present among the physician trainees.

Method
We conducted a prospective longitudinal observational study with intern physicians from Medicine, Pediatrics, and Anesthesiology over a 6-month period. Participants agreed to provide their EHR-audit logs—a trail of their work activities on the EHR—over the same time period. EHR actions from audit logs were categorized into time spent on orders, notes, and review actions completed by the participants and aggregated over two-week periods. The time spent on each workload metric category were computed as well (Ouyang 2016). In addition, shift analyses such as number of shifts, average shift duration, gap between shifts, and total hours worked were computed for each participant similar to previously described (Dziorny 2019).

We utilized principal component analysis (PCA) to extract three EHR workload patterns among the trainees. We examined the top 3 principal components, which described 80% variance in the data. The loadings of the three principal components were considered highly contributory for scores with a value ≥ 0.3.

Results

![Heat Map of Loadings](image)

Figure 1: A heat map of the loading values of each workload features (x-axis) for the first 3 PC (y-axis). Color bar indicates loading values generated by PCA.

75 interns participated in this study (76.0% participation). The top loadings with values ≥ 0.3 for the first 7 PCs revealed 7 distinct workload patterns among the participants. The first pattern was found to be number of days spent on orders or notes. The second was the fraction of time in EHR spent on reviewing patient data. The third was related to time spent in EHR afterhours. The fourth was related to the time spent on inbox messages. The fifth was related to time spent on written notes. The sixth was related to the number of patients seen and the time spent on notes and orders for those patients. Finally, the seventh pattern was related to average time spent in the EHR.

Conclusion
The loading values calculated for each principal components revealed areas of great effort variability. The higher loading values signified areas where some trainees are exerting more effort than others, with greater risk for exhaustion. Thus, the uncovering of these areas allows for potential training interventions for residents on how to minimize efforts and could be monitored for trainees at high risk of exhaustion. Next steps would include clustering the participants to see the underlying subgroups of residents differing in exertion among the workload patterns.

References
Using Natural Language Processing to Identify Laterality of Nephrolithiasis Surgical Procedures

Cosmin A. Bejan, PhD,1 Siwei Zhang, MS,2 Yaomin Xu, PhD,1,2 Ryan S. Hsi, MD3
1Department of Biomedical Informatics, 2Department of Biostatistics, 3Department of Urology; Vanderbilt University Medical Center, Nashville, TN, USA

Introduction
Laterality of surgical procedures could provide important insights to better understand clinical phenotypes and patient outcomes. In the Electronic Health Record (EHR), laterality of procedural codes, such as Current Procedural Terminology (CPT) codes, can be represented by CPT modifiers. However, their utilization and validity are unknown. The purpose of this study is twofold. First, we assess the coverage of CPT modifiers expressing laterality of nephrolithiasis surgical procedures (i.e., surgery performed on the left, right, or both kidneys) and evaluate their accuracy in the EHR. Second, we develop a natural language processing (NLP) method to automatically extract such laterality attributes from clinical notes when CPT modifiers are not available.

Methods
From the de-identified EHR of Vanderbilt University Medical Center, nephrolithiasis surgical procedures were extracted using the following CPT codes: 52352, 52353, 52356, 50590, 50080, and 50081. For laterality, we identified the CPT modifiers indicating left, right, and bilateral procedures. The evaluation of CPT modifiers consisted of manually reviewing 150 randomly sampled patients with at least one nephrolithiasis surgical procedure and a corresponding CPT modifier for laterality (50 each for left, right, and bilateral).

For the procedures without CPT modifiers, a previously developed NLP system for kidney stone phenotypes was adapted to automatically extract laterality for every patient with a nephrolithiasis surgical procedure. A set of 200 patients with nephrolithiasis surgical procedures was randomly selected and, for each procedure, a laterality attribute (left, right, or bilateral) was assigned after manual review of the corresponding patient record. Pattern matching algorithms were used to analyze notes associated with each surgical procedure. Regular expressions encoding left, right, and bilateral or both were implemented to analyze the textual context of keywords including ureteroscopy, ureterorenoscopy, shockwave, kidney, ureter, ureteral, percutaneous, renal, UPJ (ureteropelvic junction), hydronephrosis, and hydroureteronephrosis. The laterality was selected based on the maximum number of matching expressions. For bilateral, this condition was relaxed to a limited number of matches of the bilateral or both expressions. A grid search was performed on 50% of the data (training set) to optimize the parameters listed in Table 1. Evaluation was performed on the remaining 50% of the data (test set).

Results
We extracted a total of 11,537 nephrolithiasis surgical procedures. Of these, 2,276 have an assigned CPT modifier for laterality and 3,219 were performed since the date of the first modifier. The evaluation of CPT modifiers for laterality indicates a precision of 98.3%, 98.1%, and 100% for left, right, and bilateral, respectively.

From the procedures without CPT modifiers available, 311 were reviewed for laterality, out of which 155, 132, and 24 were labeled as left, right, and bilateral, respectively. Table 1 lists the parameter values that achieved the best results on the training set for each laterality category. Overall, the best performance was achieved when analyzing notes at and +/- 2 days from the procedure date and matching the left context of the relevant keywords (e.g., bilateral ureteroscopy). The NLP evaluation on the test set using these values is shown in Table 2.

Conclusions
We demonstrated the potential of extracting laterality of nephrolithiasis procedures from notes when this information is not available in structured format. When available, CPT modifiers indicate laterality with high precision.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Parameter values optimized by the NLP system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
<td>Left</td>
</tr>
<tr>
<td>Days apart from the procedure date for note selection</td>
<td>2</td>
</tr>
<tr>
<td>Context laterality (left, right, both)</td>
<td>left</td>
</tr>
<tr>
<td>Context size (characters)</td>
<td>20</td>
</tr>
<tr>
<td>Minimum number of matches for bilateral expressions</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2</th>
<th>System evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laterality</td>
<td>P</td>
</tr>
<tr>
<td>Left</td>
<td>89.94</td>
</tr>
<tr>
<td>Right</td>
<td>84.21</td>
</tr>
<tr>
<td>Bilateral</td>
<td>48.72</td>
</tr>
</tbody>
</table>

P: precision, R: recall, F: F1 score

Funding: This study was supported by the National Institutes of Health (NIH) grant R21DK127075.
Standardizing PRO Data to OMOP

Rimma Belenkaya, MS, MA, Dmitry Dymshyts, MD, Michael Kallfelz, MD, Shantha Bethusamy, MS, Zachary Rachlin, MS, Julian Schwartz, MS

1Division of Health Informatics, Memorial Sloan Kettering, New York, NY.
2Odysseus, Boston, MA.

Introduction

Incorporating patient reported outcomes (PROs) in clinical oncology practice and research has become a key component of ensuring high-quality patient-centered care and evaluating new interventions. Despite the spread and scientific maturity of PRO data collection instruments, national and international methodological guidelines for choosing and using PRO instruments in the intended population, and advancements in the development of electronic PRO tools, semantic standards for PRO instruments are limited and disjoint from other terminological standards. This poses a barrier to the integration of PRO data with other clinical data. In the absence of structural standards for representing PRO data, analytical tools are not standardized and bound to proprietary PRO tools.

In this report, we are presenting an approach for semantic and structural standardization of PRO data to the OMOP Common Data Model (CDM) and Standardized Vocabularies. It addresses the challenge of integration of PRO with other clinical data and creates a foundation for open source standardized PRO analytics.

Methods

Preserving structure of the PRO instruments to continue supporting analytics intended for use with each PRO instrument. We preserved the original question-answer structure of the PRO instruments by incorporating and linking question-answer concepts in the OMOP Vocabulary. Supplementary metadata attributes, including data type and temporality, were added. PRO data was converted to the OMOP CDM Observation table based on the attribute-value structure.

Improving semantics and structure of the PRO instruments. We augmented the text of those survey questions where the original text alone, without the context of the previous question or the survey section header, was ambiguous or incomplete. Redundant concepts were deduplicated by mapping them to one standardized concept.

Integration of the PRO data with other types of health data. We precoordinated original questions and answers into concepts representing certain conditions, symptoms, and observations and mapped them to the respective standard concepts where possible. PRO data was converted to OMOP by vocabulary driven ETL and stored uniformly with the data sourced from electronic health records (EHR) and tumor registry while preserving the provenance of each data source.

Testing. We tested our approach by building patient cohorts based on integrated EHR, Tumor Registry, and PRO data using open-source standardized OMOP tools.

Results

We integrated and mapped four custom-built MSK surveys into the OMOP Vocabulary and converted data from these instruments using vocabulary driven ETL. PRO data from these instruments was combined with the data from EHR and Tumor Registry in one OMOP database. We built several patient cohorts based on the integrated data using the same standard concepts, extending the mere clinical data by patient reported data reflecting subjective aspects such as well-being or psychological impact of disease and treatment.

Discussion

We provided an approach for solving the challenge of interoperability and integration of the PRO data with other health data on the OMOP platform. This opens opportunities of using PRO data to generate evidence leveraging OMOP open-source standardized analytics. However, the challenges of standardizing PRO content to standard terminologies and difficulties of data conversion will persist until methodology for development and implementation of PRO instruments will include informatics guidelines for creating structured and well-formed content.
Classification of Dementia and Mild Cognitive Impairment from Clinical Notes

Arnav Bhandari MS, Mia Yang MD, Kathryn Callahan MD, Eric Kirkendall MD, Kristin Lenoir MPH, Nicholas Pajewski Ph.D., Umit Topaloglu Ph.D., FAMIA
Wake Forest University School of Medicine, Winston Salem NC

Introduction

The timely diagnosis of dementia and Mild Cognitive Impairment (MCI) is a challenge. In some estimates, half of the patients with dementia have not been correctly diagnosed [1]. Even fewer are coded in the Electronic Health Record (EHR) as billing codes or discrete data elements. We utilized several machine learning models to parse clinical notes in order to test a pragmatic method for classifying Dementia and MCI. The purpose of this project is to highlight the implementation of Machine Learning model classification on the set of notes belonging to a patient prior to and exclusive of the date of dementia diagnosis.

Methods

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Patient count</th>
<th>Notes Extracted</th>
<th>Average Note per Patient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dementia (DSM IV criteria)</td>
<td>435</td>
<td>10,808</td>
<td>24.85</td>
</tr>
<tr>
<td>MCI (Peterson criteria)</td>
<td>171</td>
<td>4,386</td>
<td>25.65</td>
</tr>
<tr>
<td>Other</td>
<td>101</td>
<td>3,859</td>
<td>38.21</td>
</tr>
<tr>
<td>Total</td>
<td>707</td>
<td>19,053</td>
<td>26.95</td>
</tr>
</tbody>
</table>

Table 1: Gold standard counts of patient and notes for the model diagnosis

Approval was granted by the Wake Forest Institutional Review Board. We extracted ambulatory care notes for patients who have been seen in our Kulynych Geriatric Consult Clinic, where clinicians perform detailed history and physical exams, as well as neuropsychological testing, as part of the current “gold standard” of memory testing. We extracted 19,053 Clinical notes from the EHR, belonging to 707 patients. Table 1 shows that, 56.7% notes of patients are diagnosed with dementia, and 23.0% notes of patients diagnosed with MCI. Two separate classification models are created for Dementia and MCI based on this data.

The notes are pre-processed/cleaned to standardize the data by removing punctuation, removing stop words and lemmatizing the words. Term Frequency- Inverse Document Frequency (TF-IDF) is applied to each note to create a feature vector weighing how significant a word in a document is relative to the other documents. These feature vectors are randomly split (80/20) into a training and a test set. Performance is measured by calculating the precision, recall, F-measure and accuracy of the models. Several Machine Learning approaches including Naïve Bayes, Random Forest, Neural Networks, Logistic Regression and Support Vector Machines (SVMs) were tested.

Results

For both diagnoses, python’s sklearn Neural Network with 8 hidden layers had the highest accuracy of 88.7% (Dementia) and 92.3% (MCI) predicting memory diagnoses compared to clinicians’ “gold standard” (Table 2). The Area Under the Curve for both models is approximately 96% which is indicative of a high prediction ability. Naïve Bayes performed the worst with under 80% accuracy. The study is in early phases and will be expanded on by studying clinical utility of the results i.e., management of patients who have been identified to be at higher risk for cognitive impairment by our models. Future work includes validating our model’s accuracy across our EMR.

<table>
<thead>
<tr>
<th>Model</th>
<th>Target</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dementia</td>
<td>Negative for Dementia</td>
<td>0.87</td>
<td>0.86</td>
<td>0.87</td>
<td>0.887</td>
</tr>
<tr>
<td></td>
<td>Positive for Dementia</td>
<td>0.89</td>
<td>0.91</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>MCI</td>
<td>Negative for MCI</td>
<td>0.93</td>
<td>0.97</td>
<td>0.95</td>
<td>0.923</td>
</tr>
<tr>
<td></td>
<td>Positive for MCI</td>
<td>0.86</td>
<td>0.77</td>
<td>0.81</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Neural Network Performance of different classification models

Topological Data Analysis of Emergent Patterning within Embryonic Pluripotent Stem Cells

Laura Bialozynski¹, Jaxon Green¹, Vanessa Newsome-Slade¹, Stephen Wessel¹

¹California Polytechnic State University, San Luis Obispo, California, USA

Introduction

The last decade has seen progress in the development of innovative technologies to translate the morphogenic potential of stem cells as models and treatment of disease and injuries. An area of interest is how stem cells organize in space due to cell-to-cell communication. Transitions in early development from pluripotent systems to differentiated, structurally-ordered tissue provide an ideal testbed to study generalizable properties of cellular coordination. However, accurately comparing pattern variations resulting from different experimental conditions is still an open problem. While pattern formation has been studied extensively using experiments and mathematical models, methods for quantifying self-organization are limited to manual inspection or global measures in many applications.

Methods

While convolutional neural networks are powerful tools to detect patterns, topological data analysis (TDA) is a simpler, computationally inexpensive alternative that further identifies when persistent patterns emerge and does not rely on training data. Our novel approach utilizes tools from TDA, particularly persistent homology, which allows us to assign shape descriptors to large, possibly noisy data sets. We use TDA to characterize collective behavior and self-organization in order to describe local and global patterns. Extracting point clouds from images of pluripotent stem cells during development and computing persistence diagrams allows us to compare images of developing cells.

Results and Conclusion

Our goal was to analyze topological characteristics such as connected components, loops, and cavities of the cellular patterns. This allows for thorough quantitative comparison of large sets of microscopy images and possibly future automated pattern detection. By applying TDA to point clouds, structures’ birth and death indicate the significance and longevity of emergent shapes within the data. Harnessing the principles of persistent homology allows us to create distinct persistence diagrams (PD) and bar-codes which efficiently identify developing and mature structures in stem cell clusters. With these unique representations, we can deduce the origin of emergent behavior. In Figure 1, the topological features of images (a) and (b) differ in that (b) contains persistent 1-dimensional structures whereas (a) does not contain any significant 1-dimensional holes due to the close proximity of orange dots to the dotted line.

![Figure 1](image-url)

References


Case Study on Data Standard Implementation Errors and Impact on Interoperability

Kristina Bloomquist¹, Ben Smith², Matt Prestegaard¹, Trent Peterson², Cerdi Beltré²; ¹National Marrow Donor Program CIBMTR, Minneapolis, MN, ²Principia Health Sciences, Cary, NC

Introduction
The 21st Century Cures Act is designed to help bring new innovations and advances to patients faster and more efficiently.¹ While the act has helped catalyze data standards implementation, our analysis suggests that more attention is needed to ensure the standards are implemented correctly and completely. Given that alignment of data models and vocabularies is required for successful interoperability, errors in implementation of standards can undermine the work put into their adoption. When it comes to use of shared data in medical practice and research, high-quality datasets, seamless communication across systems and standard data formats are required.² Source system data must conform to data standards to reduce and avoid amplification of data entry errors. The Center for International Blood & Marrow Transplant Research (CIBMTR) has embarked on a data transformation initiative (DTI) to improve the process by which donor and transplant patient data is shared from over 400 transplant centers. The data is complex and comprehensive, offering an excellent resource to assess source data suitability for electronic sharing.

Methods
Through DTI, the CIBMTR is transitioning some data capture from manual electronic forms completion to a more automated method that integrates with source databases. This new process captures data through the CIBMTR Reporting App (CRA), available in the Epic App Orchard®, which leverages HL7 Fast Healthcare Interoperability (FHIR) standards to allow centers to extract and transmit data from Epic databases, where the data is then integrated and transferred to the registry database. Among other requirements, CIBMTR’s automated approach stipulates that laboratory result data be formatted according to the Logical Observation Identifiers Names and Codes (LOINC) standard. An initial group of eight participating transplant centers randomly selected 2-3 patients and submitted historical routinely ordered lab panel (e.g. CBC) variables using FHIR and LOINC. The data was analyzed for data quality and compliance with data standards in preparation for production use. Three out of eight transplant centers were found to have non-compliant or missing data during the assessment. The most common errors were related to units of measure, missing LOINC codes and missing values (See Table 1).

Table 1: Comparison of Common Submission Errors and Expected Data

<table>
<thead>
<tr>
<th>Type of Error</th>
<th>Example Data Types</th>
<th>Code Received</th>
<th>Expected Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect LOINC Code</td>
<td>Lymphocyte %</td>
<td>731-0, LOINC for count</td>
<td>736-9, LOINC code for Lymphocyte %</td>
</tr>
<tr>
<td></td>
<td>Neutrophil count automated</td>
<td>753-4, Neutrophil manual count</td>
<td>751-8, LOINC code for automated</td>
</tr>
<tr>
<td>Mismatch LOINC code and</td>
<td>Segmented Neutrophils % by manual</td>
<td>769-0, LOINC code indicates</td>
<td>FHIR resource text states auto method</td>
</tr>
<tr>
<td>data text</td>
<td>method</td>
<td>manual method</td>
<td>(text and LOINC code mismatch)</td>
</tr>
<tr>
<td>No LOINC Code</td>
<td>Monocyte %</td>
<td>Only local system codes received</td>
<td>LOINC code lab</td>
</tr>
<tr>
<td>Incorrect unit of measure</td>
<td>Absolute neutrophil count (ANC)</td>
<td>770-8, K/uL unit of measure</td>
<td>UCUM units expected % for 770-8</td>
</tr>
<tr>
<td></td>
<td>Reticulocyte count</td>
<td>14196-0, M/uL unit of measure</td>
<td>UCUM units expected 10³/³/uL or 10⁹/L</td>
</tr>
<tr>
<td>Some data missing</td>
<td>Pharma cell % observation</td>
<td>LOINC code, no values</td>
<td>Complete dataset (including values)</td>
</tr>
</tbody>
</table>

Conclusion
Interoperability is vital to the success of the vision of the 21st Century Cures Act and collaborative research leveraging clinical care data. Significant time has been spent defining the sharing models, technology, and data standards. These are important issues, however, the conformance of data within source systems cannot be overlooked. Errors in source data systems are typically introduced through manual data input errors, data standards mapping errors, and risk propagation into data repositories through data sharing. Early results from this case sample suggests source data systems require equal or greater attention as technical standards and data sharing model development. More participants will result in a larger sample size and better understanding of the scale of this issue.

References
Continuous prediction of organ dysfunction in general wards

James Brogan, BS¹, Boudewijn Aasman, BS¹,², Michelle Ng Gong, MD, MS¹,², Parsa Mirhaji MD, PhD¹,²
¹ Albert Einstein College of Medicine, Bronx, NY, United States
² Montefiore Medical Center, Bronx, NY, United States

Introduction

Prediction of sepsis in general wards is a major challenge because there is no gold standard for diagnosis, leading to models based on consensus definitions¹,². Predicting organ dysfunction instead of the construct of sepsis may provide an opportunity for intervention before patients develop sepsis-induced organ failure. We derived the Predicting Onset of Organ Failure (PROOF) model to identify patients at risk of organ dysfunction.

Methods

Retrospective cohort study of adults admitted to three hospitals within the Montefiore Health System from 1 February 2020 through 31 May 2020. The cohort was randomly split 80:5:5:10 at the patient level into the training, validation, calibration, and test sets. Sequences of lab values, vital signs, observations, and medication orders were used as predictor variables. Organ dysfunction events were defined as an increase in Sequential Organ Failure Assessment (SOFA) score ≥ 2 within the next 48 hours of hospitalization and were computed every two hours for each admission. Recurrent Neural Networks, a form of deep learning sensitive to temporality and sequences of events, were used to create a predictive model. PROOF was computed every two hours for each admission.

Results

The cohort was comprised of 17,722 unique adult patients and their 20,704 admissions. Organ dysfunction, as defined above, occurred in 5,765 (27.8%) admissions. PROOF had an area under the receiver operating characteristic curve of 0.843 (95% CI 0.839 – 0.846) and an area under the precision-recall curve of 0.405 (95% CI 0.396 – 0.414) in the hold-out test set. At a threshold of 0.25 determined by F₁ score, PROOF had a sensitivity of 0.452, specificity of 0.942, and a precision of 0.410.

Conclusion

We present a deep learning model for early detection of organ dysfunction in general wards. Our model focuses on predicting organ dysfunction, rather than the construct of sepsis, and shows promise for real-time risk assessment.

Figure 1. Model performance by receiver operating characteristic and precision-recall curves.

Receiver operating characteristic (a) and precision-recall (b) curves for the risk that organ dysfunction will occur within the next 48 hours.

References

Usability Testing the Hematology Patient Explorer and the Importance of Leveraging Multiple Usability Assessment Methodologies

Zach Burningham PhD MPH\textsuperscript{1}, Kyle Hansen PhD\textsuperscript{2}, Andrew Post MD PhD\textsuperscript{2}, Vikas Patil MS\textsuperscript{1,2}, Sarah Soderborg\textsuperscript{2}, Luisa Rusta\textsuperscript{2}, Michael Deininger MD PhD\textsuperscript{2}, Richard McShinsky\textsuperscript{2}, Ahmad Halwani MD\textsuperscript{2}

1. Division of Epidemiology, Internal Medicine, University of Utah, Salt Lake City, Utah
2. Division of Hematology and Hematologic Malignancies, Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah

Introduction: The University of Utah Huntsman Cancer Institutes’ (HCI) Hematology Clinical Data Science and Health Informatics (CHI) program has developed the Hematology Patient Explorer (HPE); a health information technology (IT) solution that acts as a self-service point-and-click cohort extractor supported by a backend harmonized data lake comprised of near-real time data. The HPE allows one to identify target cohorts through selection of diagnoses (i.e., ICDO-3), patient features (i.e., gender), and disease characteristics (i.e., stage). A participatory design approach was followed in defining core requirements and in developing initial prototypes. Stakeholders included division leadership, clinicians, researchers, and the development team. As part of our development process, we chose to implement formal usability testing among a sample of potential end-users before releasing the web app for routine use. Here we summarize our usability testing approach, feedback, and significance for leveraging multiple usability assessment methodologies.

Methods: Methods of usability assessment included task analysis, direct observation, and questionnaires [1]. Usability testers were selected from a pool of research investigators and assistants within HCI’s Division of Hematology. We developed an interview protocol administered over video conferencing software which supported real-time observation via screenshare. The protocol was comprised of 10 pre-defined tasks designed to ensure the tester interacted with all core components of the HPE. A series of pre-defined follow-up questions were also administered following the completion of tasks to capture user satisfaction with the system. The questions administered were based on the System Usability Scale (SUS) developed by Brooke [2]. Test administrators leveraged both think-aloud and retrospective probing moderating techniques.

Results: Six potential end-users completed usability testing of the HPE; each interview took approximately 35 minutes. A mean SUS score of 95.0 (SD = 3.8) was achieved, indicating superior usability of the HPE in comparison to the recommended standard of acceptability (SUS score: 68). All participants completed each task with a high degree of accuracy. Testers praised the app’s ease of navigation, accessibility, querying performance, and the ability to export patient-level reports for further analysis. Even though the SUS score was superior, the qualitative thematic review revealed improvement could be made in the following 4 domains: visual clarity, feature distraction, intuitiveness, and accuracy.

Discussion: Leveraging multiple methods of usability assessment (i.e., task analysis, direct observation, questionnaire) has allowed us to gather rich information and insight into productive solutions for addressing negative feedback that may not have been achieved if relying on questionnaire instruments alone. For example, negative experiences were not reflected in the SUS score and subsequently were only captured through direct observation and from qualitative feedback obtained through think-aloud moderating during task completion.

Conclusion: Usability testing the HPE was vital to ensuring the app’s content and functionality were appropriate for the tool’s intended use. Usability testing should be a standard practice and a prerequisite to releasing Health IT solutions for routine use. Furthermore, multiple assessment strategies should be leveraged as relying on heuristic checklists or questionnaires alone can produce misleading results.

References
Using Synthea™ Derived Synthetic Data to Simulate Randomized Clinical Trials: the SPRINT Trial Experiment

Francis X. Campion, MD, Brendan Abraham, MS, Dylan Hall, BS, Rute Martins, MS, Jason Walonoski, MS, Eldesia Granger, MD MPH
The MITRE Corporation, Bedford, Massachusetts

Introduction

Learning from randomized controlled trials (RCTs) can be limited by study population characteristics, long study timelines and resource limitations. We explored the potential to reliably replicate and extend learnings from the well-known SPRINT trial1 using Synthea™-2 derived synthetic patient data.

Methods

We generated a synthetic population using existing disease progression and standards of care modules from Synthea, refined to reflect critical characteristics of the SPRINT trial population. Additional Synthea modules were built using the Synthea Module Builder3 to mimic the study enrollment and randomization protocol. Newly developed Java-based modules were created to represent trial interventions for each treatment arm. This enabled logic to mimic clinical decision making of medication selection and titration used to treat hypertension in the trial. We then used the full stack of Synthea modules for population generation, study enrollment and interventions to generate multiple iterations of the clinical trial. Exported Synthea patient-level data for each simulation was then processed and analyzed using Python scripts and Jupyter Notebook4.

Results and Discussion

We were able to reasonably approximate the synthetic trial population to the actual SPRINT population across critical variables of interest, such as age, race and gender, as well as clinical characteristics of renal function and blood pressure levels at time of enrollment. This required several iterations of refinements to Synthea modules combined with a post-simulation stratified down-sampling strategy consistent with that used for patient enrollment and randomization in the original SPRINT trial. We found consistent blood pressure control patterns including time to onset of BP control and the number of medications needed to achieve control across runs of varying trial size populations (10,000, 20,000, 40,000, 100,000). We found more heterogeneity in trial outcomes, including divergence in the statistical significance of certain outcomes. Some of this divergence can be explained by simplifications and modeling choices in Synthea modules, which could be refined to produce more accurate and reliable results. In addition, Synthea modules also reflect more current clinical practice, which can also explain divergence in the prevalence of certain cardiovascular outcomes between the original SPRINT trial and the results of synthetic trial runs.

Conclusion

We demonstrated an end-to-end approach to simulating a randomized clinical trial using Synthea derived synthetic data. With further refinement we believe that synthetic clinical trials may become an important adjunct for planning and extending learning from RCTs by simulating variable enrollment and intervention characteristics.

References


https://synthetichealth.github.io/module-builder/


Approved for Public Release: Case No. 21-2520. Distribution Unlimited. ©2021 The MITRE Corporation
Identification of Post-acute Sequelae of COVID-19 from Electronic Health Records

Victor M Castro, MS1, Vivian S Gainer, MS1, Justin Manjourides, PhD2,3, Roy H. Perlis, MD3, Shawn N Murphy, MD, PhD1,4

1Mass General Brigham, Somerville, MA; 2Northeastern University, Boston, MA; 3OHDSI Center at The Roux Institute, Northeastern University, Portland, ME; 4Massachusetts General Hospital, Boston, MA

Introduction

Identifying signs, symptoms and other ill-defined conditions is notoriously difficult in electronic health record (EHR) and claims data. While billing codes, problem list and other structured fields exist, they typically suffer from low sensitivity and variable use across data sources. Accurate identification of symptomology has become increasingly important during the COVID-19 pandemic. The goal of this work is to build a generalizable and portable library of symptoms defined from EHR data linked to common vocabularies for use in observational studies of COVID-19 and post-acute sequelae of COVID (PASC).

Methods

We identified a set of 12 COVID and PASC-related symptoms that were reported in the literature and had available validation data. For each symptom, we identified appropriate ICD-10-CM diagnosis codes in the symptom chapter (ICD-10 R*) and other related disease codes. We also developed a natural language processing (NLP) approach to identification of symptoms by manually and iteratively curating token lists derived from survey and other literature sources. These token lists were combined with rules for negation and family history exclusion and implemented as structured query language (SQL) full text index queries. All clinical notes during a COVID-19 admission at MGH were then loaded into a SQL server database with full-text indexing enabled to allow rapid evaluation and iteration of the NLP rules. Full-text SQL querying is implemented in most major relational databases allowing portability and implementation in both i2b2 and OMOP common data models. The developed ICD-10 and NLP rules are evaluated against a ‘gold standard’ manually annotated registry of consecutive Massachusetts General Hospital (MGH) COVID-19 admissions between March and July 2020 [1].

Results

Table 1 includes performance of 3 symptom definitions validated against the gold standard data. NLP rules were significantly more accurate compared to ICD-10 rules. Sensitivity of ICD-10 only definitions of anosmia, headache and fatigue was poor compared to NLP. We are in the process of validating additional symptoms.

<table>
<thead>
<tr>
<th>Symptom</th>
<th>Source</th>
<th>Balanced accuracy</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>anosmia</td>
<td>ICD10</td>
<td>0.503</td>
<td>0.999</td>
<td>0.008</td>
<td>0.500</td>
<td>0.863</td>
</tr>
<tr>
<td>anosmia</td>
<td>NLP</td>
<td>0.745</td>
<td>0.810</td>
<td>0.680</td>
<td>0.363</td>
<td>0.941</td>
</tr>
<tr>
<td>headache</td>
<td>ICD10</td>
<td>0.553</td>
<td>0.941</td>
<td>0.165</td>
<td>0.403</td>
<td>0.824</td>
</tr>
<tr>
<td>headache</td>
<td>NLP</td>
<td>0.643</td>
<td>0.457</td>
<td>0.830</td>
<td>0.268</td>
<td>0.918</td>
</tr>
<tr>
<td>fatigue</td>
<td>ICD10</td>
<td>0.518</td>
<td>0.891</td>
<td>0.144</td>
<td>0.343</td>
<td>0.725</td>
</tr>
<tr>
<td>fatigue</td>
<td>NLP</td>
<td>0.558</td>
<td>0.201</td>
<td>0.914</td>
<td>0.311</td>
<td>0.856</td>
</tr>
</tbody>
</table>

Conclusion

In this work we have begun to develop and assess EHR symptom definitions to be used in observational and population health studies of COVID-19 and PASC. We use manually collected cohort study data to validate these EHR definitions so they can be accurately applied to larger populations available in EHR databases. Future work will assess performance of the definitions in under-represented populations.

References

Translating Hypertension Treatment Guidelines into Patient-Facing Clinical Decision Support

Ellie Y. Chang1,2, Christopher A. d’Autremont, MBI2, David A. Dorr, MD MS2, Robert Rope, MD2

1Princeton University, Princeton, New Jersey; 2Oregon Health & Science University, Portland, Oregon

Introduction
Nearly half of adults in the United States have high blood pressure (hypertension); of those with known hypertension, roughly half are uncontrolled,1 leading to increased risk of heart disease and stroke, two leading causes of death for older adults.2 Although clinical guidelines for the treatment of hypertension represent a wide breadth of biomedical knowledge, their implementation remains inconsistent and resource intensive. Guidelines often conflict on the appropriate diagnosis and treatment for hypertension, which can result in undertreatment and impaired understanding of treatment options.3 Comprehensive hypertension management can be complex, relying heavily on patients to adopt healthy behaviors and self-manage. Multiple factors, including a patient’s co-morbid health conditions, medication side effects, social support, and demographic factors can significantly influence the approach needed for effective treatment.3 Behavioral and lifestyle factors, including alcohol consumption and physical activity, also substantially affect treatment and outcomes. Clinical decision support (CDS) can address these challenges, presenting more complete information, tailored to a patient’s specific treatment needs.

Methods
Creating computable and personalized guideline-based recommendations can streamline hypertension monitoring and diagnosis, as well as improve care, by better engaging patients in the care management process. We are developing a CDS tool to translate hypertension guidelines into practice using Clinical Quality Language (CQL) modules and incorporate patients’ self-selected preferences into treatment. This approach will elicit the information needed for patient-facing clinical decision support, as well as allow patients to self-assess treatment progress and adherence through a workflow sensitive solution. Our specific innovations include the use of blood pressure sets to match guidelines (e.g., at least 4 office or 6 home blood pressures);4 a counseling, goal, and preference system for non-pharmacologic recommendations; and specific patient and care team versions of the recommendations.

Discussion
We created recommendations that patients will be given based on their engagement with the tool. In total, we use eight clinical guidelines that encompass 71 recommendations as the foundation for the logic behind six distinct workflows. The diagnosis and monitoring workflow (Figure 1) makes an initial evaluation based on a set of blood pressure readings and shows the patient a recommendation associated with said evaluation. For patients with high blood pressure who have not been diagnosed with hypertension, the patient is first recommended to reach out to their care team for a diagnosis. For patients already diagnosed with hypertension, the tool would suggest non-pharmacologic interventions through the non-pharmacologic workflow if such interventions are applicable for the patient. The recommendations (example in Figure 2) are concise and easy to understand and encourage patients to work alongside their care team in the treatment process. Ultimately, the goal of these recommendations is to empower patients to manage their own care – a prerequisite for the successful treatment of hypertension.

References
Evaluation of Apache cTAKES, MetaMap, and QuickUMLS for SDoH extraction from clinical narratives

Lokesh K. Chinthala, MS¹, Billy Barnett, MS¹, Cindy K. Lemon, MS¹, Charisse Madlock-Brown, PhD¹
¹University of Tennessee Health Science Center, Memphis, TN

Introduction
Social Determinants of Health (SDoH) are strongly associated with health risks and outcomes. Most SDoH information in Electronic Health Record (HER) systems is stored in narrative text, which requires extraction to be useful for research and surveillance. The purpose of this paper is to evaluate the performance of three widely used clinical Natural language processing (NLP) applications - Apache cTAKES, MetaMap, and QuickUMLS in identifying SDoH concepts in clinical narratives.

Methods
Using the UTHSC research Enterprise Datawarehouse (rEDW), we identified a corpus of 88,459 clinical documents associated with the 1,395 patients having SDoH conditions based on the ICD-10-CM SDoH diagnosis codes in the rEDW data. We then applied cTAKES, MetaMap, and QuickUMLS on the corpora to compare the performance of extracting the SDoH concepts from the clinical narratives. We manually reviewed a 1% subset from each SDoH domain for accuracy. The performance metric used in the evaluation is the positive predictive value (PPV).

Results
As shown in Table 1, the systems varied in the number of identified SDoH concepts. For instance, in the problems related to employment and unemployment subset, MetaMap tagged half as many as cTAKES. Similarly, in the problems related to social environment subset, cTAKES performed poorly compared with the other two NLP applications. Additionally, the overall PPV ranged from 0.68 to 0.86.

Table 1. The systems identified overall documents with SDoH conditions.

<table>
<thead>
<tr>
<th>SDoH Condition</th>
<th>ICD Code</th>
<th>Number of Documents tagged</th>
<th>PPV of 1% subset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>cTAKES</td>
<td>QuickUMLS</td>
</tr>
<tr>
<td>Problems related to employment and unemployment</td>
<td>Z56</td>
<td>674</td>
<td>714</td>
</tr>
<tr>
<td>Problems related to housing and economic circumstances</td>
<td>Z59</td>
<td>9323</td>
<td>9311</td>
</tr>
<tr>
<td>Problems related to social environment</td>
<td>Z60</td>
<td>27</td>
<td>56</td>
</tr>
<tr>
<td>Other problems related to primary support group, including family circumstances</td>
<td>Z63</td>
<td>114</td>
<td>0</td>
</tr>
<tr>
<td>Problems related to psychological circumstances</td>
<td>Z65</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

Conclusion
Our results show that the popular NLP applications can identify SDoH with varying accuracies. While cTAKES tagged the most documents, MetaMap had an overall better performance. We observed consistent text for each SDoH tagged accurately and consistent phrasing related to inaccuracies during the manual review. Additionally, we identified documents indicating a patient was screened for an SDoH, while the SDoH was not present. Our findings can form the basis for an SDoH-specific NLP dictionary for accurate extraction of SDoH concepts.
Comparative Protein-Protein Interaction Network Analysis of Urothelial Carcinoma Subtypes

Charissa Chou, Alper Uzun1, PhD, Ece Gamsiz Uzun-4, PhD
(1) Department of Biology, Brown University, Providence, RI. (2) Department of Pediatrics, Alpert Medical School, Brown University, Providence, RI. (3) Rhode Island Hospital, Department of Pathology, Providence, RI (4) Department of Pathology and Laboratory Medicine, Alpert Medical School, Brown University, Providence, RI.

Cancer is a complex disease and the second leading cause of death in the United States. According to the American Cancer Society, there were 17,900 deaths from bladder cancer in 2021, which is four times more prevalent in men than women. Urothelial carcinomas can be classified as two subtypes, papillary and non-papillary carcinomas, which present histological and molecular differences. The genomic landscape distinguishing the subtypes remains to be studied; our approach to demystify the genomic landscape uses the tool Proteinarium3, which is a novel multi-sample protein-protein interaction (PPI) tool, to identify clusters of patients with shared PPI networks (Figure 1).

This novel network biology approach was applied to two datasets1,2 to investigate how the PPI networks of papillary urothelial carcinomas differ from the PPI networks of non-papillary urothelial carcinomas. To create the individual patient PPI networks, for each patient, we ranked the available RNA-seq z-score data and used the top 100 upregulated genes (seed genes) for input into Proteinarium. After generating the network graphs for each patient, the similarity between each pair of graphs is calculated by using the Jaccard distance and recorded in a matrix. The similarity matrix is used as the input to cluster the set of graphs. Unweighted Pair Group Method with Arithmetic Mean (UPGMA) was used to build the dendrograms. The further details of Proteinarium are described in Armanious et al 20203. The papillary and non-papillary PPI networks contained genes unique to each network, which were implicated in phenotype-associated pathways (Figure 1). We found functionally enriched pathways distinct to the shared gene networks of the papillary and non-papillary groups using gProfiler. The MAPK signaling and FGFR signaling pathways were uniquely enriched in the papillary cancer while PI3K/Akt pathway and TP53 regulation pathway were uniquely enriched in the non-papillary cancer network. Our findings suggest that papillary and non-papillary urothelial carcinomas may proliferate via parallel pathways – the MAPK-Ras-Raf-ERK pathway and PI3K/Akt signaling pathway, respectively. Our findings for non-papillary urothelial carcinomas are supported by the analysis of a second dataset5. Separation scores of the distance between the papillary and non-papillary networks in the interactome showed that papillary and non-papillary urothelial carcinomas are distinct molecular entities with biomarkers that can be used as therapeutic targets and predictors of disease development.

References
Development of an Environmental Scan to Assess an Electronic Clinical Quality Measure for the Diagnosis of Venous Thromboembolism in Primary Care

Erin Coghlan1, Avery Pullman, BS1, Ania Syrowatka, PhD1,2, Michael Sainlaire, MS1, Wenyu Song, PhD1,2, Patricia C. Dykes, PhD, RN1,2, Department of Medicine, Brigham & Women’s Hospital, Boston, MA; Harvard Medical School, Boston, MA

Background

This study demonstrates the process for developing a comprehensive environmental scan for an electronic Clinical Quality Measure (eCQM) assessing venous thromboembolism (VTE) diagnostic delay of adults aged 18 years and older in Primary Care at the provider group level. The environmental scan is necessary for measure development to collect and interpret relevant data for strategic planning and project design.

Methods

A literature review was conducted to identify relevant publications assessing VTE management and diagnostic delays. The Harvard Countway Librarian assisted with these searches, conducted in PubMed, to compile papers that met the eCQM criteria: related to VTE in last 10 years and focused on adult population. Titles and abstracts of each paper were collected and included in the literature review. Additionally, using the Centers for Medicare and Medicaid Services (CMS) Measure Inventory Tool (CMIT), National Quality Forum (NQF) Positioning System, and Harvard Librarian, existing measures and guidelines related to VTE diagnosis or treatment practices within the last 5 years were examined. These searches were conducted to inform the specifications of the proposed eCQM and to determine if there were any competing measures in existence or under consideration. A stakeholder and expert feedback section was added to the environmental scan to provide a supplementary analysis for the measurement development process. Stakeholder feedback is ongoing and will include a series of five clinician and five patient interviews and Technical Expert Panel meetings. These methods, in addition to a background analysis and review of regulations, provides the framework for eCQM advancement to assess diagnostic delays of VTE in primary care.

Results

Literature Reviews: A total of 39 articles were extracted that met the inclusion criteria.

Summary of Clinical Practice Guidelines: Five guidelines that are relevant to the diagnosis and management of VTE were identified.

Review of Competing and Related Measures: No competing measures for this eCQM were developed. Two related measures were identified: Perioperative Care: Venous Thromboembolism (VTE) Prophylaxis and Perioperative Pulmonary Embolism or Deep Vein Thrombosis Rate. A measure summary is provided in the table below (Table 1).

Stakeholder and Expert Feedback: Of the three providers and one patient interviewed to date, all participants are in support of the measure development to eliminate VTE diagnostic delays.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Data Source</th>
<th>Competing</th>
<th>Related</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perioperative Care: VTE Prophylaxis</td>
<td>Centers for Medicare &amp; Medicaid Services (CMS)</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Perioperative PE or DVT Rate</td>
<td>Agency for Health Research and Quality (AHRQ)</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Proposed eCQM</td>
<td>Brigham and Women’s Hospital (BWH)</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 1: Related Measure Summary

Conclusion

This environmental scan includes a literature review, guideline review, competing measure analysis, and stakeholder feedback to construct an eCQM focused on assessing delayed VTE diagnosis in Primary Care. In addition to the eCQM, this information will ultimately be utilized in the development of a Clinical Decision Support (CDS) alert system that notifies clinicians when a patient is at high risk for VTE to promote timely diagnosis of the disease.

References

A Framework to Integrate, Explore, and Visualize Patient-Generated Health Data, Patient Reported Outcomes and Clinical Data: PPP-Integrator

Scott Cukras, MS, Rodrigo Carvajal, BS, Rhoda Gai-Cherry, MS, Sashanna Roman, MPH, Laura Oswald, PhD, Brian D. Gonzalez, PhD
Moffitt Cancer Center, Tampa, FL

Cancer researchers are challenged by the amount, frequency and heterogeneity of the data required to perform translational cancer research. The challenges increase with the integration of continuous data from wearable devices and self-reported information. The integration, exploration and visualization of longitudinal patient data demands a carefully designed, multifaceted infrastructure and robust technology stack for ease of use for the study participants and researchers. However, herein lies the challenge. There is a need to provide a solution capable of handling data acquisition, data integration, and data rendering with the purpose of facilitating the analysis and navigation of integrated patient-generated health data (PGHD), patient reported outcomes (PRO), and clinical data extracted from the electronic health records (EHR).

We built the Patient PRO PGHD Integrator (PPP-Integrator), a secure and HIPAA compliance framework that integrates clinical data, PGHD from Fitbit devices and PROs captured via REDCap. PGHD can include heart rate, physical activity, dietary intake, and sleep outcomes. PRO measures in use include PROMIS-29+2, PRO-CTCAE, and FACT-G instruments. PPP-Integrator is currently being used in two studies that collectively follow 90 patients with cancer for a duration of 60 to 90 days.

Moffitt Cancer Center (MCC) built a clinical data warehouse® that PPP-Integrator uses to associate patients’ clinical data to their PGHD and PROs. Fitbits are activated and registered to the study prior to being provided to a patient. Once a patient has their own Fitbit, they wear it during the observation period. The device automatically syncs with a loaner tablet, which will then upload their data to Fitbit’s cloud server (See Figure 1). PPP-Integrator uses Fitbit’s secure API and makes data pulls via a PHP-based server, transforms the data to facilitate interoperability, and finally stores the data in MongoDB. PPP-Integrator’s web interface combines multiple data visualization and reporting microservices built on ReactJS and NodeJS that at the query stage integrate requests for data across MongoDB (Fitbit), Oracle (HER) and a GraphQL server that resolves data queries against REDCap’s API that access MySQL. All databases are behind the firewall. See Figure 2 for an example visualization of PROs using a heatmap; green indicates low levels of pain and red indicates high levels. Other visualizations display time series data for heart rate, activity, and sleep efficiency. See Figure 3, change overtime (five days period) in overnight sleep efficiency, where green indicates good level of sleep and red bad levels of a patient.

Patients can either enter the PRO at the clinic or receive by e-mail the REDCap’s URL of the instruments. No PHI data is captured in REDCap. PPP-Integrator provides a robust data architecture for acquiring information, transforming, and writing that information to a database and using microservice-based data visualization web apps to support cancer research.


**Figure 1.** PPP-Integrator Architecture

**Figure 2.** PRO heatmap

**Figure 3.** Fitbit overnight sleep efficiency
Predicting Patients Most Likely to Respond to Mobile Health Messaging
Esha Datta, PhD, Colleen Bouey BS, Courtney Ng MS, Arielle Slam MPH, Raj Behal, MD, MPH
One Medical, San Francisco, CA

Introduction: Digital health interventions are an effective way to engage patients with healthcare and promote preventative health. However, several qualitative studies have shown that many of these patients are prevented from responding to digital interventions by barriers such as technology literacy, personal motivation, or impersonal messages. In this study, we created a model to predict which patients were most likely to respond to a digital request to book a physical. Using this quantitative approach, we then determined the most important features for the predictions.

Methods: 132,728 members of a primary care practice who had never engaged with the practice received a suggestion to book an annual physical with their provider through the practice’s mobile app. A gradient boosted classifier model was used to predict whether a member would book a visit within 14 days. Model performance was evaluated using k fold cross validation and relative feature importances were calculated using Shapley values. The features considered in this model included patient demographics, diet and exercise habits, patient engagement, and available capacity of appointments.

Results: 2.7% of these patients who had never engaged booked a visit within 2 weeks of receiving a message and the model predicted the bookings with an AUC of .8. Among the most important features in the model were the number of days since they had installed the mobile app, the number of days since they had last used the mobile app, the patient’s age, whether they lived in the San Francisco Bay area, and whether the patient had selected a primary care provider.

Conclusions: Digital health interventions may be most effective for young, tech-enabled patients who are engaged in their healthcare. For less engaged, older, or technologically challenged patients, more aggressive and more personal interventions, such as health coaching or personalized phone calls could be more effective.

References
Predicting Out of Control Hypertensive Patients With Blood Pressure History
Esha Datta, PhD, Colleen Bouey BS, Raj Behal, MD, MPH
One Medical, San Francisco, CA

Introduction: Early and fast control of blood pressure for hypertensive patients is an important way to manage cardiovascular risk and prevent more serious conditions\(^1\). Predictive models have been shown to be a promising way to identify hypertension patients at risk of having uncontrolled blood pressure. Previously, predictive models have relied on a vast number of features, such as patient demographics, diagnoses, medications, vitals, labs, and other health history\(^2\), that are not always easy to access in many systems. This study shows that we are able to achieve similar predictive power by relying solely on blood pressure history.

Methods: This study aims to predict which patients would have uncontrolled blood pressure if they were measured today and also to identify the factors that contribute to that prediction. The control status from the most recent blood pressure reading of 42,911 hypertension patients who had at least two readings was predicted using features engineered from blood pressure history data. 11,711 of these patients (27.3\%) had out of control readings. A gradient boosted classifier model was used to predict whether the member had a systolic blood pressure greater than 140 or a diastolic blood pressure greater than 90. Model performance was evaluated using k fold cross validation and relative feature importances were calculated using Shapley values\(^3\).

Results: The model predicted the control status of the hypertension patients with an AUC of .72, which is comparable to previous studies. In comparison, predictions based solely on the previous reading resulted in an AUC of .65. Among the most important features of the model were the last blood pressure reading, the average of previous blood pressure readings, number of days since the last reading, number of days since last in or out of control, and number of blood pressure readings in the history.

Conclusions: On its own, blood pressure history has strong predictive power for determining whether a hypertension patient is out of control. Including the full picture of previous readings increases the performance of the model above the use of just the most recent reading.

References
The Iowa Health Data Resource data enclave for processing patient data: Initial design and pilot results

Heather Davis, MS, MLIS1, Joseph Hetrick, BS2, Michael Frangi, BBA, MA, PMP3, Joseph Wagner, BS, MA, MBA3, Boyd Knosp, MS, FAMIA4
1Institute for Clinical & Translational Science, University of Iowa, Iowa City, IA; 2Information Technology Services, University of Iowa, Iowa City, IA; 3Healthcare Information Systems, University of Iowa Hospitals and Clinics, Iowa City, IA

Introduction
Analyzing patient data plays a central and critical role in advancing research across the broad spectrum of healthcare. Access to these data in a manner that enables studies involving machine learning and other artificial intelligence techniques is limited at the University of Iowa (UI) due to the lack of infrastructure to process the patient data in an environment acceptable to institutional data governance. Nationally the acceptability of cloud based solutions is for sensitive data is variable, and depends on the organization’s approach to mitigating risk. One way to address this is to install computational resources behind the health care organization’s firewall – thus keeping all the data within the Health Insurance Portability and Accountability Act covered entity while replicating expensive compute resources that are available in other parts of the campus. This project designs a more cost-effective architecture as part of the Iowa Health Data Resource (IHDR)1 where data is stored within the hospital firewall but is accessible – through policy and security settings – to existing compute resources installed outside of the hospital in UI data centers. This leverages existing computational resources while maintaining higher level of protections required to work with this data. This poster describes the design of this data enclave and reports on pilot uses of this resource.

Methods and Results
The business requirements of the IHDR data enclave were determined by UI Healthcare (UIHC) leadership and included two key requirements: 1. Datasets extracted from patient data need to remain in the hospital data center and 2. Access to this data must be controlled by Health Care Information Systems (HCIS). The enclave was created by a unique collaboration between University IT, HCIS, and IHDR team members to establish compute-resource linked storage devices in the UIHC data center. The environment designed allows processing of the data (governed by IRB protocols) by direct connection to compute resources outside the hospital firewall. The need to download data is eliminated as the computational scientists who are on the IRB protocol are issued credentials to access data extracts from existing computational resources such as the UI Interactive Data Analytics Service2 or the UI central computational cluster, Argon3. The IHDR data enclave eliminates the need to replicate computational infrastructure inside the hospital firewall, saving costs.

The enclave design consists of 1) data secure storage within healthcare data center, 2) direct network from storage to compute resources, 3) access control via healthcare managed credential group. We will report on the progress of the data enclave and on initial pilot work of two projects including predictive analytics applied to maternal-fetal health conditions leading to preeclampsia and a text analytics project that performs natural language processing of notes looking for social determinants of health.

Conclusion
We have designed an architecture that addresses the scientific needs of data driven health research projects that takes into account the structural, political and technical environment at our institution to provide previously unavailable computational capabilities. UI researchers are positioned to participate in new areas of research, including interinstitutional collaborations. Next steps will include; collecting feedback from pilot work, process evaluation, and service usage metrics. Future research is needed to compare approaches across institutions.

References
Effects of Limited Sample Size on the Performance of Commonly Used Neural Networks: A Case Study Using PubMed Hepatitis Case Report Abstracts

John Del Gaizo, MS, Ahmed Alameldin, BS, Jihad S Obeid, MD
Biomedical Informatics Center, Medical University of South Carolina, Charleston, South Carolina, USA

Introduction
We present three common AI model architectures for screening of hepatitis from physician-authored PubMed reports (title and abstract). For many translational researches, obtaining large amounts of clinical text may be infeasible. To identify appropriate models for smaller sample sizes, we examined the performance of these models at different sample sizes, from 500 to 2000 points per class. The 3 tested architectures were a CNN (convolutional neural network) (1), an LSTM (bi-directional Long Short-Term Memory Recurrent Neural Network), and a HAN (Hierarchical Attention Network) (2). We compare performance in terms of accuracy, AUC, and training time. We also examined X-AI (eXplainable – AI) methods to identify key terms and symptoms: CNN through the Grad-CAM (Gradient – Class Activation Mapping) (3) algorithm, and HAN through visualization of attention weights.

Methods
We downloaded 6,726 reports (title and abstract) from PubMed, 2,268 of which were labeled as Hepatitis through the MeSH system. The data was cleaned through regex (regular expressions), which performed steps such as punctuation removal; date substitution with a token; and separation of contextual information, i.e. “89yo” to “89 yo”. For any given sample size our experiment was as follows: 1) sample the same number of positive and negative reports from the overall dataset, where the total number of reports equals a specified sample size. 2) Separate the sample into train (67.5%), test (7.5%), and validate sets (25%). 3) Train a tokenizer from the train set. 4) Tokenize the terms. 5) Separate the sentences (HAN only). 6) Pad or cutoff the samples so each is of equal length: 400 tokens for CNN/bi-LSTM, 45 tokens/sentence and 20 sentences for HAN. 7) For each model, train on the train set until the test set performance stops improving. 8) For each model, record the results on the validation set.

The above 8 steps were run 25 times for each sample size. The experimented sample sizes were 1000, 2000, 3000, and 4000; 500, 1000, 1500, and 2000 samples per class, respectively. We therefore ran a total of 100 experiments. Finally, we visually inspected the output of Grad-CAM (CNN), and attention weights (HAN) on sampled reports.

Results
Table 1. AUC and Accuracy (Acc.) results for each model and sample size.

<table>
<thead>
<tr>
<th></th>
<th>1K AUC</th>
<th>2K AUC</th>
<th>3K AUC</th>
<th>4K AUC</th>
<th>1K Acc.</th>
<th>2K Acc.</th>
<th>3K Acc.</th>
<th>4K Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>.97±0.01</td>
<td>.98±0.01</td>
<td>.98±0.00</td>
<td>.98±0.00</td>
<td>.94±0.01</td>
<td>.94±0.01</td>
<td>.94±0.01</td>
<td>.95±0.01</td>
</tr>
<tr>
<td>Bi-LSTM</td>
<td>.97±0.01</td>
<td>.97±0.01</td>
<td>.97±0.00</td>
<td>.98±0.00</td>
<td>.93±0.01</td>
<td>.94±0.01</td>
<td>.94±0.01</td>
<td>.94±0.01</td>
</tr>
<tr>
<td>HAN</td>
<td>.94±0.02</td>
<td>.94±0.01</td>
<td>.93±0.06</td>
<td>.94±0.01</td>
<td>.94±0.01</td>
<td>.93±0.02</td>
<td>.92±0.09</td>
<td>.94±0.01</td>
</tr>
</tbody>
</table>

The mean training times (seconds) were CNN: 18, 32, 45, 58; LSTM: 22, 38, 53, 78; HAN: 22, 37.5, 54, 76.

Discussion
The 3 models did not have statistically significant differences in accuracy. However, the CNN and bi-LSTM had higher AUC then HAN, and CNN had less computation time. Importantly, there was only a slight difference in mean classification performance across sample sizes. Hence, these models can be applied to smaller datasets. This CNN architecture lends itself to Grad-CAM analysis as the convolution filters are run in parallel. If the filters are in sequence (such as deep nets), then interpolation would blur Grad-CAM output. A limitation of Grad-CAM is that it can only visualize the output of one convolution layer at a time. The HAN model calculates attention weights for both sentences and tokens. These attention weights are discrete probably distributions that sum to 1. Term highlights through Grad-CAM and HAN attention weights are shown in our accompanying poster. While we heuristically verified X-AI, future work can use these techniques to empirically identify key n-grams.

Conclusion
The CNN equally or outperformed HAN and bi-LSTM at each sample size with less computation. While HAN models produce visualizable attention weights, Grad-CAM can be used to visualize CNN n-gram importance.

References
Risk Prediction of Black Fungus in COVID-19 Survivors

Eshita Dhar, MA(PhD)¹, Shabbir Syed-Abdul, MD¹, A. Shoban Babu, MS², Raja Shekhar Bellamkonda, PhD³, Ramaiah Itumalla, MD⁴, GVRK Acharyulu, PhD³, Surya Krishnamurthy, BTech⁵, Y. Venkat Santosh Ramana, MSc⁶, Naresh Mogilicharla, MD², Shwetambera Malwade, MSc¹, Yu-Chuan (Jack), MD Li¹

¹Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan; ²Department of ENT, Gandhi Medical College and Hospital, Secunderabad, Telangana, India; ³School of Management Studies, University of Hyderabad, Hyderabad, Telangana, India; ⁴Department of Health Management, College of Public Health and Health Informatics, University of Hail, Hail, Kingdom of Saudi Arabia; ⁵IQGateway, Bengaluru, Karnataka, India; ⁶Department of Hospital Administration, Gandhi Hospital, Secunderabad, Telangana, India.

Introduction

India had around 4300 deaths from COVID-19-related black fungus (also known as mucormycosis) as of July 21, 2021(1). This study aimed to create artificial intelligence based models that can predict the risk of black fungus in COVID-19 patients at the time of discharge from the hospital.

Methods and Evaluation Results

This retrospective study was performed in a tertiary hospital in Telangana, India (Ethical Approval: IEC/GMC/2021). The dataset included 1229 COVID-19 positive patients, aged 30-75 years, treated between March 19 and June 30, 2021, and 214 COVID-19 positive inpatients who were later infected with black fungus (mean number of days between COVID-19 diagnosis and admission for black fungus: 11.1 (boostrapped 95% CI 10.1 – 12.0)). For predictive modeling, 35 of 74 variables were selected after encoding of categorical variables, input to our models and evaluated by 5-fold validation. We used class weights to handle the imbalance in class distribution. We implemented logistic regression (LR), decision tree, and random forest algorithms and extreme gradient boosting techniques (XGBoost). Figure 1A depicts the LR, XGBoost, and random forest all fared similarly well, with Area under receiver operating characteristic curve (AUROC) values of 95.0, 94.0, and 94.0, respectively. XGBoost had best accuracy and precision, with 0.91 ± 0.026 and 0.67 ± 0.0526, respectively. LR was the best model with highest AUROC and recall, 0.95 ± 0.023 and 0.87 ± 0.057 respectively. Figure 1B shows the impact of variables on the risk of black fungus.

Conclusion

This is the only study that we are aware of that has developed an artificial intelligence-based model to predict the risk of black fungus among COVID-19 survivors. Obesity, anosmia, de novo diabetes, myalgia, and nasal discharge were identified as the leading five factors having a significant impact on the likelihood of black fungus. Previously, uncontrolled diabetes, was found to be a substantial predisposing factor for post-COVID black fungus (2).

References

Detecting Dose Changes of Oral Antineoplastic Medications Using an Open Claims Dataset

John E. Doole, Pharm.D., MFA¹, Lydia Gonzalez MS¹, Antonio Solano, MFA¹, Matvey B. Palchuk, MD, MS, FAMIA¹,²
¹TriNetX, LLC, Cambridge, MA; ²Harvard Medical School, Boston, MA

Introduction

Administering oral chemotherapeutic medications exemplifies the balance between outcomes and toxicity management. Dose titration is traditionally employed but applying it to oral agents is more challenging than traditional injectable chemotherapy¹; dose titration may impact efficacy, especially for medications with a narrow therapeutic index. Therefore, it is critical to establish a process to derive these elements from claims data. Accurate dosing history obtained through real-world data (RWD) analysis can provide valuable insight into phase IV side effect profile and outcomes. We present a method to derive and detect dose changes, then compare to published reports of dose reduction in select oral chemo agents.

Methods

We identified a patient cohort (n = 20,136) from paid claims using a TriNetX Network of clearinghouse data (representing 62 million patients) taking at least one of eight commonly dose-reduced oral antineoplastic medications¹. A total daily dose was calculated as Strength * Frequency, where Frequency = Quantity Dispensed/Days' Supply. We extracted drug strength from NDC codes using RxNorm API calls and harvested quantity dispensed and days' supply from NCPDP-complaint data in pharmacy claims². We partitioned the data by patient, medication ingredient, and date of prescription authorization and summed all doses of the same ingredient prescribed on the same day to capture combined product dosing (e.g., Sutent 25 mg + 12.5 mg = 37.5 mg daily dose).

Results

Overall, 30% (5,952) of patients experienced a dose decrease, nearly identical to the median of 33% reported in the literature. However, comparison by ingredient (Figure 1) shows that our dataset had fewer dose reductions on average than literature reports; practice changes and toxicity mitigation by interval adjustment provide a possible explanation.

Conclusion

Claims data offers a relatively approachable way of calculating the total daily medication dose, and we used it to detect and quantify dose reduction for selected oral chemotherapeutic drugs. We have demonstrated a solid alignment to the literature on the patterns of dose reduction and feel that this approach provides a good foundation for persistence and medication profile evaluations. Better availability of detailed medication data from EHR systems is necessary to perform similar analyses on clinical data. Additional characterization of these dose-change patterns is needed to enhance our understanding of the prescribing practices and their relationship to the clinical outcomes.

References

Unexpected somatic mosaicism and significantly reduced penetrance of NF1 pLOF variants in a large medical biobank: evidence for widespread mosaicism of pLOF variants throughout the genome in otherwise-healthy individuals

TG Drivas MD PhD, TT Nomakuchi MD PhD, ZC Bogus MS CGC, M Bonanni CRNP, ARaper MS CGC, SKallish DO, KL Nathanson MD, MD Ritchie PhD

University of Pennsylvania and Children’s Hospital of Philadelphia, Philadelphia, PA, USA

Predicted loss-of-function (pLOF) variants in the NF1 gene cause the 100% penetrant Mendelian disorder Neurofibromatosis type 1 (NF1), thought to affect roughly 1 in 3,000 individuals. We have recently been referred four patients with incidentally discovered pathogenic NF1 pLOF variants, but with no features of the syndrome on exam or history. To understand the true population-level incidence and penetrance of NF1 pLOF variants when ascertained in an unbiased medical biobank population, we turned to the Penn Medicine BioBank which includes EHR and whole exome sequence data from 43,731 adult subjects, all patients of the UPHS healthcare system.

We identified 51 individuals with any of 43 unique NF1 pLOF variants, equating to an incidence of 1 in 857; this rate is more than 3 times higher than the commonly accepted estimates. Only 20 of these 51 individuals carry a diagnosis of NF1 in their EHR, suggesting that ~60% may be undiagnosed. NF1 pLOF variant allele frequencies were significantly lower in the NF1-undiagnosed group, implying that in some undiagnosed individuals these variants might exist in a somatic mosaic state. Subsequent PheWAS analysis across 9,436 binary ICD-code-derived patient phenotypes revealed numerous significant (p < 5.3e-6) associations between NF1 pLOF variants and a number of classic NF1 phenotypes (regression model adjusted for age, sex, and PC1-10). These significant associations included phenotypes such as scoliosis and neoplasams of the peripheral nerves, connective tissue, and spinal cord. However, no significant associations were seen when limiting our analysis to the set of patients with NF1 pLOF variants but without a known NF1 diagnosis, suggesting that the presence of an NF1 pLOF variant in blood does not, per se, increase risk for disease.

Our experience with NF1 led us to examine the incidence of somatic mosaicism on a larger, genome-level scale. Within PMBB, there is clear evidence that nearly all individuals harbor multiple likely somatic-mosaic variants in various genes, with certain genes being significantly enriched for somatic mosaic variants, at least in peripheral blood. The genes we found to be enriched for somatic mosaicism are, in general, associated with cell proliferation/survival, and include the set of genes known to be associated with clonal hematopoiesis of indeterminate potential. This identification of widespread somatic mosaicism has major implications for future genetic testing and biobanking efforts, and the further investigation of this finding will be critical for accurate counseling of patients and families.

Learning Objective: To identify the under-recognized high frequency and reduced penetrance of Mendelian disease gene pathogenic variants in large medical biobanks, and to appreciate the implications of these finding for future genomic studies.
EHR-based Detection of Youth with Suicide-related Emergency Department Visits:
Variation in Structured Data Indicators

Juliet Edgcomb, MD PhD¹, Chi-hong Tseng, PhD², Neal Ryan, MD³, Bonnie Zima, MD MPH¹

Affiliations: ¹UCLA-Semel Institute for Neuroscience and Human Behavior, Department of Psychiatry and Biobehavioral Sciences, University of California at Los Angeles, Los Angeles, CA; ²Department of Medicine Statistics Core, University of California at Los Angeles, Los Angeles, CA; ³Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA

Suicide is the second leading cause of death of people 10-34 years old in the United States, with higher rates of morbidity and mortality than many childhood diseases¹. More than 1.12 million annual suicide-related ED visits for 5-18-year-olds occur in the U.S.² Children experience developmentally varied pathways to suicidal behavior and are priority population to target for prevention of suicidal behavior throughout the lifespan. Electronic health record (EHR)-based risk prediction algorithms may be a promising means to identify precursors of suicidal behavior among youth³-⁴. Predictive models frequently rely on structured data (e.g., billing codes) to screen EHR datasets for cases of suicidal ideation and behavior. However, structured data likely underestimates the true prevalence⁵-⁶ of suicide-related events. We sought to examine the variation among common structured data indicators of suicide-related ED visits by children and adolescents, and measure variation by demographics and legal status.

Using EHRs from a large academic health system, eligibility criteria were: ages 10-17-years-old, ≥1 ED visit associated with ≥1 mental health (MH)-related International Classification of Disease Version 10 (ICD-10) code⁷, chief complaint of “suicid***”, involuntary hold order, or Columbia Suicide Severity Rating Scale (c-SSRS)⁸ score ≥1. Visits were restricted to the most recent visit by each child with complete data, occurring 10/1/2015-10/1/2019 to correspond to adoption of ICD-10. Each MH-ED visit was screened for Child and Adolescent Mental Health Disorders Classification System (CAMHD-CS) ICD-10 subgroup for suicide or self-injury, chief complaint of “suicid***”, and c-SSRS ≥1. We compared the prevalence of MH-ED visits with and without these indicators of suicidality, stratified by demographics and legal status. The study was exempt by the University of California, Los Angeles Institutional Review Board.

We identified 1,699 MH-ED visits by unique children, of which 56.4% (n=959) visits had no structured indicators of suicidality (Table 1). Among visits with a single indicator of suicidality (n=740; 43.6%), visits with c-SSRS ≥1 (n=170; 23.0%) exceeded visits with a suicide-related ICD-10 code (n=52; 7.0%). Of the 526 children a suicide-related ICD code, 85.7% (n=451) had a c-SSRS ≥ 1, 53.0% (n=279) had a corresponding chief complaint, and only 48.7% (n=256) had all three indicators. Compared with Hispanic youth (n=183, 39.1%), non-Hispanic youth (n=554, 45.5%) were more likely to have ≥1 suicide-related indicator (p=0.018). Among children with ≥1 indicator, males (p=0.0001), preteens (p=0.016), black youth (p=0.02), and voluntary patients (p=0.04), had significantly fewer indicators of suicidality compared with females, adolescents, white youth, and involuntary patients, respectively.

In this single site analysis of structured EHR data, we discovered substantial variation in documentation of common indicators of suicide-related ED visits by children and adolescents. Moreover, we identified heterogeneity in completeness of documentation of these indicators by sex, age group, race, and legal status. Results suggest that careful examination of the types of structured data used to detect cases of pediatric suicidality is prudent to safeguard against perpetuating inequities in identification and prevention, and development of a comprehensive computable phenotype may help to improve identification of children at risk.

<table>
<thead>
<tr>
<th>N (%) visits</th>
<th>Total</th>
<th>None</th>
<th>ICD-10 only</th>
<th>cSSRS ≥1 only</th>
<th>ICD-10 and cSSRS ≥1</th>
<th>ICD-10, chief complaint, and cSSRS ≥1</th>
<th>χ²</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>809</td>
<td>499</td>
<td>17 (2.1)</td>
<td>103 (12.7)</td>
<td>74 (9.1)</td>
<td>93 (11.5)</td>
<td>30.2</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>F</td>
<td>889</td>
<td>460</td>
<td>35 (3.9)</td>
<td>67 (7.5)</td>
<td>121 (13.6)</td>
<td>162 (18.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 to 12</td>
<td>335</td>
<td>199</td>
<td>8 (2.4)</td>
<td>45 (13.4)</td>
<td>29 (8.7)</td>
<td>41 (12.2)</td>
<td>10.32</td>
<td>0.016</td>
</tr>
<tr>
<td>13 to 17</td>
<td>1364</td>
<td>760</td>
<td>44 (3.2)</td>
<td>125 (9.2)</td>
<td>166 (12.2)</td>
<td>215 (15.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>1086</td>
<td>588</td>
<td>31 (2.9)</td>
<td>114 (10.5)</td>
<td>145 (13.4)</td>
<td>170 (15.7)</td>
<td>8.99</td>
<td>0.029</td>
</tr>
<tr>
<td>Black</td>
<td>165</td>
<td>95</td>
<td>10 (6.1)</td>
<td>15 (9.1)</td>
<td>12 (7.3)</td>
<td>25 (15.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>436</td>
<td>264</td>
<td>12 (2.7)</td>
<td>41 (9.4)</td>
<td>37 (8.4)</td>
<td>59 (13.5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>468</td>
<td>285</td>
<td>8 (1.7)</td>
<td>45 (9.6)</td>
<td>46 (9.8)</td>
<td>59 (12.6)</td>
<td>2.74</td>
<td>0.43</td>
</tr>
<tr>
<td>Non-Hispanic</td>
<td>1218</td>
<td>664</td>
<td>44 (3.6)</td>
<td>125 (10.3)</td>
<td>148 (12.2)</td>
<td>195 (16.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Legal Status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voluntary</td>
<td>1381</td>
<td>898</td>
<td>35 (2.5)</td>
<td>120 (8.6)</td>
<td>130 (9.4)</td>
<td>148 (10.7)</td>
<td>8.31</td>
<td>0.04</td>
</tr>
<tr>
<td>Involuntary</td>
<td>318</td>
<td>61</td>
<td>17 (5.3)</td>
<td>50 (15.7)</td>
<td>65 (20.4)</td>
<td>108 (34.0)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Structured electronic health record indicators of suicidality among children and adolescents with MH-related emergency department visits, and variation by patient demographics and legal status. Shaded cells indicate comparison groups for chi-square test (df=3). Comparisons of CC are omitted from test of significance due to low sample size. Abbreviations: ICD-10: International Classification of Disease, Version 10; c-SSRS: Columbia Suicide Severity Rating Scale; CC: Chief Complaint of “Suicide***”.

Economic Analysis of Single Institutional Review Board Data Exchange Standards in Multi-Center Clinical Studies

Eric L. Eisenstein, DBA1, Katrina Donovan, MA2, Meredith N. Zozus, PhD3, Anita Walden, MS4, Feliciano B. Yu, Jr., MD2, Vivian L. West, PhD1, W. Ed Hammond, PhD1, Lawrence H. Muhlbaier, PhD1

1Duke University, Durham, NC; 2University of Arkansas for Medical Sciences, Little Rock, AR; 3University of Texas Health Science Center at San Antonio, San Antonio, TX; 4Oregon Health & Science University, Portland, OR

Introduction

National Institutes of Health (NIH) policy and the Common Rule require that a single institutional review board (sIRB) review all NIH-funded, non-exempt multicenter domestic human subject research1. Reviewing IRBs, coordinating centers, and relying sites currently exchange sIRB documents manually. HL7 FHIR sIRB standards for data exchange were balloted in September 20212. We evaluated the potential economic impact of those standards.

Methods

We compared sIRBs operating with and without the proposed sIRB data exchange standards. Four workflows were included: (1) Initial Study Protocol Application, (2) Site Addition and Approval, (3) Continuing Review and Approval, and (4) Reportable Events. The study team documented current Duke University Medical Center sIRB workflows, collected labor hours and costs, and determined which tasks would be eliminated with the proposed sIRB data exchange standards. We assumed IRB software vendors would update their systems to include the HL7 FHIR sIRB standard. This study was funded by the National Center for Advancing Translational Sciences (UL1TR002553) and approved by the Duke University Medical Center Institutional Review Board (Pro00101455).

Results

Implementing the proposed sIRB data exchange standards did not eliminate tasks in the Initial Study Protocol Application workflow (executed once per study) or the Reportable Events workflow (executed once per event). However, tasks would be eliminated in the Site Addition and Approval workflow (executed once per study site) and the Continuing Review and Approval workflow (executed once per site per study year). Site Addition and Approval total workflow hours were reduced by 2.5 hours (from 15.5 to 13.0 hours). Similarly, Continuing Review and Approval total workflow hours were reduced by 9.0 hours (from 36.5 to 27.5 hours). sIRB data exchange standard associated cost savings were $251 per site addition (costs reduced from $1609 to $1358) and $1033 per continuing review (costs reduced from $4110 to $3077). Relying site cost savings were $101 per site addition and $764 per continuing review. In a 50-site clinical trial with sites active for 5 years, Total sIRB cost savings with the proposed data exchange standards would be $270,800 ($12,550 for site additions and $258,250 for continuing reviews).

Conclusion

Data exchange standards can significantly reduce sIRB-associated labor hours and costs.

References

Achieving Interoperability for Electronic Access of POLST Information

Kate Fultz Hollis, MS, MBI1, William R. Hersh, MD1, Abby Dotson, PhD2
Oregon Health & Science University, 1Department of Medical Informatics and Clinical Epidemiology, 2Department of Emergency Medicine, Portland, OR, USA

Abstract

The ability to quickly obtain electronic Portable Orders for Life-Sustaining Treatment (POLST) documents is particularly needed in hospital emergency rooms and in pre-hospital emergent situations. The Oregon Health & Science University Department of Emergency Medicine and the Department of Medical Informatics and Clinical Epidemiology are working together to improve interoperability between the POLST Registry and electronic health record systems. We describe work to discover new approaches to access registry information in the electronic health record.

Introduction

The Portable Orders for Life-Sustaining Treatment (POLST) document includes patient wishes for end-of-life care and treatment, documentation of medical orders, and the commitment by a health care professional to honor these wishes. The Oregon POLST Registry (OPR) project began in response to a need expressed by Emergency Medical Services (EMS) to access POLST orders when they arrived on the scene of a medical emergency and could not immediately locate the original POLST form. A specific need addressed by EMS was for interoperability of POLST forms between electronic health record (EHR) and registry systems. The Oregon Health & Science University (OHSU) Department of Emergency Medicine and Department of Medical Informatics and Clinical Epidemiology (DMICE) have been working together to study and improve interoperability between the POLST Registry and EHR systems. The goal of our study was to discover how organizations might design, develop, and implement POLST interoperability.

Methods

Built in 2008, the Oregon POLST Registry was rebuilt in 2020 as a document-oriented database server with a web-based front-end. The platform recently developed the capability for automated POLST form entry into the Registry yet interoperability with healthcare system EHRs is limited. How might forms on the registry be available in the EHR and vice versa? In January 2021, we first undertook literature searching for studies about interoperability of registry forms between EHRs and registry entities. We discovered information about POLST form interoperability in studies1,2 but these studies did not provide insight about interoperability of the data entry forms.

Next, we developed surveys using Qualtrics® for two different professional groups, emergency medicine and informatics professionals. The surveys included quantitative data (surveys will be open until all interviews and recommendations are made) and there are additional in-person questions to supplement the surveys. A sample electronic survey question is pictured in Figure 1. The poster will include survey results, the study methods diagram, and how in person interviews contributed to assessment of the POLST interoperability environment and institutional readiness for electronic POLST records.

Results and Conclusion

POLST accessibility and interoperability needs differ greatly depending on the provider, their workflow and healthcare facility resources. Small clinics, home health, hospice providers, and assisted living facilities seem to prefer a stand-alone web portal to access POLST information without the need for additional IT effort and costs. For large healthcare systems, it is recommended to have seamless integration with the EHR and “one click” access to a patient’s POLST form. The interoperable design should be compatible with the upcoming development of a POLST HL7 C-CDA and FHIR based API. We found strong interest by EMS personnel and first responders for electronic POLST information and the requirement for multiple POLST access approaches to cover all EMS needs in the field. There was also interest in sending POLST documents quickly to a state registry and little information available in journals on interoperability of registry documents in the EHR.

References

Evaluating Comprehensiveness of the HL7® FHIR® Standard in Supporting Data Exchange in Clinical Research: A Pediatric Trial Network Use Case

Maryam Y. Garza, PhD, MPH, MMCi1, Michael W. Rutherford, MS1, Bhargav S. Adagarla, MS2, Eric L. Eisenstein, DBA2, Karan R. Kumar, MD2, Kanecia O. Zimmerman, MD, MPH2, Umit Topaloglu, PhD3, Zhan Wang, PhD4, Meredith N. Zozus, PhD, CCDM4

1University of Arkansas for Medical Sciences, Little Rock, AR; 2Duke Clinical Research Institute, Durham, NC; 3Wake Forest University School of Medicine, Winston-Salem, NC; 4University of Texas Health Science Center at San Antonio, San Antonio, TX

Introduction

For data standards to be useful, they must support their intended use cases. Historically, clinical research studies have relied on manual approaches to data collection, such as medical record abstraction (MRA). However, MRA often increases the cognitive burden on data collection personnel, and results in research data with high discrepancy rates.1,2 Use of the HL7® FHIR® standard3,4 has the potential to reduce site burden and data discrepancies by partially automating the data collection process by automating the direct exchange of information from EHRs to EDC systems. We aim to quantify the current maximum potential coverage of the HL7® FHIR® standard in supporting data collection for clinical research studies. Limited work has been done in this area,5-7 and even less so for trials targeting pediatric populations. Therefore, we focused on pediatric studies for our use case.

Methods

A systematic mapping4 of study data elements for three federally sponsored pediatric studies was used to assess the HL7® FHIR® standard’s coverage as implemented in four academic medical centers. Review of study data elements was conducted to identify (1) data elements available within the site’s EHR and (2) data elements available within the EHR that were also available via FHIR®. An initial, independent review of mappings for the three pediatric studies evaluated mapping consistency across studies. A comparison was also performed across the four sites to identify similarities and differences in data availability, and to determine study mapping “shareability” across sites. This work will influence the direction of standards development to better support data collection for multicenter pediatric studies.

Results

Across all three studies, 1,116 total data elements were mapped. Preliminary results indicate that, on average, nearly half (47%) of the full set of data elements were available within site EHRs. Of those data elements available within the EHR (n = 527), on average, 61% (n = 320) were available via FHIR®; indicating that there is the potential for pediatric research study sites to electronically abstract over half of the EHR-based data elements using FHIR®.

Funding

This work was supported in part by the National Institute of Child Health and Human Development (NICHD) contract HHSN275201800003I and by the National Center for Advancing Translational Sciences (NCATS) grant UL1TR003107 of the National Institutes of Health (NIH). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Risk Factor Assessment and Control for Patients with Diabetes and the Impact of Telemedicine Use

Anne Grauer, MD,1 Andrea T. Duran PhD,1 Nadia A. Liyanage-Don MD,MS,1 Lucille M. Torres-Deas MD,1 Gil Metser, MD,1 Nathalie Moise, MD, MS,1 Ian M. Kronish, MD, MPH,1 Siqin Ye, MD, MS,2
1. Department of Medicine, Columbia University Irving Medical Center, New York, NY, USA
2. Department of Medicine, Cardiology, Columbia University Irving Medical Center, New York, NY, USA
3. Center for Behavioral Cardiovascular Health, Columbia University Irving Medical Center, New York, NY, USA

Background: Assessment and control of cardiovascular risk factors is essential to reduce complications of diabetes1. Increased telemedicine use may negatively impact risk factor management for patients with diabetes2. We examined whether telemedicine use in primary care affected cardiovascular risk factor assessment and control for diabetes patients.

Methods: This was a retrospective cohort study conducted between February 2020 and December 2020 in a large primary care network at a large urban academic medical center in New York City. Our participants were patients ages 18 to 75 with at least one primary care visit and an ICD-10 diagnosis of diabetes during the study period. Our exposure was categories of telemedicine use, defined as in-person visits only (Group 1), only 1 telemedicine visit (Group 2), and 2 or more telemedicine visits (Group 3). As an outcome we determined whether hemoglobin A1c (HbA1c), blood pressure (BP) and low-density lipoprotein cholesterol (LDL-C) were assessed for each patient during the study period. For each risk factor, we determined whether the risk factor was controlled during the study period when they were assessed (i.e., last HbA1c level <8.0%, BP <140/90 mmHg, LDL-C <100 mg/dL). For each risk factor (HbA1c, BP, and LDL-C), we first fitted a logistic regression model on the entire sample of patients to determine the association between categories of telemedicine use and whether or not the risk factor was assessed.

Results: Group 1 had a total of 2,325 visits (all in-person). Group 2 had a total of 1,407/3,098 (45.4%) telemedicine visits and group 2 had a total of 7,382/10,481 (70.4%) telemedicine visits. For 4,932 patients with diabetes, telemedicine use was associated with a lower proportion of patients with all three risk factors assessed (544/1,296 [42%], 362/1,407 [26%], and 496/2,229 [22%], for Groups 1, 2, and 3 respectively; p<0.001). In a multivariable model adjusting for clinical and demographic characteristics, telemedicine use was also associated with lower odds of risk factor assessment for individual risk factors. When compared with Group 1, the odds ratios (OR) for HbA1c assessment were 0.35 (95% confident interval [CI] 0.29 to 0.42, p<0.001) for Group 2 and 0.13 (95% CI 0.11 to 0.16, p<0.001) for Group 3. Similarly, the ORs for BP assessment was 0.02 (95% CI 0.01 to 0.32, p<0.001) for Group 2 and 0.0006 (95% CI 0.0003 to 0.001, p<0.001) for Group 3, and the ORs for LDL-C assessment was 0.36 (95% CI 0.30 to 0.43, p<0.001) for Group 2 and 0.17 (95% CI 0.14 to 0.21, p<0.001) for Group 3, all when compared to Group 1. However, when assessed, telemedicine use was not significantly associated with risk factor control (HbA1c <8%: odds ratio [OR] 1.02, 95% CI 0.78 to 1.33, p=0.88 for Group 2 and OR 0.94, 95% CI 0.70 to 1.27, p=0.68 for Group 3; BP <130/80 mmHg: OR 1.12, 95% CI 0.92 to 1.36, p=0.26 for Group 2 and OR 1.07, 95% CI 0.85 to 1.33, p=0.58 for Group 3; LDL-C <100 mg/dL: OR 1.07, 95% CI 0.78 to 1.47, p=0.66 for Group 2 and OR 0.89, 95% CI 0.62 to 1.28, p=0.53 for Group 3; reference is Group 1 for all models).

Conclusions: Telemedicine use was associated with gaps in risk factor assessment for patients with diabetes, but had limited impact on whether risk factors were controlled. Effective strategies are needed to optimize diabetes risk factor management in the telemedicine era.

Supporting the Research Enterprise with Geocoding as Infrastructure

Winston L. Guo, BS1, Evan T. Sholle, MS2,3, Xiaobo Fuld, MS1, Sean T. Pompea, MFA3, Ryan McGregor, MS3, Thomas R. Campion, Jr., PhD2,3,4
1Weill Cornell Medical College, Weill Cornell Medicine, New York, NY; 2Department of Population Health Sciences, Weill Cornell Medicine, New York, NY; 3Information Technologies & Services Department, Weill Cornell Medicine, New York, NY; 4Department of Pediatrics, Weill Cornell Medicine, New York, NY

Introduction

Social determinants of health (SDoH) profoundly impact patient well-being. In the hospital, SDoH data can help providers understand a patient’s situation and arrive at diagnoses that appreciate the root causes of that patient’s health issues1. However, this data is rarely available to clinicians at the point of care, due to factors such as gaps in history-taking and an electronic health records (EHR) system not configured to capture SDoH data1. Although efforts to capture individual-level SDoH are promising2, it is currently possible to supplement the EHR with neighborhood-level SDoH data, which has already been used to study disparities in contexts such as hospital outcomes3,4. We hypothesized that by building a geocoding process relying on the US Census application programming interface (API), incorporating SDoH data into our research IT infrastructure, and comparing it against a prominent geocoding method (DeGAUSS)5, we could better equip investigators to conduct novel investigations involving SDoH.

Methods

We built a pipeline that, using SQL and Python scripts, continually integrates geocoding data from the US Census API with patient data from our research data warehouse at Weill Cornell Medicine (WCM). Our process involves the following steps. First, we run a SQL script that extracts address data from our local copy of the Clarity data model and removes unnecessary ancillary elements such as apartment or floor numbers. A second script conducts the actual geocoding process by connecting to our database infrastructure, passing the processed address data to the Census API, and re-integrating the data from the API into our SQL environment. Finally, we compared runtime and map rate between our method and DeGAUSS on a sample of 11,762 patients recently admitted with COVID.

Results

Our method geocoded the 11,762 patients as part of a larger batched API call for 301,013 patients, which completed in a total of 1324 minutes and successfully geocoded 86.4% of records. DeGAUSS geocoded the 11,762 patients in 1 minute 23 seconds on a personal laptop, and successfully geocoded 92.6% of records. In addition to making geocoded patient data available for research, we have connected this data to the FACETS data set4 and made it available to WCM researchers to conduct downstream analyses.

Conclusion

Geocoding is currently feasible for collecting neighborhood level SDoH data, allowing researchers to conduct new analyses. Areas of refinement include data granularity and consistency; future studies should seek to obtain individual level SDoH data and map geographic data to as many patients as possible. DeGAUSS serves as a promising method for conducting geocoding at scale without passing any data outside the institution5.

References

Phenotypic analysis within the socio-economic exposome

Christopher Hawthorne, MSc\textsuperscript{1}, Guillermo H. Lopez-Campos, PhD\textsuperscript{1}
\textsuperscript{1}Wellcome-Wolfson Institute for Experimental Medicine, Belfast, Northern Ireland, United Kingdom

Introduction

The socio-economic exposures which constitute the exposome (all exposures experienced during our lifetime)\textsuperscript{1} have an influence on human phenotype development. To our knowledge, informatically the relationship between socio-economic exposures and phenotypes is less well explored in the exposome field compared to other exposure data such as chemical compounds. To counteract this, we are developing a knowledgebase that utilizes text-mining and annotation resources to identify the co-occurring relationship between phenotypes, chemicals or diseases associated with a socio-economic exposure using an example showcase of “Income” and “Educational Status”.

Methods

A literature corpus derived from PubMed utilizing the R package RISMed, was created using the query “((x[MeSH] AND x[tiab]) NOT(SARS-CoV-2[All Fields] OR coronavirus[MeSH Terms] OR coronavirus[All Fields])) AND 1950:2019(dp)” where x represents either “Income” or “Educational status”. This corpus was then annotated for disease and chemicals using PubTator (2020-02-15) and phenotypes (Human Phenotype Ontology 2019-02-12) using the R package Onassis (using ConceptMapper). Co-occurrence analysis between chemicals, phenotypes and diseases was calculated through shared PubMedID.

Results and Discussion

For the socio-economic factor Income, 15305 abstracts were retrieved, 546 phenotypes, 506 chemicals and 1335 diseases were identified. Co-occurrences consisted of 882 phenotype and chemical pairs, 2108 chemical and disease pairs and 3690 chemical, phenotype and disease groupings. For Educational Status, 2725 abstracts, 260 phenotypes, 149 chemicals and 576 diseases were found. Co-occurrences for this socio-economic factor had 258 phenotype and chemicals, 486 chemical and diseases and 1566 chemical, phenotype and disease groupings.

Results of top 5 phenotype and chemical pairs for Income and Educational status are shown in Table 1.

![Table 1. Top 5 phenotype and chemical pairs for Income and Educational Status.](image)

Phenotype and chemical pairs presented known relationships between these entities highlighting relationships associated to the selected socio-economic factor. Query was constructed to not return coronavirus associated literature as the pandemic has caused an additional layer of effect and therefore we have decided not to include them.

Conclusion

In this work, we have presented the integration of information from the socio-economic literature to identify relationships of chemical compound and human phenotypes found within the socio-economic exposome.

References

Informatic analysis of electromagnetic fields for the physical exposome

Christopher Hawthorne, MSc¹, Guillermo H. Lopez-Campos, PhD¹
¹Wellcome-Wolfson Institute for Experimental Medicine, Belfast, Northern Ireland, United Kingdom

Introduction
The exposome is the totality of exposures an individual comes into contact with during their lifetime¹. Physical factors including radiation, electromagnetic fields (EMFs) and noise have an influence in human health and disease however informatically they are not as well explored or have comprehensive resources compared to other aspects such as the chemical component of the exposome which has been informatically boosted by a variety of chemical databases such as CTDBase, PubChem and T3DB. To advance physical factor exposome research, we developed a new methodology utilizing annotation methods, text-mining, enrichment analysis and network visualization using EMFs as a showcase.

Methods
EMFs abstracts were retrieved using the R package rentrez and an edited pubmedXML R script from PubMed using the query 'Electromagnetic Fields'[MeSH] AND 1950:2019[EDAT] AND ('gene'[TiAB] OR 'protein'[TiAB]) AND english[LA]. Abstracts were annotated using PubTator Gene dataset annotations (2020-02-15) and the R package Onassis with the Human Phenotype Ontology (2019-02-12), Onassis utilizes ConceptMapper to annotate. A gene list was created where gene annotations had to be present in at least 3 abstracts. Phenotype enrichment were calculated with the gene list using the phexpo methodology², a threshold of an adj. p-value of 0.05 and visualized through ggplot2. Visualization of networks for phenotype co-occurrence by PubMed ID was provided by the igraph R package.

Results and Discussion
Figure 1 presents the results from the two analytical approaches. 1) Phenotype prediction results, where the overlap is the intersection between those identified in the text and those derived from gene set enrichment, and 2) network visualization of phenotype co-occurrence (phenotypes are both present in the same abstract) with weights showing the number of times they appear together (Fig 1.C) and phenotype groupings through their connections (clustering using igraph’s walktrap function). We identified 89 potential phenotypes in the literature and predicted 79 based on the constructed gene list. These predicted phenotypes include hyperuricemia, skin nodule, abnormality of the lymphatic system and increased susceptibility to fractures. These predictions not previously identified in the literature are supported by EMF’s therapeutic and inducing effects.

Conclusion
This new novel methodology combines various informatic techniques to advance the study of physical factor exposures in exposome informatics research through its relationship building and hypothesis generating capabilities.

References
Design Principles Towards Higher Efficiency in Multi-document Annotation

Huan He, PhD, Sunyang Fu, PhD, Liwei Wang, MD, PhD, Andrew Wen, MS, Sijia Liu, PhD, and Hongfang Liu, PhD
Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, USA

Introduction: Building a high-quality annotation dataset requires considerable time and expertise efforts to go through the whole annotation process with annotation tools. Although existing text annotation tools may provide many features to meet different needs in annotation, limited attention is paid to how to improve the annotation efficiency and user experience. In response, we propose applying human-computer interaction (HCI) design principles to guide the design of annotation tools. Moreover, based on the design principles and the feedback summarized from our experienced domain experts, we developed MedTator, a lightweight, open-source, web-based application for multi-document annotation (Live demo and source code are available at: https://github.com/OHNLP/MedTator).

Design Principles: Since there have been several comprehensive annotation tools used in our past projects, we interviewed our domain experts who have extensive annotation experiences to get their comments on existing tools and annotation workflow. Based on their comments and task requirements, we summarized the design principles to guide our development, including: 1) reducing the cognitive overhead. To enable users to focus on the actual annotation tasks rather than other distracting tasks, we only selected those core features related to corpus annotation to be included in MedTator. And we customized visual designs to show useful information for the annotation task; all necessary functions are organized on single page with the same design language; the statistical results of the annotated tags are updated in real time to help users track annotation progress (Fig. 1(a1)); 2) low physical effort. To minimize the number of steps in each task, we studied the existing workflow and optimized the interactive process for each task. For example, to start a new annotation task from beginning, users only need to drag and drop the schema file and raw text files without any further operation; the annotation hints are optionally shown based on the completed annotations to reduce repetitive searching (Fig. 1(a2)); and adding all annotation hints as new tags only requires one click to confirm the suggestions (Fig. 1(B)); 3) flexibility in use. To address users’ requirements for different workspace preferences, we use this principle to guide the tool design to balance the needs of tool features and the interface complexity. For example, we designed a variety of mode switches and organized them in a unified display layout for changing the visual effects.

Future work: Although our tool is still in development stage, we demonstrated our visual and interactive designs to domain experts and got positive feedbacks. We are going to improve its flexibility to meet the needs for downstream tasks and conduct a formal evaluation to improve the usability.

References
Estimating Inspiratory Capacity from Spirometry

Scott A. Helgeson, MD1; Zach S. Quicksall2; Patrick W. Johnson2; Kaiser G. Lim, MD3; Rodrigo Cartin-Ceba, MD4; Daniel D. Poliszuk5; Augustine S. Lee, MD1; Alexander S. Niven, MD3

1Division of Pulmonary and Critical Care Medicine, Mayo Clinic, Jacksonville, Florida, USA; 2Department of Quantitative Health Sciences, Mayo Clinic, Jacksonville, Florida, USA; 3Division of Pulmonary and Critical Care Medicine, Mayo Clinic, Rochester, MN, USA; 4Division of Pulmonary and Critical Care Medicine, Mayo Clinic, Scottsdale, AZ, USA; 5Mayo Clinic Information Technology, Mayo Clinic, Rochester, MN, USA

Introduction: Inspiratory capacity (IC) is the maximal amount of air that can be inspired into the lungs following a normal exhalation. IC is a surrogate for lung hyperinflation and is mainly seen as elevated in obstructive lung diseases such as chronic obstructive pulmonary disease (COPD). Respiratory and all-cause mortality, along with risk of COPD exacerbation, has been predicted using the IC to total lung capacity ratio (1). Currently, the only way to determine IC is to perform a lung volume test—which many physician offices do not have the proper equipment to perform. To address this shortcoming, we evaluated the ability of spirometry to predict IC.

Methods: This was a retrospective multi-site study utilizing complete pulmonary function tests to determine IC. A total of 37,871 baseline tests were divided into training and testing data sets using a 70/30 split stratified on IC. Features (e.g., length, maximum) were extracted from the spirometry timeseries curves and tabulated for use in predicting IC. Penalized linear regression models were tuned using 5-fold cross validation with the final model fit to the full training set. Model performance was evaluated on the test data set. This study was IRB approved.

Results: The final linear model was able to accurately predict IC across all patients with a root mean square error (RMSE) of 0.33 (Figure 1). When stratified by ATS pulmonary function classification (normal, obstructed, restricted, and mixed patterns), the IC was predicted with comparable RMSEs of 0.346, 0.320, 0.288, and 0.283, respectively (Figure 2). Top features, as defined by ranked model coefficients, included series length, mean value, patient weight, and maximum value.

Conclusion: Utilizing only spirometry-based measurements and easily obtained demographic information, the IC was able to be accurately predicted. Of note, accuracy was maintained in patients with normal pulmonary function as well as those with restrictive or mixed defects. This model enables anyone with the ability to perform spirometry to accurately predict IC and provide the best prognostic data to COPD patients.

References:
Putting NCCN Guidelines on the Map: Using CMAP as a Visual Aid to Treat Non-Small Cell Lung Cancer

Sean Hernandez, MD\(^1\), Isabella Ruiz\(^1\), Meijian Guan, PhD\(^2\), Ashi Awasthi\(^1\), Umit Topaloglu, PhD\(^1\)
\(^1\)Wake Forest School of Medicine, Winston-Salem, NC, \(^2\)Janssen of Johnson & Johnson, Blue Bell, PA

**Background:** The National Comprehensive Cancer Network (NCCN) has been providing evidence based guidelines on treatment options for 97% of cancers\(^1\). A recent study on Non-Small Cell Lung Cancer (NSCLC) concluded that standard of care treatments that are concordant with the NCCN guidelines cost $7,000 less than discordant care annually\(^2\). Although the NCCN guidelines are considered to be a gold standard and may reduce financial burden for patients, one survey of 202 oncologists in 2016 found that only 31% followed the NCCN guidelines to direct care for their patients\(^3\). Further explanation was not documented by the survey. However, the difficulty of remaining current with the consistently updating guidelines is a known barrier to care\(^4\). EHR integration of the NCCN guidelines is the ultimate goal given potential for point of care recommendations. Our stepwise approach includes CMAP representation of NCCN guidelines with the intention of eventually integrating with the EHR.

**Methods:** We used the CmapTools program Version 6.04 to depict the NCCN guidelines version 4.2021 from May 3, 2021. Each decision/branching point in the NCCN guidelines was graphically represented in the CMAP. First line therapy was represented as well as subsequent treatment for progression of disease.

**Results:** We were able to represent the entirety of the NCCN guidelines for metastatic or advanced Non-Small Cell Lung Cancer requiring molecular testing. Figure 1 shows the entire map while Figure 2 represents the central branching point based on genomic marker presence. Figure 3 demonstrates the first line therapies for EGFR and the branching decision points.

**Discussion:** The concept graph representation could serve as a tool for oncologists to make clinical decisions using previous treatment history and genomic sequencing data. Furthermore, we are in the process of linking the CMAPs to our EHR to further leverage the EHR to promote the NCCN guidelines in daily practice.

**References**

Experiential Learning and Enterprise Data Warehouses for Research with the Iowa Health Data Resource

Ashlee Hoberg, BS¹, Donna Santillan, PhD², Kirk T. Phillips, PhD¹,
Boyd Knosp, MS, FAMIA¹, Heather Davis, MS, MLIS¹

¹Institute for Clinical & Translational Science, University of Iowa, Iowa City, IA; ²Department of Obstetrics and Gynecology, University of Iowa, Iowa City, IA

Introduction

Efficiently extracting and analyzing patient data from disparate sources is critical for advancing research across the spectrum of healthcare. Access to those data is limited due to a less than optimal infrastructure to gather, collate, fully protect, de-identify, and disseminate these data in ideal formats that are customized to the research question. A key goal of the Iowa Health Data Resource (the IHDR) is to address this need by significantly improving how the UI research community accesses and utilizes health science data which includes electronic health records (EHRs), payer claims, clinical trials, genomic analyses, and environmental sources¹. These resources are managed in our enterprise data warehouse for research (EDW4R) by the Institute of Clinical and Translational Science Biomedical Informatics Core (ICTS BMI). Our objective is to create an Intercollegiate Access and Implementation (IAI) team of faculty and staff to plan and further develop our EDW4R

Methods

Members of the IAI team started with experiential training by serving as a member of the EDW4R team for 2 months. This training was offered to one key faculty and one key staff who are partially funded by the IHDR (IAI team). A rotating schedule, called an IHDR Shift, was assigned to individuals to work part-time as an EDW4R team member. We provided each IAI team member with a list of prerequisites consisting of on-line training and certifications, for example HIPAA, CITI IRB-01 and information security awareness to complete prior to their shift. During their shift, each IAI team member participates in data request triage, EDW4R data request meetings and in building upon established competencies. These competencies include the ability to: 1) demonstrate and train other faculty and staff to use data self-service tool; 2) collaborate with faculty in order to create and submit a data request to the data team to evaluate; 3) gather specifications and details from the research team and translate those into a data list for the data team; 4) increase data literacy through knowledge of available tools and existing mapped data sources; and, 5) serve as a liaison to their home college or department. In addition, they will complete an evaluation of our EDW4R request processes and the experiential learning methods.

Results

IAI faculty and staff members applied their domain specific knowledge to the data request process by collaborating with faculty investigators to enhance data requests. Having a team that is facile at translating conceptual clinical questions into specific data elements for the EDW4R team to extract promotes data literacy in the home colleges and departments and decreases the time from data request to data delivery. Feedback from these early IAI team members has generated improvements in both our training plans and in our data request service design. As the key staff complete their experiential learning they implement the new data request process for their college and serve as the initial point of contact for investigators considering data requests. As the key faculty complete their experiential learning they serve as both domain experts for their college and as research mentors to faculty new to data requests/data science.

To date, eight key personnel from the five health science colleges at University of Iowa have completed approximately forty hours of experiential learning over an eight-week period. Three colleges have gone live with the new process.

Discussion

The workflow process for data requests will be further streamlined so that research questions from different domains can be extrapolated and turned into usable datasets from which the EDW4R team can extract data. In this work, we apply a novel approach to expand our team with IAI team members assisting to further refine our processes. The response to training has been very positive from both the IAI team members and the ICTS BMI team. Next steps will include; analyzing feedback from the learners, further process evaluation, and comparison of service usage metrics pre and post experiential learning process implementation.

References


688
Moffitt Cancer Analytics Platform (MCAP): A Cloud-Based Data Platform for Improved Data Curation, Integration, and Access
Rachel Howard, PhD., Jordan Creed, Phillip Reisman, Patricia Lewis, Chandan Challa, Kedar Kulkarni, PhD., Dana Rollison, PhD.
Office of Health Data Services and Information Technology
H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL

Introduction: Rapid growth in the production and collection of medical data over the last decade has driven healthcare to the forefront of the big data industry. This vast data resource holds great promise for accelerating novel research, discovery, and clinical translation, but also necessitates creative solutions for data harmonization, storage, and management. The cancer data ecosystem poses unique challenges; clinical data routinely collected in the electronic medical record must be linked to demographic, molecular, and patient-reported data, in addition to unstructured data assets associated with imaging, pathology, and clinical notes. Furthermore, individual-, tumor- and biospecimen-level data must be integrated to provide a complete view of the patient and facilitate knowledge discovery.

Platform Description: Moffitt Cancer Analytics Platform (MCAP) is Moffitt’s next-generation, cloud-based advanced analytics platform. MCAP’s infrastructure is primarily comprised of an Amazon S3-based data lake and Snowflake cloud data warehouse. MCAP efficiently unifies a broad range of multi-modality data and enables seamless end-to-end data lifecycle management by feeding diverse and typically siloed local data streams into a central data repository. This includes data from the electronic health record, cancer registry management system, biobanking management system, and patient-reported information including electronic surveys, as well as organizational data relating to billing and scheduling. A comprehensive inventory of our domain-specific institutional databases, ranging from clinical management tools to project-specific datasets, facilitates the ongoing integration of novel data resources including detailed, manually-abstracted treatment and outcomes data. In addition, we have implemented an innovative molecular data warehouse for the storage and management of genomic information ranging from raw sequencing data to identified genotypic and phenotypic alterations, annotations from public resources, and potential clinically-actionable targets. Data from these diverse sources has been harmonized to facilitate intuitive user querying; billing and diagnosis data is mapped to common medical ontologies, demographic and clinical data are conformed across systems, and derived fields have been created to support common research use cases. The initial loading of the data warehouse, along with the debut of a data science toolkit including the RStudio Team product suite (RStudio Workbench, RStudio Package Manager, RStudio Connect), integrated with Moffitt’s Snowflake and GitHub Enterprise instances, was completed in the Spring of 2021. Since this time, we have further launched two custom Curated Data Marts (CDMs) to actively support operations in the departments of Precision Medicine and Gastrointestinal Oncology at Moffitt, built on the foundational infrastructure of MCAP.

Discussion: This rich data asset allows us to fully leverage existing data for secondary use, laying the foundation for additional customizable, domain-specific CDMs and analytics solutions, as well as for the application of advanced machine learning and artificial intelligence applications which require clean, curated data from a range of sources. The novelty of MCAP lies not only in the number of unique source systems linked at the patient, tumor, and/or sample level, but also in the scope of the MCAP toolkit and documentation prepared to facilitate the interaction of both data analyst/scientist and faculty/staff end users with the available resources. We will describe the full scope of data captured in the warehouse including linkages between systems and novel approaches to data cleaning and curation, data governance and quality control considerations associated with the platform, and ongoing projects and future goals including natural language processing for extraction of disease characteristics from unstructured clinical narrative text and storage of images and associated metadata. We will additionally provide examples of the custom, group-specific CDMs and dashboards developed to date, highlighting challenges faced and solutions identified and implemented along the way.

References
Evidence points to gene-environment effects having a contribution to heritability for traits related to adiposity, with alcohol consumption, smoking behavior, diet, age, gender, and socioeconomic status, having significant gene-environment associations for body mass index (BMI) estimated through genome-wide methods. However, gene-environment effects are notorious for suffering from lack of replicability, and these findings have largely not been replicated across independent cohorts. Using three separate cohorts – UK Biobank (UKBB), Penn Medicine Biobank (PMBB), and eMERGE – we tested 51 exposures in at least one dataset for evidence for gene-environment interaction with BMI in adult European ancestry individuals, 9 of which were present in more than one dataset. Evidence for gene-environment interaction was measured by calculating a polygenic risk score (PRS) for BMI, and regressing BMI on the PRS, the exposure, an interaction term between the PRS and exposure, and appropriate covariates (age, gender, top 5 genetic principal components), with the p-value for the interaction term being the significance of the gene-environment interaction. We first reaffirmed significant gene-environment effects in UKBB for BMI with socioeconomic status (Townsend deprivation index, p=3.29x10^{-48}), alcohol consumption (days per week of alcohol consumption, p=3.69x10^{-48}), and physical activity (number of days of moderate-vigorous physical activity per week, p=7.87x10^{-48}). Next, we observed significant interaction with age and gender across all three datasets (Bonferroni p<2.17x10^{-4} (.05/23)). We also observed significant interaction with triglycerides in UKBB and eMERGE, with the effect being directionally consistent in PMBB. These analyses will provide further evidence of the significance of gene-environment effects for BMI, encouraging future work to understand and utilize differences in environment for treatment and understanding of diseases that BMI is a risk factor for. Future work includes additional replication cohorts including Genetic Epidemiology Research on Aging (GERA), Biobank Japan, and a proprietary dataset of 6,566 indigenous African individuals, replication using non-European ancestry individuals, and a novel procedure correcting for shared heritability between BMI and the exposure being tested.

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Units</th>
<th>Cohort</th>
<th>N</th>
<th>Interaction Beta</th>
<th>Interaction P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Years</td>
<td>PMBB</td>
<td>9,342</td>
<td>-0.00120</td>
<td>1.29x10^{-3}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UKBB</td>
<td>376,729</td>
<td>-0.000574</td>
<td>3.14x10^{-29}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>eMERGE</td>
<td>41,393</td>
<td>-0.00110</td>
<td>4.76x10^{-13}</td>
</tr>
<tr>
<td>Diastolic blood</td>
<td>mmHg</td>
<td>PMBB</td>
<td>9,084</td>
<td>-1.61E-04</td>
<td>.816</td>
</tr>
<tr>
<td>pressure</td>
<td></td>
<td>UKBB</td>
<td>352,804</td>
<td>2.72E-04</td>
<td>9.74x10^{-18}</td>
</tr>
<tr>
<td>Ever smoker</td>
<td>1 ever smoker</td>
<td>PMBB</td>
<td>5,689</td>
<td>-0.0193</td>
<td>.155</td>
</tr>
<tr>
<td></td>
<td>0 never smoker</td>
<td>UKBB</td>
<td>375,410</td>
<td>-0.00158</td>
<td>5.31x10^{-2}</td>
</tr>
<tr>
<td>HbA1c</td>
<td>mmol/mol</td>
<td>PMBB</td>
<td>2,095</td>
<td>0.00321</td>
<td>.188</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UKBB</td>
<td>358,798</td>
<td>0.000255</td>
<td>8.19x10^{-5}</td>
</tr>
<tr>
<td>HDL</td>
<td>mmol/l</td>
<td>PMBB</td>
<td>4,934</td>
<td>0.0147</td>
<td>.410</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UKBB</td>
<td>328,719</td>
<td>-0.0123</td>
<td>3.08x10^{-31}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>eMERGE</td>
<td>31,325</td>
<td>-0.0100</td>
<td>.105</td>
</tr>
<tr>
<td>LDL</td>
<td>mmol/l</td>
<td>PMBB</td>
<td>4,964</td>
<td>-0.00941</td>
<td>.305</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UKBB</td>
<td>358,556</td>
<td>-0.00653</td>
<td>1.92x10^{-42}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>eMERGE</td>
<td>30,502</td>
<td>-0.00598</td>
<td>8.79x10^{-5}</td>
</tr>
<tr>
<td>Never vs current</td>
<td>1 current smoker</td>
<td>PMBB</td>
<td>2,987</td>
<td>-0.0240</td>
<td>.402</td>
</tr>
<tr>
<td></td>
<td>0 never smoker</td>
<td>UKBB</td>
<td>242,693</td>
<td>7.53E-04</td>
<td>.595</td>
</tr>
<tr>
<td>Sex</td>
<td>1 male, 0 female</td>
<td>PMBB</td>
<td>9,342</td>
<td>-0.0537</td>
<td>6.36x10^{-7}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UKBB</td>
<td>376,729</td>
<td>-0.0106</td>
<td>4.18x10^{-39}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>eMERGE</td>
<td>41,393</td>
<td>-0.0380</td>
<td>6.45x10^{-15}</td>
</tr>
<tr>
<td>TG</td>
<td>mmol/l</td>
<td>PMBB</td>
<td>5,171</td>
<td>-0.0142</td>
<td>.173</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UKBB</td>
<td>358,933</td>
<td>-0.00320</td>
<td>2.20x10^{-16}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>eMERGE</td>
<td>30,983</td>
<td>-0.0130</td>
<td>8.58x10^{-5}</td>
</tr>
</tbody>
</table>
Learning about Impact on Chronic Diseases from #COVID19 Conversations on Twitter

Nishant R. Jain, MS, MHA1,2, Suzanne A. Boren, PhD, MHA, FAMIA1,2
1 MU Institute for Data Science and Informatics, University of Missouri, Columbia, Missouri, USA, 2 Department of Health Management and Informatics, University of Missouri, Columbia, Missouri, USA

Introduction

Social media can be leveraged for spreading awareness about chronic diseases such as cancer [1] as well as patient education and support [3]. Twitter can often be vulnerable to misinformation-led campaigns. Therefore, it is important to evaluate the trends, characteristics and credibility of chronic disease information on this platform during the COVID-19 pandemic. In this preliminary study, we have used social network analysis to capture the visual maps of Twitter users’ conversations about various chronic diseases in relation to #COVID19.

Methods

Twitter data was collected and analyzed using NodeXL [2], a social network analysis tool developed by the Social Media Research Foundation. Search terms included the hashtag #COVID19 coupled with the chronic disease name: #COVID19 + cancer, #COVID19 + diabetes, #COVID19 + hypertension, #COVID19 + heart disease and #COVID19 + obesity. Quantitative and network analysis was done using NodeXL. The social network analysis retrieved tweets from 12030 users from 19497 tweets posted in August 18- 29th 2021 & Dec 22nd - Jan 2nd 2022.

Results

The number of tweets retrieved ranged from 12-6982 and number of users ranged from 16-3939 users for each of the chronic disease networks. The largest overall network was that of cancer with 3939 users and the smallest overall network was for hypertension with 194 users (Aug 2021). This trend was replicated in December 2021. The largest networks from ascending to descending were in the following order: cancer, diabetes, obesity, heart disease and hypertension in August 2021 with heart disease and diabetes switching places in December 2021. Obesity (1.53) & cancer (2.27) had the greatest number of tweets per user in August 2021 & December 2021 respectively.

Conclusions

We find that during two different time periods in COVID-19 pandemic, among the five chronic diseases assessed, cancer was by far the most talked about chronic disease on Twitter. Diabetes & heart disease were the second most popular topics in August 2021 & December 2021 respectively indicating a shift in frequency of topics discussed. We found that the largest Twitter subgroup for #COVID19 + cancer, displayed a broadcast shaped network and included a Spanish speaking physician at the center of the network who tweeted that COVID-19 tests for asymptomatic patients are as valid as the administration of chemotherapy for cancer. In future studies, we will further assess content to gauge the scientific validity of the trending conversations and examine the categories of influencers. Social network analysis of Twitter data has potential to give context and assess the public health trends and thereby inform influence and support policies and research directions for chronic disease conversations related to COVID-19.

References


692
**Azure Modern Analytics Architecture for Synthetic Syndromic Surveillance**

**Arash Jalali MPH, MSHI**, **Sean Huang MD**, **Karl M. Kochendorfer, MD**

**1University of Illinois Hospital and Health Sciences System, Chicago, IL**

**Introduction**

The use of machine learning to build predictive models often requires the harmonization of disparate data sources. For instance, at our institution at the University of Illinois at Chicago, we have found a need to meld real-time geographic data with hospital data and enriched geospatial information. This required a complex architecture for data analytics over cloud services which had not been published in the literature over the past decade. This poster aims to discuss the construction and design of synthetic syndromic surveillance cloud architecture for pre-hospital and hospital data integration.

**Process and Development**

The open-source data analytics architecture that we used is shown here. Data can be obtained from various sources, including patients and populations, from an electronic health record. We can also use synthetic data obtained from sources such as the RTI U.S. Household Population™. Afterward, the data is stored in a hyper-scale repository Azure Data Lake. After filtering with ICD-10 codes, the data is ingested into an Azure Synapse SQL Server. Azure pipelines help to clean and filter the data inside the Azure SQL data warehouse. Data analysis can be performed using Azure Synapse Spark as an analytics platform and Azure Machine Learning for its analytic services. Predictive models can be created, deployed, and managed by this means. We can move the data into Azure Databricks and into ArcGIS Enterprise for geospatial analytics.

**Conclusion**

The presence of health data, synthetic data, and the lack of interoperability between these forms for research necessitates designing a new system to perform modern data analytics efficiently. Such experiments often require the use of a complex data warehouse as well as a cloud computing environment. Here we have described an open-source architecture that we have developed to successfully build our unique predictive models at the University of Illinois. This architecture supports diverse workloads and data types ranging from unstructured to structured data.

**References**


The Potential in Structured Process for Dysphagia Rehabilitation after Cerebrovascular Diseases

Shogo Kato, Ph.D., Satoko Tsuru, Ph.D., Makoto Ide, MD, Ph.D., Akira Shindo, MD, Naohisa Yahagi, MD, Ph.D., and Shu Yamada, Ph.D.

1Keio University, Kanagawa, Japan; 2National Center for Child Health and Development, Tokyo, Japan; 3The University of Tokyo, Tokyo, Japan; 4St. Mary’s Healthcare Center, Fukuoka, Japan; 5University of Occupational and Environmental Health, Fukuoka, Japan; 6Ooguno Hospital, Tokyo, Japan

Introduction

Japan has been a super-aged society as more than 28% of the population is 65 years or older in 2020. Rehabilitation is an important treatment that affects the prognosis. Standardization of rehabilitation processes is difficult because there are few quantitative indicators, and they depend largely on the individual therapists rather than medicines and/or instruments. As a result, there are differences in processes and outcomes among individual therapists and hospitals.

In this study, we aim to structure process for rehabilitation after cerebrovascular diseases. Then, we aim to computerize the operation for dysphagia rehabilitation using Patient Condition Adaptive Path System (PCAPS), which is a model and/or system to structure clinical knowledge based on patient condition to implement it into hospitals.

Method

The inherent/empirical knowledge of experienced therapists in a hospital was visualized and structured, focusing on responsible lesions, dysfunctions, and examinations using 2D-matrix format, and modified through discussion with physicians and nurses among four hospitals. We designed and improved rehabilitation processes for dysphagia, language disorder, basic motion, and work activity. We organized assessments and interventions needed for dysphagia, and identified the components of assessments, interventions, and relationships between assessments and interventions.

We elaborated the process, assessments and interventions for dysphagia rehabilitation based on PCAPS. PCAPS structures clinical knowledge by both “Clinical Process Chart (CPC)” describing the overall flow, and “Unit Sheet (US)” describing the details. We computerized the operation for dysphagia rehabilitation on “PCAPS-Administrator,” which is an existing computer application to administrate processes in PCAPS format, and implemented it retrospectively into an acute stage hospital by multiple therapists to validate the comprehensiveness of the contents.

Results

Assessments and interventions needed for dysphagia in USs was structured by 111 assessment items and 140 intervention items. The relationships between assessments and interventions were structured by 158 records.

Rehabilitation processes were identified to have patient condition-oriented structure, and the operation was rationally computerized on PCAPS-Administrator. The structured process for dysphagia rehabilitation was operated on PCAPS-Administrator in the acute stage hospital retrospectively to record 345 cases of clinical process for 21 months. The clinical process for all cases were recorded based on the CPC. In addition, all 78,714 actual assessments were recorded based on the US. Thus, 8,148 out of 8,166 (99.8%) actual interventions were recorded based on the assessments.

Conclusion

The comprehensiveness of the contents for dysphagia rehabilitation was confirmed because clinical processes of all cases could be recorded based on the CPC, and all assessments and almost all interventions could be recorded based on the US. Thus, it becomes possible to operate a rehabilitation process structured based on patient conditions. Furthermore, by accumulating structured assessment data, it would be possible to evaluate the effects of interventions by outcomes such as “food form,” which is the form of the meal the patient can eat, and/or process indicators such as “progress and/or speed on CPC,” in addition to compare and verify the transition of patient conditions.
Deep Learning-based Approaches for Relation Extraction between Cancer Treatment and Temporal Information from Electronic Health Records

Youngjun Kim, Wenzhe Yang, Timmy O’Connell, Arielle Redfern, Xiang Zhou, Rong Chen

1Sema4, Stamford, Connecticut, USA

Introduction: Extraction of clinical events with temporal information from clinical notes is important as a fundamental step to optimize patient care and prevent adverse events. As a use case, we focus on extracting relations between cancer treatments and associated dates from electronic health records (EHRs). Statistical machine learning methods have been leveraged to identify relations with temporal information. Our target cancer treatments include chemotherapy, immunotherapy, hormone therapy, and a combination of these. A Cancer Treatment mention is associated with none, one, or more Date (e.g., “Feb 21, 2019”) and Duration mentions (e.g., “two weeks”). We put the relations of interest into two categories: Positive, which accounts for present or past events, and Negative, which accounts for negated, uncertain, hypothetical, not referring to the patient, or generic events. Therefore, this task involves recognizing six types of relations: Date-Pos (positive), Date-Neg (negative), Duration-Pos, Duration-Neg, None-Pos, and None-Neg. Unlike previous studies, we train the model using two different types of examples: a treatment only (i.e., None-Pos and None-Neg examples, e.g., “We held his Lupron”) or a pair of two mentions (i.e., all other relations, e.g., “Lupron given 02/21/2019”). This data augmentation allows us to simultaneously determine the attributes and relations of each treatment.

Methods: We used 200 EHR text notes randomly selected from prostate cancer patients treated at Mount Sinai. All mentions of these treatments and their relations with dates and durations were annotated by medical experts. The inter-annotator agreement was good with F1 scores of 91.6% for mentions and 84.9% for relations, respectively. Our corpus was randomly split in a training subset of 120 notes and a test subset of 80 notes. Date-Pos and Duration-Pos relations accounted for 33% of the instances in the test data. We trained three relation classification models. As input to each model, we used the words contained in each mention, 10 preceding and 10 following words for each concept, and the words between the two mentions. We created a multi-class ECT (Error Correction Tournament) classifier as a shallow (or classical) learning method. We developed two relation classification models that use deep learning. We trained an RNN (recurrent neural network) model with static word embeddings. The model can regard the training example as a sequence of words and efficiently capture the local syntactic structure of the sequence. We pretrained static word embeddings with the MIMIC-III clinical dataset (version 1.4). We created a 300-dimensional skip-gram model using the fastText library. For another deep learning model, we fine-tuned the transformer model on the test set. We used the uncased BlueBERT-Large embeddings pretrained with MIMIC-III and PubMed abstracts. Detailed hyperparameter configuration will be provided in a more extended publication.

Results: Table 1 shows the overall performance of each model and Table 2 details the transformer model performance for each category. Precision (P), recall (R), and F1-scores (F1) were calculated to evaluate relation extraction of given reference standard concepts. Deep learning models outperformed the ECT multi-class model. The transformer with BERT embeddings achieved the best accuracy (82.6% F1 score). This model reached over 80% F1 score in all categories except the Date-Neg class.

Conclusion: We demonstrated that our deep learning-based approach can efficiently and accurately classify the attributes and relations of each cancer treatment simultaneously. We are creating new text collections from different cancer types and our future goal is to further increase generalizations across these heterogeneous data.

References
Methods Used to Assess mHealth Applications for Cardiovascular Disease: First Results From a Quasi-Systematic Scoping Review

Jennifer Kircher, BA¹, Walter Swoboda, MD¹, Felix Holl, M.Sc.¹,²
¹DigiHealth Institute, Neu-Ulm University of Applied Sciences, Neu-Ulm, GERMANY; ²IBE, Ludwig Maximilian University of Munich, Munich, GERMANY;

Introduction

The demographic change and a change towards chronic diseases have led to increasing healthcare costs in high-income countries. A major burden of illness is attributed to coronary heart disease. 331,211 deaths were attributable to cardiovascular diseases (CVD) in Germany in 2019¹. CVDs are also the highest cause of healthcare cost in Germany, 46.4 billion Euros in 2015². Mobile health (mHealth) applications can improve care and decrease costs. This project examines the extent to which existing evaluation criteria and methods of mHealth applications for CVD.

Methods

A quasi-systematic scoping review was conducted. The first stage was a scoping review to develop a search strategy and define inclusion and exclusion criteria. Then, three databases ("PubMed", "Livivo" and "ProQuest") were search for literature published in English or German between 2000 and 2021 using a fixed search strategy. Subsequently, the data of the identified studies were extracted and analyzed.

Results

After duplicates removal, a total of 4066 studies were screened, of which 38 met the inclusion criteria, all of which were published in English. One third come from the US, 13% from Australia, and 10% from China. Quantitative and qualitative research designs are included in the review. 18 studies consist of randomized controlled trials (RCT). The study duration of each intervention varies. Most studies used standardized questionnaires (n=31). The overall aim was to assess participants' perceptions of treatment and subjective health (n=33). 63% evaluated mHealth intervention using various medical measurements. The aim was to analyze health parameters before and after the intervention. Interactions with mHealth on the part of patients (n=19) and health care providers (n=2) were measured by usage protocols (n=19). Conclusions could be drawn about the motivation (n=17), adherence (n=18) and self-efficacy (n=14) of the participants. The usability evaluation (n=14) of mHealth interventions included several methods and measures. Several publications investigated the effectiveness and efficiency of mHealth (n=14).

Discussion

Of the 38 studies, RCT studies were most frequent. With the active involvement of patients in the treatment process, a large proportion of the evaluations dealt with the user perspective. Different methods and criteria were used. Standardized questionnaires were the most common. Quantitative methods mean less time and cost for researchers because of validated and meaningful data. Almost all studies determined the added value of an mHealth intervention through clinical results. Laboratory diagnostics and physical tests were found to be the most meaningful in assessing objective physical health. Subjective quality of life was also evaluated in standardized surveys. In a direct comparison of an mHealth treatment with a standard intervention, economic evaluations were carried out only sporadically. Cost calculations in the form of cost-effectiveness analyses should be carried out to prove efficiency and effectiveness. Limitations of this work are that new work about this topic is being published frequently.

Conclusion

Criteria such as usability, motivation, and user experience were most frequently evaluated using standardized questionnaires and usage protocols. Clinical outcomes were assessed by laboratory diagnostics and quality of life questionnaires. Economic evaluations were sporadically conducted in the context of cost-effectiveness analyses.

References

Building a Scalable Knowledge Management Approach to Support Evidence Provision for Precision Medicine

Taneya Y. Koonce, MSLS, MPH1; Mallory N. Blasingame, MA, MSIS1; Annette M. Williams, MLS1; John D. Clark, MS1, Spencer J. DesAutels, MLIS1, Dario A. Giuse, Dr. Ing, MS, FACMF2, Jerry Zhao, MS, MLIS1, Jing Su, MD1, Helen M. Naylor, MS1, Nunzia Bettinsoli Giuse, MD, FACMI1,2

1Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN; 2Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

Introduction

Maintaining understanding of a clinical topic requires regular effort to incorporate new knowledge to ensure practice is aligned with the most current evidence. This can be a time-consuming process, which has led to a need for automated assistance. Automation of article abstract and full-text review for relevance to a clinical question has been implemented for contexts such as systematic review screening1, but replicable automated approaches are also indicated to efficiently identify new evidence impacting patient care. Awareness and use of genetic information is increasing for both clinicians and patients, making knowledge maintenance in this field both challenging and important. Furthermore, high-quality and timely genetic information comes from a variety of specialized sources beyond journal articles, necessitating an approach to automation that can be flexibly applied regardless of information and metadata structure. The Center for Knowledge Management (CKM) has provided clinicians with evidence syntheses for over two decades, with an increasing number of requests related to genetics over that time. CKM information scientists have established practices for fully documenting and archiving evidence syntheses, along with search strategies used to identify relevant journal articles and grey literature in the CKM-developed Clinical Support Knowledge Acquisition and Archival Tool (CS-KAAT)2. This tool supports ongoing consultation, update, and reuse. As part of an institution-wide initiative to integrate genetic testing more fully into clinical practice, the expertise of CKM information scientists is used to create and maintain evidence summaries on actionable and strategic genetic topics.

Project Description

To improve and scale the maintenance of genetics evidence syntheses, CKM is enhancing CS-KAAT by incorporating regular update alerts and applying knowledge management principles to the selection of newly published evidence for inclusion in revised evidence syntheses. Information scientists will capture metadata highlighting the most important elements of the completed evidence summary. Simultaneously, we will solicit and translate into enhanced metadata the value-added reflection of subject matter experts to refine our ongoing evidence maintenance strategy. Currently we are piloting the use of natural language processing based on this metadata to recommend relevant new information in scheduled alerts, expediting the evidence synthesis revision process. While approaches have been described to automate the filtering of genomic literature from PubMed/MEDLINE3 and aid in variant interpretation4, this project will extend and semi-automate evidence provision, filtering, and maintenance across both traditional literature databases and specialized genetic resources to flexibly support ongoing update of genomic evidence targeted to both specific patient cases and broader clinical contexts/populations. The key aim of this project is to build upon the existing and proven knowledge acquisition techniques of evidence provision and filtering to build a fast infrastructure able to support the ever-changing, complex world of precision medicine.

References

Using N3C Enclave for Drug-Drug Interaction studies: the case of DOAC-Dexamethasone Interaction and the risk of Thrombosis

Olga V. Kravchenko, PhD1, Ainhoa Gomez-Lumbreras, MD, PhD2, Paul T. Kocis, PharmD, MPH3, Daniel Malone, PhD, FAMCP2, Charles E. Leonard, PharmD, MSCE4, Richard D. Boyce, PhD1
1University of Pittsburgh, Pittsburgh, PA; 2University of Utah, Salt Lake City, UT; 3Penn State Health, Hershey, PA; 4Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA

Abstract
We conducted a population-based drug-drug interaction (DDI) study in the National COVID Cohort Collaborative (N3C) Data Enclave. The study aimed to evaluate the association between concomitant dexamethasone and direct oral anticoagulant (DOAC) use and thrombosis. Contrary to our prespecified alternative hypothesis, we did not find statistically significant interaction using the latest data in the N3C Enclave. However, our approach can be applied to similar studies utilizing data sets harmonized with OMOP common data model.

Introduction
Dexamethasone is a corticosteroid that has been shown to improve outcomes of patients diagnosed with moderate to severe COVID-19 with respiratory failure who received supplementary oxygen or mechanical ventilation. Many of these patients are also treated prophylactically with anticoagulant therapy to prevent thromboembolic events. Prolonged exposure to dexamethasone may induce metabolism of cytochrome P-450 (CYP) 3A4 thus creating a potential to reduce the efficacy of DOACs, such as rivaroxaban and apixaban. It is currently unknown if co-exposed patients are more likely to experience thromboembolic events. We conducted a retrospective nested case-control study in the N3C Data Enclave to answer this important question.

Methods
We conducted a retrospective nested case-control study in which we assessed the odds of co-exposure to dexamethasone/DOAC (for ≥ 5 days) among persons with (cases) and a risk-set sampled subset of persons without (controls) a thromboembolic event. The data for this study are in the N3C Data Enclave, a cloud-based federated research platform that was created in 2020 to address the need in expedited data-driven translational research focused on COVID-19. The platform is updated weekly with electronic health records data from more than 40 contributing hospitals and health care sites from across the country harmonized into OMOP common data model. We have used Limited Data Set (LDS) version of the data that included diagnoses, drug exposure, observations, laboratory measurement results, procedures, and demographics. Our statistical analysis included McNemar Chi-Square test with continuity correction and conditional logistic regression. We conducted sensitivity analyses to evaluate the effect of the following parameters on the result: age window, data partner date shift, order of initiation of the putative interacting pair, and duration of dexamethasone exposure. We used SQL to construct a reproducible data workflow, Python for the final logistic regression analysis and R for McNemar Chi-Square test.

Results
We did not observe a statistically significant association between dexamethasone + DOAC (rivaroxaban and apixaban ($\chi^2(1, N = 333) = 0.5, p = 0.4795; OR=1.15 (95% CI: 0.32, 4.18)$)). However, sensitivity analyses showed significant association with aspirin and heparin, as well as several health conditions such as heart failure and hypertension. While insufficient data precluded us from testing dose-response, there was sufficient data to evaluate exposure and timing of the drugs. We created more than 20 concept sets for drugs, conditions, and measurements for the study that could be repurposed for other studies.

Conclusion
We developed an approach to conduct a DDI study in the N3C Data Enclave to extract and analyze data for retrospective, observational study for a potential DDI that could impact patients with COVID-19. While we found no statistically significant association, our workflow can be adopted to other DDI studies of a similar epidemiological design in the N3C Enclave.
Real Time Decision Support for Achieving Maximal Compliance with Annual Suicide Screening Requirements in the U.S. Department of Veterans Affairs

David V. LaBorde, M.D., M.B.A. 1, Christina M. Alvaro, B.A.Sc. 2, Kassi L. Goodman, M.A. 2
1Iconic Data, Inc., Norcross, GA, USA, 2Document Storage Systems, Juno Beach, FL, USA

Introduction
More than 45,000 Americans died of suicide in 2017 and this figure includes 6,139 U.S. Veterans. 1 In 2018, the U.S. Department of Veterans Affairs (VA) Office of Mental Health and Suicide Prevention’s (OMHSP) Suicide Risk Identification Strategy (Risk ID) was developed and implemented 2. By OMHSP policy, Columbia Suicide Severity Rating Scale (C-SSRS) screening must be attempted annually for every Veteran receiving care at the VA. At present, VA facilities utilize clinical reminders to prompt providers to complete screenings when they are due. However, there is not currently an easy way for facility suicide prevention teams to identify which Veterans are due for screening and, of these due screening, when they are next scheduled for care and where. There is also no real-time reporting that can aid suicide prevention teams and clinic staff and managers in determining whether the required screening was actually completed. If VA facility clinic staff at the various points of care are able to easily know which patients on the schedule are due for C-SSRS screening without having to open each chart, this could increase the probability that the screens are attempted. Moreover, real-time monitoring would allow stakeholders in the clinic, for example when the patient is checking out, to easily determine whether a screen was due and, if so, the completion status – without opening each chart. This additional check can potentially help avoid scenarios where a Veteran due screening leaves their visit without the screen being attempted. This would be meaningful as completing a screen: i) helps identify at risk Veterans that may benefit from additional suicide prevention intervention; and ii) avoids the need for service recovery – namely, attempting to complete the screening after the Veteran has departed the visit, which may or may not succeed. The current OMHSP dashboards that support this workflow describe the percent of time that screening attempts were made when due. However, these dashboards are only updated monthly, and they provide no decision support to assist with compliance, making maximal compliance with this policy an operational challenge.

Methods
To enable the VA to achieve high reliability with this key aspect of the Risk ID Strategy, a VA hospital information system integrated solution (Suicide Prevention Manager™, Iconic Data, Inc., Norcross, GA) was created to address this enterprise-wide vulnerability. Requirements were established and a Veterans Health Information Systems and Technology Architecture (VistA) remote procedure call (RPC) interface was developed utilizing the Massachusetts General Hospital Utility Multi-Programming System (MUMPS). The solution was then tested and deployed in a VA VistA test environment prior to subsequent deployment in a VA production environment.

Results
A list of the subset of Veterans coming due for screening is maintained via automation. The system constantly compares this population against clinic schedules and patients that present as walk-ins to clinic or to the emergency department or urgent care center. Real-time decision support reporting filterable by multiple properties, including the point of care, is available. This provides an always up to date list of the Veterans currently checked in or that are scheduled for care and the status of the required screening; completed screenings are monitored in real time.

Discussion
Suicide prevention policies will only attain their full potential when compliance is maximized. The probability of achieving maximal compliance may increase with proactive management of the workflow. Situational awareness of what action is needed when is key to achieving highly reliable care processes. A real-time solution for managing annual suicide screening processes can help the VA attain the highest possible compliance with this aspect of their Risk ID Strategy. Future work could aim to understand and quantify the impact of such VA systems.

References
Machine learning on multivariate time-series clinical data for early prediction of prostate cancer metastasis

Morgan Lafferty, BA, Isabel Friesner, BA, Sasha Yousefi, BA, Julian C. Hong, MD, MS
University of California, San Francisco, San Francisco, CA, USA

Introduction:
Prostate cancer is the most common cancer and responsible for the second most cancer deaths among men in the United States. Advances in imaging such as Prostate-Specific Membrane Antigen (PSMA) Positron Emission Tomography (PET) have enabled early detection of metastatic prostate cancer, particularly oligometastatic prostate cancer, which may allow intervention for durable control or even cure. In this study, we aim to use longitudinal electronic health record (EHR) data to predict future prostate cancer metastasis.

Methods:
We identified 4,464 patients with at least a 6-month history of prostate cancer (ICD-10 C61) preceding/without a diagnosis of metastatic cancer (C79), seen at UCSF between 2013 and 2020; 252 (5.6 %) developed metastatic prostate cancer during the study period. Structured EHR data including demographics, encounters, medications, lab results, and ICD-10 diagnoses were extracted and patient multivariate time series with 30-day intervals were created. Training (60%), validation (20%) and testing (20%) data sets were generated on a temporal basis, with the training and test fold containing the earliest and latest patients, respectively. All data was standardized with respect to the training set. A temporal convolutional network (TCN) was fit to the training data and hyperparameters were tuned on the validation set. For comparison, a LASSO logistic regression model was fit to the training data at the current timestep and tuned by validation. Feature importance was interpreted using the SHAP and DeepExplainer for the TCN and by model coefficients for LASSO logistic regression. Kaplan Meier plots were generated to assess model stratification for right-censored time to metastasis by separating test patients into four groups based on ascending prediction probabilities of developing metastatic prostate cancer.

Results:
Kaplan Meier curves of TCN-based and LASSO-based stratification were similar (Figure 1). Feature-wise Shapley values computed with SHAP and combined across all time steps showed the most influential features in the TCN model were active systemic therapy, specifically bicalutamide or degarelix, testosterone level, and patient age. LASSO logistic regression coefficients showed similar importance to bicalutamide therapy, age, and testosterone level, in addition to radiation therapy (CPT code 77385).

Discussion:
Both TCN built from temporal and LASSO logistic regression based on single time point EHR data demonstrate strong predictive and stratification abilities for time to metastasis. Limitations include rare events, dependence of metastasis identification on potentially delayed ICD-10 entries, and potential stage migration due to increasing PSMA PET adoption. Despite absence of a significant performance boost with time-series data, both models show promise in early prediction of prostate cancer metastasis which may allow early imaging and intervention. Further, feature importance analyses reveal model predictions follow reasonable clinical expectations. Ongoing work includes continued iterative improvement of the TCN approach, integration of other data, including natural language processing-derived features from pathology reports, and characterizing the impact of clinician data collection on timing of EHR-based metastatic diagnoses.

Acknowledgements:
This study was supported by the American Cancer Society, American Society for Radiation Oncology, and Prostate Cancer Foundation. Funders had no role in design and conduct of the study, nor the decision to prepare and submit the abstract. JH is also supported by a career development grant from the American Society of Clinical Oncology.
Leveraging EHR Data to Support a Hospital GIS for Hospital-Onset Clostridioides difficile (C. difficile)

Kaiting Lang (Ph.D. student at College of Public Health, Division of Epidemiology, The Ohio State University, Columbus, Ohio)
Dr. Courtney Hebert, MD,Ms (Assistant Professor, Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio)
Dr. David Kline, Ph.D. (Assistant Professor, Department of Biostatistics and Data Science, Wake Forest School of Medicine, Winston-Salem, North Carolina)
Dr. James B. Odei, Ph.D. (Assistant Professor, Division of Biostatistics, The Ohio State University, Columbus, Ohio)
Dr. Elisabeth Root, Ph.D. (Principal Research Scientist, Institute for Disease Modeling, The Bill and Melinda Gates Foundation, Seattle, Washington)

**Background:** Healthcare associated infections (HAIs) cause a significant amount of morbidity and mortality and can occur in outbreaks within medical facilities. The standard way these outbreaks are investigated includes regular review of numbers and locations of cases, and when there is a concern, manual chart review of cases. As part of a larger study, we are developing a hospital GIS (GeoHAI) to help infection preventionists investigate hospital outbreaks in real time with the help of spatiotemporal modeling and visualization. In order to accomplish this goal, the research team needed to determine how to transform electronic health record (EHR) data into an individual room-based, time stamped dataset. This poster describes our detailed process to allow this to be replicated at other institutions.

**Methodology:** Data were extracted from the EHR for a 5-year period (2014-2018), including admission-discharge-transfer data, encounter level information, basic demographic data, and C. difficile PCR test (result, date of test). A reusable pipeline was developed within R, with iterative feedback from subject matter experts (physicians, infection preventionists, geographers, biostatisticians). Patient level information was aggregated to quantify room contamination as defined by a room occupied by a patient 3 days prior to and 7 days after a C. difficile positive test. The final table was structured so that each row showed an hour for a particular room and if that room was considered contaminated during that hour. This structure facilitated the building of disease mapping models for the hospital floors.

**Results:** 295,984 encounters were included in the final dataset, with the final table including 18,215,254 rows. The process was as follows:

<table>
<thead>
<tr>
<th>Step</th>
<th>process</th>
<th>result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Extract data, create cohort: Read and join hospital admission discharge transfer (ADT) data, Demographic data, hospital encounter data and hospital lab data using individual visit ID.</td>
<td>Each row shows a patient’s stay in an individual room</td>
</tr>
<tr>
<td>2</td>
<td>Classify hospital onset (HO) or community onset (CO) C. difficile: For all positive C. difficile tests in a room, classified HO (positive test at least 4 days after admission) CO (less than 4 days after admission).</td>
<td>Step 1 plus CO or HO</td>
</tr>
<tr>
<td>3</td>
<td>Identify contaminated vs. non-contaminated period in a room. Separate time considered “contaminated” (3 days before to 7 days after) from each room stay with a positive test</td>
<td>Each row is a time period in a room considered contaminated vs. non-contaminated (n=957,768 rows)</td>
</tr>
<tr>
<td>4</td>
<td>Transform data into hourly timestamp format. A patient is considered to be in a room in an hour time-stamp if they spent at least 30 minutes of that hour in that room.</td>
<td>Each row is a one-hour period in a room plus all above (n=18,215,254 rows, only 2017, 2018 data)</td>
</tr>
</tbody>
</table>

**Conclusion:** EHR data can be transformed into data that can be leveraged in a typical GIS model. This information can be beneficial for improving future hospital acquired infection disease investigation.

*This project was supported by grant number R01HS027200 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality.*
Evaluating the Bias and Fairness of Machine Learning-based Clinical Prediction Models

Fuchen Li¹, Patrick Wu, PhD²; Josh F. Peterson MD, MPH³; Wei-Qi Wei, MD, PhD³; Juan Zhao, PhD²
¹College of Art and Science, Vanderbilt University, Nashville, TN, USA ²Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA.

Introduction

Machine learning (ML) promises to revolutionize clinical diagnose and prognosis prediction. However, in practice, health data is subject to bias in multiple ways through building and analyzing datasets (e.g., sampling bias). ML models trained on such data might amplify the bias and yield decisions skewed toward some groups of people, e.g. populations overrepresented in training data.¹ Currently, there is a lack of work exploring the bias and fairness of ML-based clinical prediction models. In this study, we explored this topic using ML models for cardiovascular diseases (CVD) risk prediction.² The objective of the study is to 1) understand the importance of detecting bias and assessing the fairness of ML-based models; and 2) evaluate metrics that can quantify the fairness of ML-based clinical predictive models.

Method

Design, Setting, and Data: We used a cohort derived from Vanderbilt University Medical Center (VUMC) de-identified EHR including outpatient adults who have a 10-year follow-up from Jan. 2007 to Dec. 2016. Machine learning models (logistic regression [LR], decision tree [DT], random forest [RF], and gradient boosting trees [GBT]) were trained using a 7-year EHR data prior to Jan. 2007 to predict CVD in the 10-year follow up period. CVD were ascertained via reference codes (ICD-9, 411.9 or 433.*). Predictors included demographic or lifestyle data, physical measurements, labs, diagnosis, and medication.² We split the dataset into training (80%) and test sets (20%). Race and gender, respectively, were the protected attribute that correspond to privileged status. Protected attributes were used to partition the population and were not used as predictors when training ML models.¹ The cutoff threshold for binary prediction was determined by the J-score. The model accuracy was evaluated using the area under the receiver operating characteristic (AUROC). We compared ML models with a clinical model in wide use – the American College of Cardiology and the American Heart Association (ACC/AHA) Pooled Cohort Risk Equation.

Fairness Evaluation: We assessed two group fairness metrics. 1) Equal opportunity difference (EOD) – difference in true positive rate between privileged and unprivileged groups. 2) Disparate impact (DI) – ratio of favorable outcome (predicted CVD) percentage between privileged and unprivileged groups. For algorithmic purpose, we defined White being the privileged group to Black, and Male being the privileged group to Female. In an ideal fair model, EOD = 0 and DI = 1.

Results

The study cohort included 109, 490 individuals (mean [SD] age 47.4 [14.7] years; 64.5% female; 86.3% White, 13.7% Black). From Figure 1, all the models are fairer across racial groups than gender groups. We found that models favor men (EOD>0 and DI>1), who have a higher percentage of CVD diagnosis in our dataset. Compared to ACC/AHA equation, most ML models are less biased, especially for gender group. Among all models, GBT has the highest AUROC (GBT:0.790 > LR:0.784 > RF:0.770 > DT:0.753 > ACC/AHA: 0.730), but it only has a moderate EOD and DI value, indicating that a model with a high accuracy is not necessarily very fair. We should also note that the reference value 1.0 for fairness for DI depends on applications, because the prevalence of certain diseases like CVD tends to vary across racial and ethnic groups.³

Conclusion

We evaluated the bias and fairness of multiple models predicting 10-year CVD risk. Almost all ML models have superior performance in the chosen fairness metrics than the ACC/AHA equation. Models with the highest AUROC may not perform the best in fairness, indicating that fairness should also be considered for performance evaluation.

Reference

Leveraging Standard Terminologies to Quantify Duplication in Patient Problem List

Sina Madani, MD, PhD, Yaa Kumah-Crystal, MD, MPH
Jonathan P. Wanderer, MD, MPhil, Asli Weitkamp, PhD
Vanderbilt University Medical Center, Department of HealthIT, Nashville, TN

Introduction

As a standard of care, providers are asked to maintain an up-to-date list of patient problems in the Electronic Medical Record (EMR) systems. Problem list maintenance is a shared responsibility among physicians, where completeness and accuracy may impact patient safety, quality measures, and reimbursement. Though standard terminologies are meant to facilitate diagnosis entry, different workflow considerations and variation in diagnosis terms may lead to duplication on problem lists, creating a “cluttering” effect. Furthermore, compositionality of interface terminologies, amplified by a representational layer of provider friendly terms, may also cause clinicians to document duplicate problems. We have used backend coding schema cross-walks from interface to standard terms, at the concept level, in order to evaluate semantic resemblance of patient problems to identify instances of potential duplicates.

Methods

We conducted a retrospective analysis of our EMR problem list from 10,000 randomly selected “inpatient” records from which 3,834 had 2 or more active problems. All ICD9/10CM and SNOMED CT codes were examined for similarity and is-a & attribute relationships. Our EMR interface terminology leverages pre-coordination to correlate a term (ex: “Type 1 diabetes mellitus with ketoacidosis”) to 1 or more SNOMED concepts (ex: “Diabetes Mellitus type 1” & “Ketoacidosis”). It also uses provider friendly terms, with the same ICD9/10CM codes, to offer a richer selection of terms to the users for documenting patient diagnosis. SNOMED CT ontological relationships are leveraged to find duplicate problems with hierarchical (where one of the concepts is typically more granular than the other one) and non-hierarchical (attribute) associations.

Results

Duplicated problems were found in 11.2% of patients (n=430). Rate of duplication based on the underlying coding system in these patients from interface terminology ranged from 44% (same ICD9/10CM code), to 29% (SNOMED CT relationships) including overlaps (Figure1). Examples of likely duplicated problems include: “Acne” & “Acne Vulgaris” for ICD9CM, “GERD” & “Chronic GERD” & “GERD without esophagitis” for ICD10CM, “Atrial Fibrillation” & “Permanant Atrial Fibrillation” and “Gout” & “Chronic Gouty Arthritis” for is-a and due_to SNOMED CT relationships, respectively.

Conclusion

Though standard terminologies offer “computable” terms, they are less expressive and potentially less intuitive to document at the point of care compared to interface terminologies with provider friendly terms. On the other hand, despite the fact that interface terms create a buffer zone between coding schemas and provider diagnoses, they may inadvertently cause duplicate problems with various levels of granularities in patient charts. While it can be argued such duplicative instances can be prevented through alerts or cleaned up afterwards there are still technical difficulties in implementation, as well as workflow integration, for such features in EMRs. We are currently evaluating likely duplicated patient problems based on their interface terms to develop a notification system for preventing providers adding diagnosis where a similar one already exists in the chart.

Reference

Supporting Electronic Health Record Data Usage in Research for Teams with Varying Data Science and Clinical Knowledge
Tanja Magoc, PhD1, Yujun Chen, MS1, Yankuic Galvan, PhD1, Kevin Hanson, MEd1, Rebecca Liao, BS2, Gloria Lipori, MT, MBA3, Lei McDevitt, PhD1, Khoa Nguyen, PharmD4, Frantzler Pierre, MS1, Alexis Quinones, MS3, Ben Staley, PharmD3, Christopher A. Harle, PhD1
1University of Florida College of Medicine, Gainesville, FL, 2University of Florida College of Medicine-Jacksonville, FL, 3UF Health, Gainesville, FL, 4University of Florida College of Pharmacy, Gainesville, FL

In academic health centers, research requests for Electronic Health Record (EHR) data typically require an honest-broker process for specifying, extracting, and transmitting data to study teams. That process can be time- and labor-intensive, and this is at least partially driven by the need to accommodate teams with varying levels of experience with data request processes, data standards, analytics expertise, and knowledge of EHR data generation processes. This challenge is an example of a more general operations management problem of how to break the trade-off between efficiency and service when customers introduce variability into a service process. When designing service processes, data providers, such as a team of honest brokers, must make decisions on if and how to accommodate or reduce customer-introduced variability. In this study, we present and discuss strengths and weaknesses of two strategies recently implemented by our academic health center’s honest broker team to overcome the challenge.

For the first strategy, we followed reduction strategy by implementing a process for producing local research registries. These registries contain comprehensive de-identified EHR data for patient cohorts of interest, such as patients with COVID-19 or patients with cancer. Each registry conforms to the Observational Medical Outcomes Partnership (OMOP) common data model (CDM). OMOP and other CDMs have been developed to standardize EHR data in support of observational research, including multi-site research. Once a unique cohort definition is defined, each registry can be created and updated using similar export, transform, and load (ETL) processes. However, while these CDM-formatted registries allow for efficient data delivery to researchers, the size of some tables that contain multiple data dimensions (e.g., measurement table includes labs and all vitals) and complexity of linking EHR data to vocabulary tables require unique data management and analysis skills to create usable analytic datasets and produce study results. We have found that some researchers, especially clinically-trained or junior researchers, do not have the necessary skills to perform this type of analysis. Despite available documentation and communities that support each CDM, learning a CDM requires significant commitment by a study team. Overall, implementing research registries using OMOP CDM allowed us to standardize data structure and increase our efficiency in honest brokering process.

While learning to use CDM-formatted registries is compelling for research teams with data management and analysis expertise, and for junior researchers who plan to dedicate significant time and effort to research over many years, other valued team members in an academic health center, such as clinical faculty and trainees, might not have time or resources to dedicate to this learning. Therefore, we developed a locally focused data standardization process that aims to provide usable data to accommodate the later group. In this approach, we deliver each data dimension as a separate table, which reduces the size of individual tables and requires less computational resources for processing. We also allow variable number of columns between tables, so that we can include all necessary attributes of an event in a single row. Each table is accompanied with an internally developed data guide that describes all details about columns and possible values. This data guide provides locally-focused data description in comparison to widely-used CDMs whose vocabularies provide only high-level description of data. This simpler and more detailed data guide allows for quicker learning about, processing of, and analysis of local data. Notably, this approach is not well-suited for data linkages across institutions.

In conclusion, we applied an operations management framework for managing customer-introduced variability to the creation of new, more efficient service processes in research data honest brokering. In an academic health center, researchers and trainees with different project aims, skills, and financial and computational resources benefit from different data formats. Meanwhile, in resource constrained environments, honest broker teams must identify process improvement opportunities that balance efficiency and customized service to different types of customers.

Generalizability of a Sepsis Machine Learning Algorithm Across a Multi-Hospital Healthcare System

Sunny Mahesh, MBI1,2, Sean Yu, MS1, Aditi Gupta, PhD1, Albert M. Lai, PhD1, Marin H. Kollef, MD1,3, Philip R.O. Payne, PhD1, Andrew P. Michelson, MD1,3
1Institute for Informatics, Washington University in St. Louis, St. Louis, MO
2Saint Louis University School of Medicine, St. Louis, MO
3Division of Pulmonary and Critical Care, Washington University in St. Louis, St. Louis, MO

Introduction
Sepsis is an acute life-threatening condition where a dysregulated host immune response to infection causes multi-organ dysfunction. It contributes to about one in every three inpatient deaths and costs the US healthcare system $38.2B annually. Early and effective therapy is the cornerstone of management and machine learning (ML) is increasingly recognized as a useful tool for early identification of sepsis, but generalizability from one healthcare system to another remains poor, preventing widespread use. Today, many hospitals are part of larger healthcare systems that share similar, but distinct patient populations. The performance degradation seen when ML models are ported between hospitals within the same healthcare system, however, remains unknown. Therefore, the objective of this analysis was to determine the degree of performance loss when a sepsis model trained on data from a large academic medical center was cross applied to within-network community hospitals.

Methods
All patients ≥18 years of age admitted to a large healthcare system (1 academic and 11 community hospitals) between 1/1/2015 and 12/31/2020 were eligible for inclusion. Encounters were excluded if admitted to Psychiatry or Obstetrics services, due to highly variable rates of physiologic data collection, there were no billing code, vital sign, laboratory, service, room, or medication data to indicate a complete hospitalization. Sepsis-3 criteria were evaluated on general ward patients and only the first occurrence per encounter was included. The models were designed to predict sepsis within 6 hours of onset using Sepsis-3 criteria. Onset was defined as the earlier of culture collection or antibiotic administration. An eXtreme gradient boosted decision tree was trained using a 75:25 train:test split, before 5-fold cross validation. Models were trained on academic encounters and validated on community encounters, then trained and tested on both. Missing data imputation was not performed.

Results
In total, 71,711 encounters met inclusion criteria; 67.5% (48406) were academic (Table 1). Academic patients were younger (61.8[50.5-71.6] vs 68.3[56.2-79.4]; p< 0.01) and had higher comorbidity scores (10.0[2.0 - 20.0] vs 8.0[0.0 - 17.0]; p< 0.01), and rates of in-hospital mortality (1.3% vs 1.0%; p< 0.01). The MLM trained on the academic data and validated on community data had an area under the receiver operating characteristic curve (AUROC) and area under the precision recall curve (AUPRC) of .875[.867-.882] and .295[.272-.318], respectively. Both the AUROC and AUPRC were significantly lower when cross applied to the community hospital encounters (AUROC .853[.838-.868]; p<0.001 & AUPRC .196[.166-.229]; p<0.001).

Conclusion
Overall, the academic hospital population was younger, but had higher comorbidity scores with higher inpatient mortality compared to the community hospital population. The MLM trained on the academic encounters and validated on the community encounters had a smaller, but significant decline in performance. However, considering the difference of demographics between cohorts, this demonstrates the potential of generalizability of such models. Further studies are needed to better quantify the etiology and degree of discrepancy and analyze multiple variations of train, validate, and test splits of cohorts for improving generalizability.

References
Creating a data repository of sociomic factors to further characterize clinical outcomes and disease progression in patients with asthma

Sonya Makhni, MD MBA, Bruno Barbarioli, MS MS, Sanjay Krishnan, PhD, Jonathan Ozik, PhD, Kevin Lazenby, Marc Sleiman, Ankur Bhargava, MD, MPH, Julian Solway, MD, Ellen Cohen, MPP, Samuel Volchenboum, MD, PhD, MS
1University of Chicago Medicine, Chicago, IL

Background
Beyond focusing on and treating biological mechanisms of disease, advancing health also requires addressing the adverse consequences of social, environmental, behavioral, and psychological factors. Altogether, these factors comprise the entire lived experience, which we call the “sociome.” These sociome factors can impact clinical outcomes in a variety of ways, interacting with human biology to cause or exacerbate disease. Further, sociome factors have not been comprehensively codified and quantified in a way suitable for large scale co-analysis with biological and clinical data; such analyses would likely lead to new insights into factors influencing wellness or disease. There is a significant need for population-wide studies that effectively correlate sociomic data to clinical outcomes. The ultimate goals of this research are to (1) assemble and integrate publicly-available, proprietary, geocoded, and real-time datasets about social, environmental, behavioral, and psychological exposures experienced by children with asthma who live on the South Side of Chicago, (2) to assess how environmental factors resulting from health disparities impact clinical outcomes of asthma and (3) build predictive models using sociome data in conjunction with clinical data from the University of Chicago electronic medical record (EMR), which can ultimately help inform practical and actionable long-term and immediate interventions to minimize asthma severity and exacerbations. The primary aims of this project will be to create the data commons as a service (1) so that we can later assess the clinical impacts of environmental racism (2) and build relevant and actionable predictive models (3). Creating a data common as a service infrastructure for sociome data is a unique and innovative approach to the problems of combining large amounts of non-clinical data with patient information and making these data usable in multiple contexts by many investigators.

Methods
Here we assembled publicly available data sets that include census data, crime, green space, building permits, vacant and abandoned buildings, business licenses, traffic (the City of Chicago data portal), pollution and weather (the National Oceanic and Atmospheric Administration), and noise (Array of Things project). These datasets not only reflect factors known to influence asthma outcomes (e.g., pollution), but also variables that – when combined together – represent the lived experience. We additionally placed a local instance of the freely available Pelias geocoder on the UChicago Center for Research Informatics HIPAA-compliant infrastructure. The UChicago Clinical Research Data Warehouse will be leveraged to obtain clinical information for children diagnosed with asthma at UChicago Medicine between 2007 and 2021, which will include clinical outcome data in addition to information pertaining to race, ethnicity, insurance information, and other variables related to social determinants of health. The home addresses of each child (who over time might live at multiple locations) will be subjected to geocoding, and this information will be aligned with imported contemporaneous sociome data. A model will be built to account for each sociome element’s contribution to asthma outcomes.

Description
Here we are creating sustainable and scalable ways for collecting, standardizing, and sharing real-world sociome data, simultaneously linking those data back to patient information. Through this work, we aim to demonstrate feasibility of a data commons-as-a-service for clinical and sociome data and to provide technical specifications and descriptions of processes employed. Creating generalizable and scalable infrastructure to support research of social and environmental impacts on clinical outcomes is critical for efficiently collecting, standardizing, and sharing real-world sociome data that can be integrated with clinical outcomes and variables. Our work will provide a framework to be used in other disease states. Further, this infrastructure will facilitate the application of advanced analytical tools and visualization platforms that will enable researchers to vet hypotheses, refine study criteria in real-time, facilitate sharing of data and research tools, and ultimately accelerate discovery of new insights into sociome factors influencing clinical outcomes of disease.

References
Convergence Towards a Unified Asynchronous i2b2/SHRINE User Experience
Anupama Maram, MS¹, Griffin M. Weber, MD, PhD¹, Douglas MacFadden, MS¹
¹Harvard Medical School, Boston, MA

Introduction
The Shared Health Research Network (SHRINE) development team at Harvard Catalyst at HMS unveiled a modern user interface for the SHRINE web client in May 2020, seen in Figure 1. The data driven and user centric focus design was achieved through use of needs and landscape analysis, wireframe iterations, focus groups, the i2i2/iS user interface working group. The release was a success, and was accepted at 54 national network sites and allowed researchers to search for patient cohorts among greater than 130M+ unique patients.

Discussion
SHRINE is a federated tool that simultaneously queries the i2b2 systems at multiple institutions in order to find patients to recruit for clinical trials. Qualified researchers can query the EMR data of all sites on the network. Researchers have successfully used the network for hypothesis validation, cohort identification, public health surveillance and as the starting point to in depth informatics analysis. Informatics for Integrating Biology at the Bedside (i2b2), based at Partners HealthCare System, is a scalable informatics framework designed to bridge clinical research data and the vast data banks arising from basic science research in order to better understand the genetic bases of complex diseases. With the success and close dependency and reliance of these two platforms, the two platforms are working to converge on a singular UI.

The technical architecture of the two platforms poses a challenging integration to fit the SHRINE UI onto the i2b2 web client. i2b2 web client is a collection of client-side components designed as an YUI AJAX based plug-ins that communicate with i2b2 cells and allows investigators to query and display the data of the hive. AJAX stands for asynchronous JavaScript and XML (extensible markup language). The SHRINE UI is based on a completely different technical stack which includes a JavaScript based React library to interpret the XML and allows interoperability between the two platforms.

Conclusion
The success of both applications in the open source informatics community is well known. Integration the two platforms into one cohesive user experience will pose significant technical and user design tradeoffs, but the benefit greatly outweighs the risks. Convergence of the two applications will streamline open source informatics technologies and can help facilitate rapid information and data sharing to identify patient populations, locate patient sets in both local and geographically diverse areas, while maintain patient confidentiality.

References
3. i2b2. [Internet]. 2021 Aug 20. Available from https://www.i2b2.org/
Consumer Attitudes Toward Vaccine Credentialing Systems

Mollie M. McKillop, PhD, MPH¹, Bedda L. Rosario, PhD¹, Gretchen P. Jackson, MD, PhD¹,²
¹IBM Watson Health, Cambridge, MA; ²Vanderbilt University Medical Center, Nashville, TN

Introduction
Vaccine credentialing systems or digital health passports provide validated health credentials, such as proof of vaccination. Tracking and provisioning of a digital pass, which vaccinated individuals can show to transportation officials, restaurants, bars, employers, or entertainment venues, demonstrates who is ‘safe’ for entry, potentially improving control of disease spread and providing incentives for vaccination. Success of these systems depends on consumers’ willingness to share their vaccination status. Yet, vaccine passports are controversial, and consumer attitudes remain unknown. We sought to assess consumer views on vaccine passports through a nationally representative survey and to examine how socio-demographic characteristics may influence these attitudes.

Methods
We conducted an IBM Watson Health PULSE® survey in April and May of 2021 to gauge attitudes and opinions of consumers. Participants were recruited through a third-party market research firm. We asked questions related to vaccination status, plans to receive the vaccine, concerns about the vaccine, concern about getting COVID-19, current preventative behaviors, and comfort with a vaccine credentialing system. We also collected information on age, annual income, and education level. U.S. Census data were used to weight our results to be nationally representative. Percentages for each response and corresponding 95% confidence interval were calculated overall and by socio-demographic variable. To determine statistical significance, we compared 95% confidence intervals between each socio-demographic category variable and the total overall survey results. The significance level was set to $p < 0.05$.

Results
A total of 3,005 and 3,002 participants completed the survey in April and May, respectively. Just over half (54% in April and 55% in May) reported feeling comfortable using a vaccine credentialing system. This rate tended to increase with education level and annual income. Table 1 summarizes the most current results captured in May 2021.

Table 1. Summary of Vaccine Credentialing System Questions

<table>
<thead>
<tr>
<th>Age group (years)</th>
<th>Response to “Are you comfortable joining a vaccine credentialling system that would allow you to prove your vaccinated status?” (95% confidence interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;35</td>
<td>Yes (51.5% (48.0%, 54.9%))</td>
</tr>
<tr>
<td>35-64</td>
<td>Yes (56.6% (54.0%, 59.1%))</td>
</tr>
<tr>
<td>65+</td>
<td>Yes (55.9% (52.1%, 59.7%))</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Education Level</th>
<th>Response to “Are you comfortable joining a vaccine credentialling system that would allow you to prove your vaccinated status?” (95% confidence interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High School Level or less</td>
<td>Yes (44.1% (40.4%, 47.7%) *)</td>
</tr>
<tr>
<td>Some College/Associate</td>
<td>Yes (48.4% (45.0%, 51.8%) )</td>
</tr>
<tr>
<td>College+</td>
<td>Yes (64.6% (62.1%, 67.1%) *)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Annual Income</th>
<th>Response to “Are you comfortable joining a vaccine credentialling system that would allow you to prove your vaccinated status?” (95% confidence interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;$25K</td>
<td>Yes (44.4% (40.4%, 48.5%) *)</td>
</tr>
<tr>
<td>$25k - $49.9k</td>
<td>Yes (53.3% (49.9%, 56.7%) )</td>
</tr>
<tr>
<td>$50k-99.9k</td>
<td>Yes (55.8% (52.6%, 58.9%) )</td>
</tr>
<tr>
<td>$100k+</td>
<td>Yes (70.0% (65.9%, 73.8%) *)</td>
</tr>
</tbody>
</table>

* Indicates statistically significant associations

Discussion and Conclusion
Digital vaccine credentialing systems are controversial, but just over half of Americans surveyed report being comfortable using one. Increased annual income and education level were associated with favorable responses to these systems. As digital passports are implemented, factors affecting adoption should be examined to avoid a digital divide.
A Generalized Approach to Assessing the Gaps between Information Needs and Information Readiness when Planning Informatics Solutions for Clinical guidelines

Meenakshi Mishra, MSc., MPH¹, Dana M. Womack, Ph.D., RN¹, Mary Pickett, MD¹, Nicole G. Weiskopf, Ph.D.¹

¹Oregon Health & Science University, Portland, OR, USA

Introduction

The 2016 safe opioid prescribing guidelines by the Center for Disease Control and Prevention (CDC) recommends that Primary Care Providers (PCPs) assess patients’ risks for Opioid Use Disorder (OUD) when prescribing opioids for chronic non-cancer pain. However, weighing risk versus benefit in this complex patient population is a notable clinical challenge. While many OUD-related factors are known, not all factors may be useful for OUD decision-making in this complex patient population, or furthermore, be present in the electronic health record (EHR). The potential gap between PCP’s information needs and EHR’s information readiness remains largely unexplored. We propose a systematic approach to identify and assess this gap to inform future informatics projects.

Method

We collected fifty-seven patient-specific risk factors for OUD from literature review and expert input and grouped these under nine different concepts: demographics, substance use, psychiatric, socioeconomic, pain and function, medication, aberrant drug-related behaviors, medical comorbidities, and genetics. A Delphi-style survey is underway to identify patient-specific risk factors that PCPs find most useful for determining OUD risk. PCPs who manage patients on opioids for chronic non-malignant pain in various primary care settings have been invited to participate in the survey. The usefulness of patient factors in determining their OUD risk is being collected on a 5-point Likert scale, from not useful to extremely useful. Consensus is defined as >70% of participants finding a patient factor as very useful or extremely useful and >25% of participants finding a patient factor as not useful.

Results

Complete Delphi Phase 1 responses from 47 PCPs were analyzed for consensus regarding the usefulness of various patient-specific information for determining OUD risk. A large proportion of the survey participants routinely manage patients on opioids for chronic non-cancer pain, (66%), and close to 35% of these participants are involved in research in the area of opioid prescribing and/or policy development at the local, regional or national level. Approximately 55% of the participants received CME dedicated to pain management in the past 2 years.

Table 1. Consensus post first round of survey

<table>
<thead>
<tr>
<th>Very useful or extremely useful</th>
<th>Substance Use</th>
<th>Psychiatry</th>
<th>Socioeconomic</th>
<th>Medication</th>
<th>Aberrant drug-related behaviors</th>
</tr>
</thead>
<tbody>
<tr>
<td>I illicit drug use, current substance use disorder, history of a substance use disorder, history of misuse of sedative or stimulant, history of misuse of cold and cough medication, history of non-fatal overdose</td>
<td>History of suicide attempt,</td>
<td>History of DUI or drug-related conviction</td>
<td>Total opioid dose greater than 90MME, concurrent Benzodiazepine prescription,</td>
<td>Resistant to change in opioid medication, reporting prescription loss or theft, obtaining opioids from multiple providers, increasing dose without provider instruction, requesting early refill, showing symptoms consistent with opioid withdrawal, obtaining opioids from multiple pharmacies, being in a hazardous situation due to opioids, weaning described as unsuccessful or difficult, ED visit to obtain opioids, abnormal UDS, requesting higher dose of prescription opioids, multiple phone calls for an opioid refill, taking opioids for symptoms other than pain.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Not useful</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
</tr>
<tr>
<td>Medical comorbidity</td>
</tr>
<tr>
<td>Genetics</td>
</tr>
</tbody>
</table>

Conclusion

Results from the first round of the Delphi survey indicate that most survey participants are in agreement regarding the usefulness of drug-related behaviors to determine OUD risk. There is an agreement with most risk factors in the CDC guidelines, though mental health factors are not considered most useful, contrary to the high risk for OUD associated with these factors. We will confirm the agreement in the second phase of the Delphi survey and our future work will assess EHR data readiness for the most useful patient information.

Acknowledgment – NLM Award # T15LM007088; Jonathan Robbins, and Dr. Ben Sanders
Weak supervision reduced need for manually annotated data for training a named entity recognition model to recognize pharmaceuticals

Hui-Shuang, Hsu, Julie Mocko, PhD, Xiaojian, Jiang, PhD, Kang-Ling, Hsieh, PhD, Sam Regenbogen, PhD

1Mercury Data Science, Houston, TX, USA; 2School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

Introduction: Deep learning model learns more features from a large training dataset to achieve higher performance than a small dataset. However, building a large training dataset for a deep learning model is labor-intensive and time-consuming. Weak supervision is a technique that can reduce the burden of obtaining hand-labeled data. It uses noisy or imprecise sources to provide supervision signals for labeling large amounts of training data in a supervised learning setting. Snorkel, a widely used weak-supervision framework, can automatically aggregate those intercorrelated signals to infer sentences that contain a drug name. In this study, we developed a pipeline that utilizes the weak supervision model to increase the training dataset size to improve the performance of the named entity recognition (NER) model.

Material and Methods: Our purpose is to verify the performance of the NER model that has trained the larger weak labeled dataset generated by Snorkel is better than feeding the small hand-labeled dataset. We selected 50,000 PubMed abstracts that contained mentions of pharmaceuticals (drug names) and split each abstract into sentences. We randomly annotated 283 sentences as the gold standard dataset. We applied a comparative toxicogenomic database (CTD) as the knowledge source to check whether the sentence included the corresponding drug names. Additionally, we used the frequency of drug names and prescription units as additional rules to increase the confidence of the sentences. We selected the sentences with above 0.93 precision scores as a high-confident dataset. The high-confident dataset was used to train the NER model compared with the baseline model trained only with the gold standard dataset. We also tested and compared different pre-train language models.

Results: We used the gold standard dataset as the baseline to train the spaCy NER model. The baseline result were precision: 0.84; F1: 0.77; recall: 0.72. Then we fed the larger dataset generated by Snorkel. The “en_core_sci-lg” had better performance than other language models. The precision were: 0.78, F1: 0.82; recall: 0.86. (Figure 1).

Discussion: Although the larger, noisy dataset-based model presented inferior performance than the small, precise dataset model, the former contains more information for the model to learn. Our model detected most of the pharmaceutical compounds as drugs. However, some chemical compounds referred to those naturally present in vivo rather than a type of treatment. For example, oxygen can be a type of treatment for Covid-19 or a basic element in the body.

Conclusions: The “en_core_sci-lg” model showed better performance than other pre-train models. Pre-train models plus a high-confident dataset generated by Snorkel had a better F1 score compared to the baseline model, except for the en_core_sci-scibert. The increasing F1 score is mainly from recall rather than precision. The results indicated that Snorkel can quickly generate a larger dataset for training, but it still contains some noise.

References
ICU Census: The Critical Care Management System for an Oncologic ICU
Thomas A. Nahass, MD1, Yekaterina D. Tayban, NP1, Gleb Kirminicci1, Mark Knott1, Meghana Mehta1, Neil A. Halpern, MD1
1Memorial Sloan Kettering Cancer Center, New York City, New York, US

Introduction
Management of the various aspects of critically ill patients requires precision and attention to detail. The ICU Census (“Census”) is a custom database system for the management of logistical, clinical, and reporting mechanisms as they relate to patient care in a specialized oncologic medical-surgical ICU. We report the design and implementation of the Census.

Design
The Census is a MS SQL backend with an ASP.NET frontend, with various .NET windows service connections to a multitude of interfaces (Figure 1). ICU admissions, consults, rapid responses, inter-hospital ICU transfers are tracked (Figure 2) and recorded through the standardized forms that can then be viewed in aggregate through the ICU Greaseboard located on screens throughout the ICU and any computer logged into the secured network. The data that is collected during a patient’s ICU stay ranges from procedures or interventions to novel clinical trial therapeutics for a patient’s cancer care. These items are retrospectively reviewed for research on critically ill cancer patients.

Reporting
Each morning, an aggregated rounding report for each ICU team is automatically produced as a PDF document and sent to the ICU team (Figure 3). This report organizes the lab, medication, and procedural data by organ systems for each patient. There are multiple ICU, respiratory therapy, and pharmacy workflows that utilize the Census to aggregate data in Tableau Dashboards (Figure 4). Additional ICU reports for mortality, admission and discharge data, demographic data, and procedures can be easily produced.

Conclusion
The management of patient flow coordination and critically ill patients can be effectively managed through the ICU Census. Future plans include the automation of the manual inputs and release the design as open source for other ICUs that are interested in the design.

Figure 1: ICU Census Dashboard
Figure 2: Census Data Flow
Figure 3: ICU Rounding Report
Figure 4: ICU Rounding Report
SEE-Diabetes, a Patient-Centered DSMES for Older Adults: Information Needs of Patients and Providers

Ploypun Narindrarangkura, MS, MD¹, Uzma Khan, MD², Suzanne A. Boren, MHA, PhD¹,², Eduardo J. Simoes, MD, MSc, MPH¹,², Min Soon Kim, PhD¹,²

¹Department of Health Management and Informatics; ²Institute for Data Science and Informatics; ³Department of Medicine, University of Missouri, Columbia, MO, USA

Introduction

One in four Americans over the age of 65 years has diabetes mellitus. Diabetes self-management education and support (DSMES) can improve diabetic control and reduce complications. However, current studies focus on clinical decision-making regarding drug choices rather than DSMES. We are developing an educational decision aid, SEE-Diabetes (Support-Engage-Empower-Diabetes), that will enable people with diabetes, in consultation with their providers to choose a DSMES strategy. SEE-Diabetes will support patients and providers in making shared decisions by organizing knowledge and information of seven DSMES principles presented at appropriate times to promote better health. SEE-Diabetes will be included at the end of clinic notes under the “Patient Education” tab and updated via shared decision-making during each clinic visit. In this study, we conducted a survey and focus groups among providers and diabetes patients aged 65 and over to determine the information needs of SEE-Diabetes to be integrated into the clinic notes.

Methods

We conducted surveys of 42 providers and 37 patients and six focus groups with 13 providers and nine patients from family medicine and diabetes specialty care at the University of Missouri Health Care. Using REDCap, we collected demographics, the information needs of DSMES, and the accessibility of medical charts. In the focus groups, we invited a subset of survey participants to evaluate four clinic notes of diabetes patients. Questions for providers were Q1) Is the clinic note easy to read and why or why not? Q2) Does the clinic note add to health care information for people with diabetes? Q3) How to improve SEE-Diabetes? Questions for patients were Q1) Is the clinic note easy to read and why or why not? Q2) Is this note helpful for people with diabetes? Q3) Who do you think should discuss the DSMES principles with you? Q4) When would be the best time to discuss the DSMES principles? Q5) How to improve SEE-Diabetes? Focus groups were conducted and recorded via Zoom. Thematic analysis was used to analyze qualitative data.

Results

Mean age of providers was 52 (30–88, SD = 14), and 22 (52%) were female. Mean diabetes care experience was 23 years (3–57, SD = 12). Most (31, 83.8%) of patients’ providers were from diabetes specialty care. Mean age of patients was 66 (24–82, SD = 12) and 22 (60%) were female. The diabetes duration ranged from 0 to 63 years (M = 21, SD = 15), and some of patients (25, 68%) took insulin. Survey results showed that some providers (25, 60%) were familiar with general DSMES guidelines. Most patients (28, 76%) accessed their medical charts using the patient portal through a computer (18, 64%) and mobile devices (17, 61%). The majority of patients (29, 80%) have read their clinic notes.

Each focus group lasted approximately 1.5 hours. Thematic analysis revealed that the readability of clinic notes was their primary concern from both provider and patient perspectives. Common issues for reading the clinic notes were poor readability because of many medical abbreviations and poor formatting. Both groups also shared concerns of insufficient self-care information, addressing DSMES principles. Most patients wanted to discuss DSMES principles with providers, followed by diabetes educators, peer diabetes patients, family, and friends. The best time to discuss DSMES principles depends on their diabetes duration. Both groups provided suggestions to improve the SEE-Diabetes model, such as adopting SMART goal model - Specific, Measurable, Attainable, Relevant, and Timely, adding external diabetes care resources, and using motivational words for the statements.

Conclusion

This study investigated the information needs of SEE-Diabetes to be integrated into the clinic notes. Both groups agreed the readability was the main concern. The survey showed most patients have read their clinic notes and accessed their medical charts using the patient portal which shows a potential benefit of the proposed SEE-Diabetes. For future study, SEE-Diabetes will be written following NIH recommended readability level of 6th grade. We will also iteratively evaluate readability and usability as we develop SEE-Diabetes modules.
Predicting Suicide Attempt Amongst People with Diabetes Using Generalized Linear Model

Ploypun Narindrarangkura, MS, MD1, Patricia E. Alafaireet, PhD, MHA2, Uzma Khan, MD3, Min Soon Kim, PhD1,2
1Institute for Data Science and Informatics; 2Department of Health Management and Informatics; 3Department of Medicine, University of Missouri, Columbia, MO, USA

Introduction

Research shows people with newly diagnosed type 2 diabetes had the rate of suicide attempts two times greater than the general population. However, few studies have focused on understanding suicidal behavior among people with diabetes. The long-term goal of this study is to develop a clinical decision support system that healthcare providers can use to identify patients with diabetes who are at high risk of suicide attempt. This system will provide suicide prevention measures during outpatient clinic visits. In this study, we explored Cerner Real-World Data™ to investigate factors associated with attempted suicide in patients with diabetes.

Methods

First, a systematic review using MeSH search terms to identify known risk factors of suicidal behavior among people with diabetes was conducted. Second, a diabetes patient list from Cerner Real-World Data™, using ICD-9 and ICD-10. Cerner Real-World Data™ encompassing 1.2 billion de-identified patient encounters from 101 hospitals in which Cerner has a data use agreement was extracted. Data from HealtheDataLab™, a data science ecosystem built and deployed on Amazon Web Services, was accessed. Third, a logistic regression model using Generalized Linear Model (GLM) function was performed to determine associated factors and to predict the probability of suicide attempt. Missing values were replaced by the imputed values generated from Multivariate Imputation by Chained Equations.

Results

The database contained 67 of 125 known risk factors identified from the literature review (Table 1). Data from 4,202,946 people with diabetes was used for analysis. Of these, the analysis showed 12,738 (0.3%) people with diabetes and a documented suicide attempt. Their mean age was 45 (8-89, SD=15.9). The sample included 7,192 female patients (57%) and 9,125 identified as white (74%). More than half (6,339, 54%) had a reported BMI ≥30 kg/m². Most of the patients (12,174, 96%) had a recorded diagnosis of type 2 diabetes and had been diagnosed less than 5 years (N=9,939, 78%). Of 67 variables, 50 variables were statistically significant (p<0.05). Very significant risk factors that increased the odds of a suicide attempt among people with diabetes have shown in Table 2. The accuracy of the logistic regression model using GLM was 99.7%.

Conclusion

This preliminary study suggested a higher number of diabetic complications were significant associated with suicide attempts among people with diabetes. This knowledge could be useful to help healthcare providers identify a high-risk group. Limitations of this study include 1) missing patients and factors from using diagnosis codes and 2) incorrect races, ethnicity, timestamp, and ICD coding. Overall, the Cerner database was easy to work with because data were stored in a person-centered model with all tables connected to the person ID. There were several challenges using the database, such as limited access hours, slow execution, and unstable interface due to the limited memory space. Multivariate statistical analysis is complex and unable to explain a cause-and-effect relationship. Data mining for subgroup discovery is a promising way of reducing the complexity of features. This may be very effective in analyzing complex healthcare data. As the next step, research team plans to use contrast mining and subgroup discovery to identify the patterns of a high-risk group of suicide attempt among people with diabetes.
Women’s Health Postpartum Care and Continuity

Jayson T. Osika, MSc1, Pedro B. Carneiro, MPH1, John Gresh, MSc2, Raymonde Charles Y. Uy, MD, MBA1, Lisa Masinter, MD, MPH, MS, FACOG2, AnnMarie Overholser, MD3, Michelle Witter4, Julia L. Skapik, MD, MPH, FAMIA1

1National Association of Community Health Centers (NACHC), Bethesda, MD 20814; 2AllianceChicago, Chicago, IL 60654; 4OCHIN, Portland, OR 97201

Background

Community Health Centers (CHCs) care for some of the most vulnerable patients in the United States, including pregnant women in low-income and minority populations. Ensuring appropriate prenatal and postpartum care for these populations is a core strategy in the efforts of achieving health equity in the US. Disparities in pregnancy outcomes among minority women have been widely reported in published literature, making this a priority for CHCs. Obstetric care for health center patients involves a multidisciplinary team of CHC workers, hospital staff as well as other community-based organizations. To provide the best care across organizations and throughout the pregnancy period, pregnancy episodes should be appropriately documented in electronic health records. Identifying pregnancies and exchanging this information between electronic health record (EHR) systems however is severely limited due to lack of standardization of data for pregnancy status on top of the existing health data interoperability issues in the US health care system. A variety of methods were used to identify pregnancies from data aggregated from CHCs, demonstrating inconsistencies in obstetric data documentation.

Methods

We used a combined dataset from two large health center-controlled network (HCCN) organizations, totaling more than 1,129,000 pregnancy-aged women seen in their member CHCs. Business requirements were developed to streamline data extraction by utilizing curated pregnancy-related value sets. HCCN partner feedback helped overcome challenges with inconsistent definitions across organizations, such as what consisted of a postpartum visit. The initial approach involved identifying pregnancies using the presence of a real or estimated delivery date; or in the absence of these, other surrogate identifiers were utilized (diagnosis codes, recorded postpartum visits). Indicators of a pregnancy include both a counseling encounter type that is one of either an initial, subsequent, or return obstetric visit, as well as either the presence of a pregnancy diagnosis, or a postpartum follow-up.

Results

Overall, around 170,000 (15%) individuals were found to have active pregnancies according to the method utilized (Figure 1). The majority of pregnancies (99%) were identified through having a delivery date, most (99%) had an estimated delivery date, and a smaller proportion had the actual date of delivery (4%). Only one third (33%) of pregnancies were identified using another indicator, such as an ICD-10 diagnose code – over 90% of pregnancies identified with an indicator other than date also had a date (data not shown). Lastly, only 4% of all pregnancies were also identified as having delivered a viable fetus, among those with a delivery date only 3%.

Conclusion

Pregnancy documentation is highly inconsistent across organizations and between their respective health centers, making it difficult to semantically and syntactically aggregate these values with accuracy and confidence. None of the HCCN organizations during the studied period had the ability to send or receive meaningful pregnancy-related data to and from hospitals where most pregnancy deliveries occur. The use of delivery dates, specifically the estimated delivery date, seems to yield the greatest number of pregnancies – which a subset was confirmed via chart review. Methods not inclusive of such dates are likely to miss most pregnancies. The lack of interoperability between hospitals and CHC systems contributes to the inability to determine fetus viability. Tools that can support patient care, such as clinical decision support, rely on a valid, standardized set of common data elements, compliant with terminology standards and with adequate completeness. It is imperative that organizations move towards closing some of these workflow and interoperability gaps by supporting care teams with tools and resources needed to best support patients, document care, and connect health data across systems.
A Pilot Study of Public Health Messages on COVID-19 in Twitter

Braja G. Patra, PhD1, Soumik Mandal, PhD2, Sanchari Pal, PhD3, Evan Sholle, MS1, Jyotishman Pathak, PhD1
1Weill Cornell Medicine, New York, NY, USA; 2NYU Langone Health, New York, NY, USA; 3Jadavpur University, Kolkata, WB, India

Introduction: COVID-19 has infected more than 213 million people globally as of August 25, 2021 (https://coronavirus.jhu.edu/map.html). During this unprecedented crisis, people have increasingly relied on online information sources for dealing with the pandemic. Social media holds considerable potential to change the way we communicate, fundamentally altering society and with it, the landscape of health information dissemination. COVID-19 pandemic is a rapidly evolving situation which includes new strains of the virus, vaccination, confirmed cases, death rates, preventive measures, travel restriction, COVID-19 stigma. A real-time assessment of Twitter discussions can be useful for timely addressing of public health emergency response. In this direction, we sought to use natural language processing techniques to classify information disseminated through Twitter related to the surveillance and prevention of COVID-19. We hope such automated classification approach will help users to get relevant information in a timely manner.

Data: The data used in our experiments were tweets disseminated by verified Twitter accounts (✓) between December 25, 2020 to April 01, 2021 and were obtained using tweepy api(https://docs.tweepy.org). The official Twitter accounts used to mine the tweets include CDCemergency, CDCFlu, CDCgov, and CDCtravel. We collected 25,829 unique tweets and annotated 3,428 tweets with binary labels based on whether a tweet was related to COVID-19 or not. Out of these, 1,872 tweets pertaining to COVID-19 were further labeled with the following five categories: ‘outreach/outbreak (588),’ ‘symptoms (161),’ ‘prevention (564),’ ‘travel advisory (364),’ and ‘vaccine (195).’ Examples: Outreach/Outbreak: The US just saw its lowest Covid-19 daily case count since October... Symptoms: A new @CDCMMWR finds that adults who had #COVID19 can develop a condition similar to multisystem inflammatory syndrome in children and have severe outcomes including requiring intensive care. Prevention: Handwashing, along with wearing a mask and staying 6 feet from others, are key to slow the spread of #COVID19... Travel Advisory: Travel increases your chances of getting and spreading #COVID19. Staying home is the best way to protect yourself and others from getting sick. Vaccine: #DYK? You need two #COVID19 mRNA vaccine doses to get the most protection from COVID-19... We used topic modelling (latent dirichlet allocation) for creating initial set of lexicons for each category and then manually filtered to develop final lexicons.

Methods: Several embedding approaches were used in our experiment, e.g., TF-IDF, LSA, LDA, and word2vec. Our experiments with TF-IDF (parameters: max df=0.95, min df=2, ngram range=(1, 2)) obtained superior performance. Preprocessing and stopword removal were performed before creating vectors. We also used five features for cosine similarity of lexicon vector and each tweet vector. Initially, a Support vector machine (SVMs) based classifier was used to determine the tweets’ class labels, i.e., whether related to COVID-19 or not. Next, we developed individual classification model for classifying tweets into COVID-19 related categories as described in the data section. Additional experiments with an adaptation of COVID-TWITTER-BERT1 model was conducted for the same classification task. However, the last model obtained less accuracy than the system developed using TF-IDF and SVM.

Results and Discussion: The binary classification system obtained an f-measure of 0.89 for identifying tweets related to COVID-19 or not. The overall f-measures were 0.70, 0.92, 0.82, 0.88, and 0.91 for outreach/outbreak, symptoms, prevention, travel advisory, and vaccine. In future, we plan to extend this work with regular COVID-19 tweets (verified and unverified) and evaluate how the social media discussion can be useful for easy access of specific public health messages. Further, these tweets can be summarized to provide relevant information to users using a web dashboard.

References
Data Visualization to Optimize RBCX Procedures in Sickle Cell Disease

Perumbeti A\textsuperscript{1}, Shah S\textsuperscript{2}, Hixson M\textsuperscript{3}, Ortiz P\textsuperscript{3}, McCalley S\textsuperscript{3}, Vaidya V\textsuperscript{3}

\textsuperscript{1}Lab Informatics, Division of Pathology & Laboratory Medicine, Phoenix Children’s Hospital; \textsuperscript{2}Pediatric Hematology, Division of Pediatric Hematology Oncology, Phoenix Children’s Hospital; \textsuperscript{3}IT Integration & Data, Phoenix Children’s Hospital.

Background:
Sickle cell disease (SCD) occurs in approximately 1 out of 365 African Americans. Some patients with SCD do poorly, with morbidity including stroke, frequent pain episodes, multiple hospitalizations, organ damage, and shortened lifespan. The care for these patients costs the healthcare system ~3 billion dollars/year in the United States (Mainous III, Value in Health, 2018). One method to reduce stroke risk and disease severity for some patients with SCD is chronic red blood cell exchange (RBCX). We developed a clinical decision support tool with Microsoft Power BI for visualization of RBCX inputs and patient outputs, with the purpose of optimizing RBCX procedures to reduce anemia and sickle cell levels in outpatient chronic red cell exchange. This would reduce the probability of stroke (Lee, Adams et al, Blood 2006). Additional purpose was to identify refractory iron overload.

Description of Intervention(s):
Optimal RBCX procedures to reduce SCD complications are parameter driven and thus defined as procedural inputs and clinical/lab outputs. These parameters were mapped and passed to Microsoft Power BI chart and text visualizations with each patient having a series of visualizations to monitor RBCX therapy. (Figure 1) Dashboard was optimized with user feedback.

Metrics (Measure and Analyze):
Patients are reviewed with the RBCX dashboard in weekly apheresis clinical meeting and by individual physicians seeing the patients, for clinical decision support for RBCX. There were 10 distinct users and 43 days utilized, including all apheresis physicians. The trended values include circulating amounts of sickle RBC, level of anemia, hemolysis and erythropoietic drive, ferritin levels, and frequency of hospitalizations. The visualized RBCX procedure elements include baseline and target hematocrit, packed red blood cell transfusion volume, and fraction of original red cells remaining (FCR) following RBCX.

Impact:
Post-implementation data with patients who met criteria of RBCX procedures 6 months prior and post-implementation, shows an 18% improvement of hemoglobin > 8g/dl, along with a 13% decrease HbS levels (n=8, p < .05), with no significant change in RBCX intervals in this population (p=0.34). Additionally, post-implementation, six SCD patients were flagged with ferritin>1000ng/dl as a measure of refractory iron overload and were discussed with apheresis and hematology teams.

To the best knowledge of the authors, the RBCX tool is a unique and best in class RBCX decision support tool to optimize RBCX therapy and unique to Phoenix Children’s Hospital. Next steps are development of global quality metrics, refining procedural and clinical correlation, and detecting patterns in procedure responses in individuals and clusters.

Figure 1: Red cell exchange dashboard. Top row displays sickle hemoglobin & sickle cell trends, second row displays anemia & hemolysis trends, and third row displays iron overload and procedure parameter trends.
Oral Cancer Diagnosis Prediction Using Dental Pathology Notes

Malvika Pillai, PhD¹, Di Wu, PhD¹
¹Carolina Health Informatics Program, University of North Carolina, Chapel Hill, NC

Introduction

In 2020, the World Health Organization estimated that there were over 400,000 new cases of oral cancer worldwide. Globally, the age-standardized disability-adjusted life years (DALYs) was estimated at 64.23 per 100,000 people (1). In the United States, oral cancer accounts for about 3% of all cancer incidences with a 66.2% 5-year relative survival. Due to its sensitive location, treatment is often challenging, but morbidity and mortality rates improve with early diagnosis. Patients are typically screened for abnormalities associated with oral cancer by their dentist. The screening consists of taking a patient history and identifying risk factors (e.g., tobacco use, human papilloma virus (HPV)), signs and symptoms of oral cancer (2). If any abnormalities are detected, the patient is referred to a specialist for a biopsy, which is assessed by an oral pathologist who will make a diagnosis. In the diagnosis process, the pathologist records microscopic descriptions for the sample as well as a clinical impression. The microscopic descriptions are objective descriptions of the sample, while the clinical impression can be definitive (e.g., patient has squamous cell carcinoma) or unclear (e.g., squamous cell carcinoma versus oral candidiasis). There are various presentations of cancer in the oral cavity, making interpretation of samples to inform a clinical impression difficult. In this study, we aim to use the microscopic descriptions written by oral pathologists to estimate the probability of a sample being cancerous.

Methods

Data was collected from the Oral and Maxillofacial Pathology Lab (OMPL) at the Adams School of Dentistry from the University of North Carolina (UNC) at Chapel Hill. The Adams School of Dentistry has the largest OMPL in North Carolina with oral pathology samples and dental electronic medical records (dEMR) from 2005-2019. The dataset used in this study consists of 109,848 records collected from 01/07/2005 to 01/27/2020 from 95,681 patients. 5,048 records with cancer mentions were extracted with Unified Medical Language System (UMLS) mapping, and 22% of the set was labeled by a clinician as non-cancerous, cancerous, or showing pre-cancerous symptoms.

In this work, we developed a supervised classification approach using clinician labels as the gold standard. Records labeled as cancerous were considered positive, and those labeled as non-cancerous were considered negative. The training data of positive and negative cases was divided with a 70/30 split. Records labeled as showing pre-cancerous symptoms were held out as a test set. For data transformation, two approaches were taken: a spaCy pipeline with term frequency-inverse document frequency (TF-IDF) and SciBERT. The spaCy pipeline consisted of using part-of-speech tags, lemmatization, dependency parsing, and named entity recognition.

Preliminary Results

As the project is in beginning stages, the test set comprised of records indicating pre-cancerous symptoms has not been used to evaluate the model. 10-fold cross-validation was used to train the models, and four metrics were used to evaluate performance: accuracy, precision, recall, and area under the receiver operating characteristic curve (ROC-AUC) score (Table 1). Table 1 displays a comparison between using the spaCy approach and using SciBERT prior to model prediction with random forest, SVM, and logistic regression. Between using spaCy or SciBERT, models using the spaCy pipeline performed better than they did using SciBERT.

Conclusion

We present an approach to predict whether oral pathology samples have high probability of being cancerous or not to aid pathologists in deciding to refer a patient to an oncologist or not. The project utilizes unstructured pathology data to streamline the referral process for cases with clinical features that may not traditionally be associated with cancer. Further research in the project will consist of connecting patient dental records to clinical records to understand whether patients with pre-cancerous symptoms in dental records later developed oral cancer. Linking the two EMR systems will provide ground truth for the test set and enable us to evaluate performance on the test set.

Acknowledgement: First author funded by National Library of Medicine through grant number 1-T15-LM012500.

References

Developing a Care Management System for Mental and Behavioral Health

Saeid Pournejati, MS¹, David C. Hendricks, BS¹, Julie W. Crews, BS¹, Sina Madani, MD, PhD²
¹Visualutions Inc, Houston, Texas; ²Vanderbilt University Medical Center, Nashville, TN

Introduction

With the emergence of pay-for-performance incentives, behavioral health providers must create a personalized care management plan for individual patients. Measuring such outcomes has shown some improvements in the quality of care¹. In addition, while the types of care provided based on the patient’s signs & symptoms vary, defining a traceable care strategy for each patient that helps with sharing effective feedback with the patient can improve the overall mental health treatment strategies². However, clinical data entry and extraction often lead to provider's frustration and burnout. This study aimed to determine the critical components needed to define a traceable care and treatment plan for mental health patients that can address providers' concerns as well.

Methods

Our team piloted the software in a behavioral health clinic setting. We picked Scrum as an established agile software development methodology to design the clinical and legal requirements of the software and evaluated and verified the quality and effectiveness of our design through a series of rapid prototyping sessions. The clinical team from the pilot clinic reviewed the design and helped identifying various components. In addition, we identified opportunities for analytical dashboards that can assist with tracking patients' progress, missing documentation, and legal compliance.

Results

Configuration Component: This component allows the clinic to define the information needed to manage a care plan. We included the type of needs for each patient, treatment objectives list for a given need that requires to be met, and possible interventions that clinicians can choose to address the need using predefined constraints. E.g., the system encouraged clinicians to document objectives using a predefined value set.

Care Plan Component: "Care Plan" is a form that allows clinicians to create or revise a mental health care plan strategy for a patient (Figure 1). Clinicians can choose the care team assigned to a patient and allocate a legal policy for each team member, including the necessity of capturing the care team member's signature after each patient's visit. Providers can pick one or more need(s) for patients while documenting details for such needs in free text. For each need, clinicians select appropriate objectives that they expect to meet. Ultimately, the provider completes the care plan definition by associating each need with an identified problem coded with ICD10CM in the EMR and defining an intervention strategy. The interventions have a start date, duration, and daily/weekly/monthly/annual occurrence.

![Figure 1. A sample screenshot showing how to create an anxiety care plan for a patient.](image)

Progress Note Component: This component is a structured form emphasizing recording patient's progress collected by the provider for each identified need and intervention. The user interface allows tracking the status of the selected objectives for each patient's need. In addition, providers are asked to select the status of each need from a set of predefined states (In progress, Pending, Completed, Revised, and Discontinued).

Conclusion

Our analysis led to the implementation of a configurable care management software. We plan to expand our analytical dashboards in a more provider-friendly way. A successful visualization of the collected data would allow mental health researchers and clinicians to identify primary mental health needs and providers' success in addressing such needs.

References

Assisting Exercise Survivorship Data Curation with a Computer-Vision- and Optical-Character-Recognition-Supported Web Application

Zach Rachlin, M.S.¹, Whitney Underwood, B.S.¹, Lee Jones, PhD¹, John Philip, M.S.¹
¹Memorial Sloan Kettering Cancer Center, New York, NY

Introduction
A primary focus of the Lee Jones Lab, as part of the Exercise Oncology Service at Memorial Sloan Kettering Cancer Center (MSK), is on the association between postdiagnosis exercise and all-cause mortality in patients with primary cancer. Studies to date on this topic have been limited in terms of the types of primary cancers examined; a lack of focus on the relationship between exercise and recurrence; small sample sizes and historical data points per patient; and focus on only one cancer population at a time. The MSK retrospective research protocol 20-101, “Patient-Reported Outcomes, and Clinical Events in Adult Cancer Survivors” attempts to address these limitations by leveraging exercise data collected as part of a standard intake assessment offered at every visit in the adult cancer survivorship clinics at MSK since 2012, encompassing approximately 25,000 patients. The challenge with this dataset is that several years of the intake assessments were filled out by hand, so the patient-reported answers to the exercise questions are in unstructured, handwritten text which was then scanned as an image. There was a desire to reduce the time needed to curate all of this data, so a data processing pipeline was created to automatically locate, crop, and attempt to structure the patient reported responses about their exercise habits. In addition, a web application was created to present the automated results alongside the data entry forms and the original assessment image, making the task of curation much easier and time efficient.

Methods
Over 1200 images were manually reviewed and split into two classes: those that contained the relevant exercise questions (Exercise Images) and those that did not (Non-Exercise Images). From that dataset, the 345 Exercise Images were labeled for object detection with bounding boxes around both the region containing the exercise question text and the handwritten patient answers, and the sub-region containing just the handwritten answers. A model using the MobileNetV2 convolutional neural network (CNN) architecture was frozen for feature extraction and then fine-tuned with a binary classification head to distinguish between Exercise Images and Non-Exercise Images. A model using the RetinaNet architecture was fine-tuned to classify and bound the aforementioned regions on Exercise Images. The de-identified, cropped images were sent via the Google Cloud Vision API to their handwritten text model, which returned the recognized text and confidence scores. Data curators were presented with a data entry form for each patient intake assessment, alongside the suggested answer from the pipeline above, as well as the cropped and full images of the relevant page of the assessment. There were quick-add buttons to use all or one of the automated suggestions, to convert a suggestion from hours to minutes, and to flag an assessment for further review. All data was written upon submission to a REDCap database via API calls.

Results
The binary classification CNN narrowed down the number of individual survey pages that needed to be examined from over 100,000 to 51,592 by surfacing only the pages that contained the relevant patient-reported exercise questions. The web application enabled the curators to view the relevant survey pages right alongside the data entry forms, saving them the time of having to manually access them via the EMR. 47,783 assessment curations were completed in less than 6 weeks, with no additional curation needed. 3,809 were flagged for further review by the Exercise Oncology team, which were finalized within another 3 weeks. Overall, curators used an NLP suggestion for one or more assessment fields in 48,261 of the 51,592 surveys (93.5%).

Discussion
We provided an approach to reduce the time required to extract out structured data from handwritten patient surveys. While the original intent of the project was to create a fully automated computer vision and NLP pipeline to structure the handwritten text with minimal human curation, it quickly became apparent that due to the low quality of the data (primarily from incorrect or incomplete responses to the exercise survey questions), that would not be feasible even with current state-of-the-art deep learning/NLP methods. Thus, we pivoted to create a web application that would provide the automated suggestion for easy use, but also speed up the curation in general by showing the curators exactly what they needed to see right alongside where they were entering the data. There were challenges and latency associated with heavy use of the REDCap API versus directly curating in REDCap. Future work on a more generalized curation assistance tool will address this by embedding new functionality into REDCap.
Supporting Population Health Outcomes Studies Using a Framework of Social Determinants Linked EHR Data

Md Kamruz Zaman Rana, MS1,2,3, Humayera Islam1,2,3, Tanmoy Paul1,4, Khuder Alaboud1,3, Xing Song1,2,3, Russ Waitman1,2,3, Abu S. M. Mosa, Ph.D., FAMIA1,2,3,4,*
1NextGen Biomedical Informatics Center; 2Department of Health Management and Informatics; 3Institute for Data Science and Informatics; 4Department of Electrical Engineering and Computer Science; University of Missouri, Columbia, Missouri;*Corresponding Author

Introduction
Population health research needs to link patients’ health care data with social determinants for social determinants of health studies. Research that requires such data must collect identifiable information (patient’s address) to connect the EHR data with social determinants of health. A researcher must consider the security risk of storing and computing identified data to follow HIPAA regulations in this process. From an institutional perspective, a framework of social determinants linked to EHR data can reduce this step for the researcher. This study aims to demonstrate such a framework for facilitating Healthcare research data by using Privacy-preserved linkage of EHR data with geocode-linked external social determinants data sources and chronic condition phenotyping incorporated in the data lake.

Method
University of Missouri’s NextGen BMI’s Datalake provides health care data using the PCORnet common data model in identified and de-identified format. Using DeGAUSS geocoder1, geocoded address information was populated in the CDM. Cohorts for 27 chronic condition was generated and stored in Datalake by combining multiple algorithms provided by the Center for Medicare and Medicaid Services (CMS). Chronic conditions phenotype data for each patient is generated and stored in the Datalake using this algorithm. Social determinant data such as five years (2015-2019) data of U.S. Census Bureau’s American Community Survey (ACS) and the Area Deprivation Index (ADI) generated by the University of Wisconsin School of Medicine and Public Health’s Department of Medicine are staged and linked to each patient using the geocoded information. A study on the relationship between chronic condition outcomes and ADI is conducted to demonstrate that social determinants-based research can be done by sharing only de-identified data, which preserves the privacy of patient data (Figure 1).

Results
From 1,673,145 patients in the NextGen BMI’s Datalake, a total of 899,231 unique addresses were geocoded to the Census block group level with a mean geocoding accuracy of 76% (Figure 2). Sharing the ACS data connected with EHR data risks re-identification but mitigable. Using univariate logistic regression analysis, we observed that ADI could determine the risk of chronic conditions. With the increase of ADI, the risk is increased for the following chronic conditions (see Figure 3 for diabetes as an example): Acute Myocardial Infarction, Anemia, Asthma, Atrial Fibrillation, Chronic Kidney Disease, Chronic Obstructive Pulmonary Disease and Bronchiectasis, Depression, Diabetes, Heart Failure, Hypertension, Ischemic Heart Disease, Stroke/Transient Ischemic Attack, Colorectal Cancer, Lung Cancer, and the risk is decreased for the following chronic conditions: Cataract and Hyperlipidemia. The poster presentation will describe the process and demonstrate the utility of the privacy-preserved and enhanced data lake for population health outcomes studies using social determinants.

References
Improving the Natural Language Processing Pipeline for De-identification of Clinical Notes at an Academic Medical Center

Vignesh Rao, Arnav Bhandari, Michael Horvath, Joseph Rigdon Ph.D.,
Umit Topaloglu Ph.D., FAMIA, Wake Forest School of Medicine, Winston Salem, NC

Introduction
The Health Insurance Portability and Accountability Act (HIPAA) regulates researcher’s access to clinical notes due to the presence of Protected Health Information (PHI). The existing de-identification pipeline developed at Wake Forest Baptist Medical Center uses MITRE Identification Scrubber Toolkit (MIST) models to redact PHI from clinical notes [1]. This previous pipeline achieved more than 95% recall rates for patient names and greater than 98% recall rates for overall PHI found in pathology and radiology notes. It was approved by the Institutional Review Board (IRB) and the Chief Privacy Officer of this institute based on aforementioned metrics, and allowed researcher’s access to these two types of de-identified notes before obtaining an IRB approval. However, when this same method was applied to Progress notes the overall recall rate was ~95% while the recall rate for the NAME tag was ~92%. Our aim was to increase this 92% recall rate for the NAME tags, motivating the creation of an ensemble method of de-identification. The recall rate directly corresponds to the PHI in each note and hence the presence of

Methodology
Our gold standard was a sample of 629 randomly selected unstructured progress notes that were manually labeled. The proposed pipeline involved the initial application of the MIST model which tagged and redacted the notes with labels NAME, DATE, IDNUM, LOCATION (related to patient), PHONE, HOSPITAL, and AGE. The second step involved the python integration of a deep learning technique to the post-processed MIST redacted notes, specifically for the NAME tag in the clinical notes. Future steps involve testing deep learning techniques for other PHI identifiers. The deep learning models that were compared were NLTK, SPACY, Hugging Face BERT-Base-NER and Hugging Face BERT-Large-Cased. Performance was tested by calculating the precision, recall and F-measure.

Results and Discussion

<table>
<thead>
<tr>
<th>Ensemble Methods</th>
<th>Tags</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Change in Precision</th>
<th>Change in Recall</th>
<th>Change in F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only MIST</td>
<td>Name</td>
<td>95.57%</td>
<td>92.33%</td>
<td>93.91%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>96.55%</td>
<td>96.06%</td>
<td>96.56%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>MIST + SPACY</td>
<td>Name</td>
<td>59.46%</td>
<td>95.30%</td>
<td>73.18%</td>
<td>-36.113%</td>
<td>2.973%</td>
<td>-20.727%</td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>86.90%</td>
<td>96.52%</td>
<td>91.46%</td>
<td>-9.647%</td>
<td>-0.080%</td>
<td>-5.100%</td>
</tr>
<tr>
<td>MIST + BERT Large Case</td>
<td>Name</td>
<td>94.36%</td>
<td>94.26%</td>
<td>94.28%</td>
<td>-1.213%</td>
<td>1.933%</td>
<td>0.373%</td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>96.54%</td>
<td>97.42%</td>
<td>96.98%</td>
<td>-0.007%</td>
<td>0.820%</td>
<td>0.420%</td>
</tr>
<tr>
<td>MIST + BERT base NER</td>
<td>Name</td>
<td>93.08%</td>
<td>94.32%</td>
<td>93.68%</td>
<td>-2.493%</td>
<td>1.993%</td>
<td>-0.227%</td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>96.28%</td>
<td>97.10%</td>
<td>96.70%</td>
<td>-0.267%</td>
<td>0.500%</td>
<td>0.140%</td>
</tr>
</tbody>
</table>

Table 1: Performance Metrics of the different ensemble methods relative to the basic MIST model

On average, the overall highest increase in recall rates for NAMES is 2.97% for the ensemble method with SPACY, but there is a 36.1% drop in precision for the identification of names due to over tagging. The MIST + BERT Large Case is the only method which has a rate of improvement in recall greater than the rate of decline in precision; hence leading to a corresponding increase in the F-measure as well. The IRB approved the more statistically significant ~95% recall rates we were able to achieve for the NAME tag, which was an improvement from ~92%.

References
Air Pollution and Type 2 Diabetes: A Rural and Urban Comparison
Naomi O. Riches, PhD, MSPH¹, Julio C. Facelli, PhD, FACMI¹, Adriana Payan-Medina¹, Ramkiran Gouripeddi, MBBS, MS¹
¹University of Utah, Salt Lake City, UT, USA

Introduction
Type 2 diabetes mellitus (T2DM) is a chronic multifactorial disease that afflicts millions across the US. T2DM has many comorbidities such as cardiovascular disease, neuropathy, and premature morbidity. Increasingly recognized as risk factors are environmental exposures, such as exposure to air pollution (AP). While many epidemiological studies indicate a significant positive correlation between T2DM and AP, the results are not consistent. It could be that certain mixtures of AP or sources impact health differently. Previously, our group and others have used unsupervised machine learning methods, k-means clustering, to successfully categorize AP data with health outcomes. This was done using US Environmental Protection Agency (EPA) data, which are collected in most metropolitan counties in the US. However, EPA monitors leave many of the more rural areas under or unmonitored. The aim of this research is to use machine learning methods to associate health effects with AP data estimates from NASA satellites, which provides coverage for the entire US population using data from NASA satellite instruments.

Methods
NASA satellite instruments collect daily data on PM₂.₅, SO₂, NO₂, and CO relative concentrations, which are available as geographically gridded data (0.625°x0.500° or 0.25°x0.25° spatial resolution; https://disc.gsfc.nasa.gov/datasets/). For this study, we downloaded data for years 2007-2016 and aggregated to US county-level. An unsupervised machine learning method, k-means clustering, was used to partition the multidimensional daily AP data. Additionally, the same data will be partitioned using the method, k-shape. This method considers longitudinal data to divide repeatedly measured daily air quality into similar subgroups based on trajectory shape and clusters using the Fréchet distance vs Euclidian distance. Due to the size (n > 9 million daily AP measurements) of the NASA data, k-means will be run first to choose representative AP trajectories, then data points will be reduced within those trajectories if necessary. The resulting clusters from these two methods will be matched to the US Centers for Disease Control and Prevention’s (CDC; https://www.cdc.gov/diabetes/data/index.html) T2DM county-level incidence per 1000 among adults 20 years+, for the year following air pollution exposure (i.e., 2008-2017) and the 2013 Rural-urban continuum codes (RUCC) to determine level of rurality for each cluster.

Results
The k-means analysis expanded prior work with EPA data to include rural US counties. The results support the finding from the EPA project in that increasing annual T2DM incidence was associated with higher relative mean levels of PM₂.₅ in both rural and urban counties (see Table). However, the inclusion of all counties, using the NASA data, resulted in an inverse pattern of NO₂ concentration with the change in annual T2DM from that seen with the EPA data. A lower mean RUCC score (i.e., more rural) was associated with the smallest annual increase in T2DM incidence, highest NO₂ concentration, and lowest PM₂.₅ concentration.

Conclusion
Chronic exposure to air pollution is known to cause systemic inflammation, adiposity, and insulin resistance thereby increasing the risk of T2DM. However, there is limited air pollution monitoring in rural areas by the EPA in 79% of counties across the US, accounting for approximately 30% of the US population. The use of satellite instruments to detect and monitor air pollution could improve exposure assessments for rural areas.

References
   doi:10.1371/journal.pone.0150738
Assessing the Contribution of Scanned Outside Documents to the Completeness of Real-World Data Abstraction

Dana Rollison, PhD1; Yayi Zhao1, MA, MS1; Rachel Howard, PhD1; Sandra Stewart, CTR1; Rossybelle Amorrortu, MPH1
1Moffitt Cancer Center, Tampa, FL, USA

Introduction

Electronic medical records (EMR) are widely used for precision medicine, observational research, and population health management. However, up to 80% of EMR data can be stored in scanned documents external to the main treatment facility1, posing an informatics challenge, particularly in tertiary referral settings such as academic cancer centers. Studies are needed to characterize the information uniquely contained in scanned outside documents (SOD) to understand how access to data in SOD (or lack thereof) may affect quality reporting and the use of these “real-world data” for cancer research.

Methods

Data were independently abstracted2 twice from the EMR corresponding to the same set of 127 advanced non-small cell lung cancer patients, with and without including SOD associated with those patients. In total, 32 variables were abstracted, including demographics, diagnosis, biomarker tests, disease progression, Eastern Cooperative Oncology Group (ECOG) score and oral drug use. Completeness was calculated between the two abstractions.

Results

The overall completeness of the abstraction including SOD was 79.7% as compared to 59.8% for the abstraction without SOD. The difference in completeness was largely driven by biomarker test results which were more likely to be available in SOD (Figure 1). The completeness of biomarker variables was significantly lower in the abstraction without SOD (median 31.13%) as compared to that with SOD (median = 74.19%, p-value = 0.0003). However, such difference was not observed for non-biomarker variables between abstraction without SOD (median = 97.64%) and that with SOD (median =100%, p-value = 0.0897).

Figure 1. Comparison of data completeness between abstractions including and excluding scanned outside documents.

Conclusion

There were no major differences in completeness between the abstraction with and without SOD for demographics, diagnosis, disease progression, ECOG score or oral drug use. However, missing biomarker data from the abstraction without SOD drove down the overall completeness. Study findings may help cancer centers prioritize the type of SOD to be abstracted and/or targeted for natural language processing.

References

NLP Sandbox: Overcoming data access barriers to reliably assess the performance of NLP tools

Thomas Schaffter, PhD¹, Thomas Yu¹, George Kowalski², Sijia Liu, PhD³, Connor Boyle⁴, Yao Yan¹,4, Jiaxin Zheng¹, James Eddy, PhD¹, Hongfang Liu, PhD³, Bradley Taylor, MBA², Justin Guinney, PhD¹,5

¹Sage Bionetworks, Seattle, WA; ²Medical College of Wisconsin, Wauwatosa, WI; ³Mayo Clinic, Rochester, MN; ⁴University of Washington, Seattle, WA; ⁵Cancer Genomics at Tempus Labs, Inc. Chicago, IL

Abstract

The NLP Sandbox adopts a model-to-data approach to enable NLP developers to assess the performance of their tools on public and private datasets. When a developer submits a tool, partner organizations (e.g., hospitals, universities) automatically provision the tool and evaluate its performance against their private data in a secure environment. Upon successful completion, the performance of the tool is returned by the partner organization and automatically published in the leaderboards.

Introduction

Critical patient information derived from academic research, health care, and clinical trials are off limits for traditional data-to-model (whereby data is transferred/downloaded into a new environment to be colocated with the executable model) benchmarking of NLP tools. Existing barriers include restricted access to prohibitively large or sensitive data. In addition to data access constraints, we also lack effective frameworks for assessing the performance and generalizability of NLP tools.

Methods

The NLP Sandbox adopts a model-to-data approach to enable NLP developers to assess the performance of their tools on public and private datasets (Figure 1). When a developer submits a tool, partner organizations (e.g., hospitals, universities) automatically provision a tool, execute it, and evaluate its performance against their private data in a secure environment. Upon successful completion, the partner organization reports the performance of the tool, which is then automatically published in the NLP Sandbox leaderboards.

Results

The first series of NLP tasks that the NLP Sandbox supports is the annotation of Protected Health Information (PHI) in clinical notes. These tasks have been identified through our collaboration with the National Center for Data to Health (CD2H). Submitted tools are currently evaluated on the dataset of the 2014 i2b2 NLP De-identification Challenge and private data from MCW. Additional data sites are currently being onboarded (Mayo Clinic, UW). As of March 1st, 2022, 56 PHI annotators have been successfully evaluated.

References

Neural Network for Stroke Outcome Prediction after Thrombectomy

Parvez M. Shaikh MD\textsuperscript{1}; Abby Swanson-Kazley PhD\textsuperscript{2}; Abdul R. Tarabishy MD\textsuperscript{1}; Abdulrahman Alhalak BS\textsuperscript{1}; Alfredo R. Siochi PhD; Geraldine M. Jacobson MD\textsuperscript{1}; Ansaar Rai, MD\textsuperscript{1}
\textsuperscript{1}West Virginia University, Morgantown, WV; \textsuperscript{2}Medical University of South Carolina, Charleston, SC

Introduction
Several factors that have been reported to be correlated with outcome of acute occlusive stroke including age, certain comorbidities, time to procedure, Alberta Stroke Program Early CT Score (ASPECTS) and, National Institutes of Health Stroke Scale (NIHSS) after mechanical thrombectomy (MT) in occlusive strokes (1). Here we use an artificial neural network (ANN) to analyze data from a tertiary care stroke center.

Methods
De-identified data from 215 patients undergoing thrombectomy was obtained from a large tertiary stroke center. An ANN was designed to predict the 90 day modified rankin score (MRS) for stroke patients after undergoing thrombectomy using SPSS. Sex, race, diabetes, hypertension, afib, hyperlipidemia, CHF, prior stroke, smoking, pre-stroke MRS, Anterior Circulation/Posterior circulation, Right/Left, ICA angioplasty, IV-tPA, IA-tPA, complications, distal embolization, hemorrhage present, craniectomy done, balloon guide used, balloon used were binary features. Age, admission NIHSS, ASPECTS, onset to puncture time, attempts, procedure time, length of stay (los) were continuous features. The output variable, 90 day MRS, was coded as: MRS\leq 3 (mrs\_gt\_3=0) and MRS=4-6 (mrs\_gt\_3=1). We used 70:30 partitioning of data for train:test data set, analyzed ANN’s performance and performed a feature importance analysis to see which variable were most influential outcome prediction.

Results
We used two hidden layer and the SPSS algorithms chose two nodes in each hidden layer (2). The Activation function for the hidden layers were set to hyperbolic tangent and softmax for output nodes. Cross-entropy was used for model validation. The training and test data accuracy was 80% and 79.4% respectively. The area under curve (AUC) of the receiver operator characteristics (ROC) curve (figure 1a) was 0.852 for both prediction (i.e., mrs\_gt\_3 = 0 or mrs\_gt\_3 = 1). Feature importance analysis (figure 1b) revealed that the most important predictor for 90 day MRS > 3 was age followed closely by NIHSS admission score. The next seven were length of stay, onset to puncture time, procedure time, number of attempts, hyperlipidemia, ASPECTS, presence of hemorrhage.

![Figure 1](image)

\textbf{Figure 1}: Receiver operator characteristics curve (a), Important features for 90 day MRS prediction (b)

Conclusion
ANN performed very well in predicting 90 day MRS > 3, which is a poor outcome. The ability to predict poor outcome in a patient is valuable as knowledge of these variables can help mitigate some of the risk particularly of the modifiable factors. For example, the onset to puncture time in particular is important – it is among the top five most important predictor of poor outcome. Moreover, hyperlipidemia, smoking and diabetes appears to intermediate effect on outcome and can be optimized for patient that are at high risk of stroke at the primary care level. Factors such as race, hypertension, afib, IV-tPA, IA-tPA and balloon guide use had minimal impact.

Reference
Testing a Novel Tool for the Development of Drug-drug Interaction Clinical Decision Support

Max A. Sibilla¹, Daniel Malone², Guilherme Del Fiol², Richard D. Boyce¹
¹University of Pittsburgh, Pittsburgh, PA; ²University of Utah, Salt Lake City, UT

Abstract
Enhancements were made to the CDS Connect Authoring tool with the goal of making it easy for drug interaction experts to develop shareable DDI CDS artifacts that are widely usable by CDS systems using CDS Hooks, CQL and FHIR.

Introduction
Drug-drug interactions (DDIs) are preventable adverse events that are responsible for 5–14% of adverse drug reactions (ADRs) in hospitalized patients. Exposure to life-threatening DDIs continues to occur despite the widespread use of clinical decision support systems (CDS). Clinicians override up to 90% of potential DDI alerts, primarily because clinicians do not consider the alerts to be relevant. Recent advances in health information technology standards such as HL7 Clinical Quality Language (CQL) and HL7 Fast Healthcare Interoperability Resources (FHIR) make it possible for creators of DDI CDS to use electronic health records data to create patient-specific alerts that provide key contextualized information. A recently published study showed the potential for eight contextualized drug interaction algorithms to reduce alerts that interrupt clinician workflow by > 50%¹.

A barrier to the widespread creation and use of contextualized DDI alerts is that DDI experts have no tools to help them author, test, and share new contextual CDS artifacts. In this study, we enhanced the Agency for Healthcare Research and Quality CDS Connect Authoring tool with the goal of making it easy for drug interaction experts to develop DDI CDS artifacts that are widely usable by CDS systems through the use of CDS Hooks, CQL and FHIR.

Methods
We elicited the requirements for a highly usable DDI CDS rule development tool using two clinically significant and frequently occurring pharmacodynamic drug interactions involving high-risk anticoagulations as use cases. Several new features were added to the CDS Connect Authoring tool to address the requirements. Think-aloud session were used to evaluate the usability of the interface. Several drug experts were recruited to complete pre-specified DDI CDS rule writing tasks using the enhanced CDS authoring tool and the sessions were completed via web conference.

Results
After modifications were made to the tool including an overhaul of the user interface, usability sessions were conducted with 7 drug experts. At the end of the sessions participants were asked to complete a System Usability Scale (SUS) questionnaire regarding the application with 1 being “strongly disagree” and 5 being “strongly agree”. The results of the questionnaire (Table 1) showed that participants felt relatively confident using the tool and would be able to learn to use it quickly. While participants did not consider the system unnecessarily complex and they would like to use it frequently, the average SUS score of 3.23 indicates a learning curve.

Conclusion
Implementing a diagram-based user interface improved the usability of the CDS Authoring Tool in the realm of CDS rule authoring. It allowed for this novel tool to bridge the gap between domain expertise and rule creation. Additional modifications allowed these rules to be immediately tested against an already existing electronic health records FHIR server. With this tool contextual CDS artifacts can now be authored, tested, and shared quicker and easier.

Table 1. Averages from the results of the System Usability Scale questionnaire

<table>
<thead>
<tr>
<th>Question</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>I would like to use this system frequently</td>
<td>3.83</td>
</tr>
<tr>
<td>I found the system unnecessarily complex</td>
<td>2.67</td>
</tr>
<tr>
<td>I thought the system was easy to use</td>
<td>3.67</td>
</tr>
<tr>
<td>I would need the support of a technical person to use this system</td>
<td>3.17</td>
</tr>
<tr>
<td>I think that most people would learn to use this system very quickly</td>
<td>3.67</td>
</tr>
<tr>
<td>I found the system very cumbersome to use</td>
<td>2.17</td>
</tr>
<tr>
<td>I felt very confident using the system</td>
<td>3.33</td>
</tr>
<tr>
<td>I needed to learn a lot of things before I could get going with this system</td>
<td>3.33</td>
</tr>
</tbody>
</table>

References
Automated Clinical Quality Language (CQL) Generation

Andrew P. Stevens, MS¹, Michael K. Riley, MS¹
¹Georgia Tech Research Institute (GTRI), Atlanta, GA, United States of America

Abstract

Clinical Quality Language (CQL) is a domain-specific query language that represents clinically-focused knowledge, focusing on applications such as clinical decision support and public health reporting. The developed system allows for a conversion of a clinician’s knowledge and needs into CQL with minimal technical knowledge, and can be evaluated against a FHIR-based system to deliver results consistently. This system will be further developed to expand its use cases as well as increase ease-of-use through a user-interface.

Purpose of the System

FHIR recently adopted CQL as one of its standard logical expression languages and the Centers for Medicare & Medicaid Services (CMS) announced that its electronic clinical quality measures will be transitioned to use CQL, indicating strong support for CQL in major health interoperability standards. When developing clinical context for CQL, we found the clinician wanted to use consistent strategies for retrieval and evaluation. To support these strategies, we developed a compressed data structure in Python and used it to generate consistent CQL across multiple contexts. This compressed data structure was implemented with a set of templates and generators for different FHIR versions to create version-appropriate CQL statements. This system allows a clinician to repeatedly change the input to generate consistent CQL, resulting in a shorter turn around between wanting to retrieve and evaluate.

The input JSON requires minimal technical knowledge beyond the editing of JSON files. The input takes in an index event (what condition or “event” in a patient’s timeline would you like to include), any inclusions (what would you like to include in addition to the index event or relative in time to the index event), and any derived data (how would you like to return the data from the inclusions in a structured format). For example, your index event would be a patient’s diagnosis of syphilis, you would want to include any medications or lab tests that relate to this syphilis diagnosis, and you would want to return certain fields from those resources into a structured tuple. In future development, this could be transformed into a user interface to remove the need to understand JSON data structures.

In Figure 1, we show an example workflow using our CQL generation tool. A clinician incorporates their knowledge and defines concept sets with the OHDSI Atlas tool¹ (or manually if preferred) into a JSON format that is the input to the generator. The generator produces an output CQL script, which is then evaluated against a CQL engine (in this example, the engine calls upon an OMOPonFHIR² instance, but it can be any FHIR server). This CQL engine returns the desired output, which the clinician now has without needing any knowledge of how to write CQL.

![Automated CQL Generation Diagram](Image)

**Figure 1.** Example workflow involving the CQL generation tool, going from clinical knowledge to clinical results

References

Analyzing Users Workload in Clinical Projects Conducted via REDCap

Lina Sulieman, PhD, Beverly Woodward, MSN, RN, Paul Harris, PhD
Vanderbilt University Medical Center, Nashville, TN, United States

Introduction

Research teams use REDCap to design data management plans and collect data for clinical trials and studies. Training users to perform their duties is an essential part of clinical studies timeline. Moreover, reporting the operational hours spent on a project can be essential for budget management, especially for studies funded by grants. Currently, there is no public tool that quantify actions and hours that team members dedicate for project setup and operational use. Our team recently added a REDCap API method allowing programmatic extraction of audit trail log details. For this project, we developed methods for automated reporting of research study personnel activities using information extracted from this API endpoint.

Methods

We developed a Python script to extract REDCap logs for a multi-institutional research study. Each log entry includes timestamp, action, user who performed the action, and instrument. We calculated the number of hours that the user worked and accounted for multiple logins per day by defining sessions. If the difference between two consecutive actions was two hours or more, we assumed that the user started a new session. We calculated the hours spent per session. We divided the project timelines into weeks and extracted the number of hours per user and the number of actions grouped by type.

Results

The project lasted for 22 months and had 62,484 records. During project period, 30 users performed 69,592 actions that were categorized into 10 different actions. The users had on average 26.2±19.6 sessions and worked for 0.84±1 hour (50.2 minutes). The number of users who worked on the project as well as the hours peaked in the sixth week where 11 users worked for 15 hours (depicted in Figure 1). The fourth week had the highest number of actions which was 938. During the first 6 weeks, more than half of the actions were managing and designing while at least 30% of actions in the following weeks were updating record.

Discussion and Conclusion

Analyzing users’ actions in clinical studies can help principal investigators and research coordinators plan team responsibilities, track study progression, identify bottlenecks, and plan for training and operational budgeting. The methods we have developed are scalable and transferrable, requiring only a REDCap API project token as input for deployment on other projects.
How Much Time Do Nurses Really Spend at the Bedside?
Carolyn J. Sun, PhD, RN, ANP-BC1,2, Caroline J. Fu, MPH1, Alan Levin, DNP, MBA, RN, CPHQ, NEA-BC3, Kenrick Cato, PhD, RN2,3
1Hunter College, New York, NY; 2NewYork-Presbyterian Hospital, New York, NY; 3Columbia University, New York, NY; 4NewYork-Presbyterian Queens, New York, NY

Introduction
As the largest component of the healthcare workforce, nurses spend more time with patients than any other profession. However, the exact amount of time a nurse spends at the bedside remains largely unknown. Existing studies rely mainly on self-report or methods such as “time-in-motion,” Work Observation Method by Activity Timing (WOMBAT), or “work sampling” which are approximations but do not obtain a detailed record over an extended period. Nurses giving report at the patient bedside (bedside shift report (BSR) and hourly rounds (HR) are recommended to improve patient outcomes. Study Aim: Using data generated by a novel technology (“AUGi” inspiren.com) with the ability to capture bedside interactions, characterize nurse surveillance including frequency and duration of BSR and HR and other interactions at the bedside, 24 hours/day, 7 days/week for one year.

Methods
The AUGi device was validated by the authors over a six month pilot period. Devices were installed at 37 patient beds in two medical/surgical units; 99 users were tracked. Data were collected 4/15/19 – 3/15/20. Nurses’ time and activity at the bedside were characterized using AUGi which integrates obfuscated computer vision with a Bluetooth beacon on the nurses’ identification. Data were aggregated and key variables like BSR and HR were extracted. Descriptive statistics for duration of BSR, HR and other nurse-patient interactions were conducted. A T-test comparing mean interaction times between day and night shift was calculated. SAS software was used to analyze data. The study was approved by Institutional Review Boards.

Results
A total of N=408588 interactions from 37 beds and 49 nurse users over 670 shifts were analyzed. We observed over 1.5 times more interactions during day shifts (DS) (n=247273) compared to night shifts (NS) (n=161135), but the mean interaction time was 3.34 seconds longer during nights than days (p<0.0001). Each nurse had an average of 7.86 (standard deviation [sd]=10.13) interactions per bed each shift and a mean total interaction time per bed of 9.39 minutes (sd=14.16). On average nurses covered 7.43 beds (sd 4.03) per shift (DS: mean 7.80 beds/nurse/shift, sd 3.87; NS: mean 7.07 /nurse/shift, sd 4.17). The 37 beds were occupied on average for 90.4% (range from 76% to 97%) of shifts during the study period. The mean time per HR was 69.5 seconds (sd=98.07) and 50.1 seconds (sd=56.58) for BSR (combined interactions are presented in Table 2). A Mann-Whitney U-Test comparing mean interaction time between DS and NS revealed that DS mean interaction time was 3.34 seconds shorter than NS, which is statistically significant (sd 0.32, p<0.001).

Conclusion
To our knowledge, this is the first study to provide continuous surveillance of nurse activities at the bedside over a year long period, 24 hours/day, 7 days/week. We detected nurses spend less than one-minute giving report at the bedside, and this is only completed about 20.7% of the time. Additionally, hourly rounding was completed only 52.9% of the time and nurses spent only 9 minutes total with each patient. Further study is needed to detect whether there is an optimal timing or duration of interactions to improve patient outcomes. In the future, such technology may allow nurses to cluster visits for optimal timing and duration of patient interactions.

References
Relation Extraction from Biomedical Literature on Pharmacokinetic Natural Product-Drug Interactions

Sanya B. Taneja¹, Patrick W. Ndungu¹, Mary F. Paine², Sandra L. Kane-Gill¹, Richard D. Boyce¹

¹University of Pittsburgh, Pittsburgh, PA, ²Washington State University, Spokane, Washington

Introduction

Co-consumption of botanical and other natural products (NPs) with pharmaceutical drugs can lead to pharmacokinetic NP-drug interactions (NPDIs), which can further lead to unwanted drug response. Computational approaches such as semantic relation extraction from in vitro and clinical pharmacokinetic studies related to the topic can provide insights about the mechanisms of these NPDIs. We extracted predications (subject-predicate-object triples) from full text articles focused on pharmacokinetic drug interactions involving the model NP green tea using two relation extraction systems, related the information to existing biomedical resources, and compared the results to human data extraction.

Methods: Full texts of 13 PubMed-indexed articles related to pharmacokinetic green tea-drug interactions were obtained from PubMed Central and PDF files. SemRep (v1.8) and the Integrated Network and Dynamical Reasoning Assembler (INDRA) with the REACH biological reader were used to extract predications from the full texts. All subjects and objects in the predications were mapped to the Unified Medical Language System (UMLS) concepts using MetaMap. The predications were compared with data from the Center of Excellence for Natural Product Drug Interaction Research (NaPDI Center) public repository. Data extracted in the repository by humans from the same articles was transalted to predications using relations from the Open Biological and Biomedical Ontology (OBO) Foundry ontologies and established as the ground truth.

Results

Counts of pharmacokinetic-related predications from the NaPDI Center repository, SemRep, and INDRA/REACH are shown. Translation of the repository data resulted in 179 predications. 522 predications were extracted from SemRep. 123 predications were extracted from INDRA/REACH, of which 94 were relevant to pharmacokinetic NPDIs (76.4%); 65 were green tea specific (52.8%). 193 predications from SemRep were related to pharmacokinetic interactions, 128 were relevant to pharmacokinetic NPDIs (66.3%), and 77 (39.9%) were green tea specific. The recall values for SemRep and INDRA/REACH were 0.31 and 0.20, respectively.

Conclusion: We present a baseline evaluation of relation extraction from full texts of a subset of articles focused on green tea-related pharmacokinetic interactions. The results are analogous to prior studies, however, certain predicate types such as ‘substrate of’ were not captured by either of the systems as they are not included in the relations extracted by the systems. It may be possible to map these predications to alternative relations. We are currently assessing the correctness of inferences and evaluation beyond the recall for the broader set of articles from our search strategy. Although there are several challenges in extracting relevant information, semantic relation extraction is a scalable approach to find associations between biomedical entities beyond simple named entity recognition to inform scientists of evidence of prior work.

References

A Multi-Task Learning Approach to Chemical Property Prediction

Carson Tao, Ph.D.1, Yoann Mamy Randriamihaja, Ph.D.1
1Sumitovant Biopharma, New York, NY, USA

Introduction

Deep learning has been applied to various chemical property prediction tasks with improved results. While lab-level accuracy is still out of reach, machine predictions are getting closer to the measured values. When compared to lab measurement, such a prediction offers general understanding of key chemical properties for a given molecule with significantly-reduced cost and manual labor. In this poster, we evaluate the efficacy of multi-task learning (MTL) in training chemical property prediction models as compared to single-task learning (STL). When it comes to evaluation metrics, we propose a new approach that offers better real-world practicality for discussion with domain expert.

Dataset and Method

Our approach combined message passing neural network1 with a self-attention layer2, both of which have claimed competitive results in chemical property prediction. The input to our deep learning pipeline is canonicalized molecular structures in SMILES format. The output is a numerical prediction to each of the property type in the training dataset. We evaluate our methods using publicly available datasets (top 4 in Table 1). To further estimate the efficacy of MTL, we prepared a new dataset by querying ChemBL, a database of bioactive molecules with drug-like properties, on reported chemical ADME stability values from scientific publications. We expect the addition of ADME training data (834 samples) to help the prediction of other related chemical properties when simultaneously trained.

We first cross-validated our deep learning pipeline using lipophilicity dataset. Our pipeline achieved an average RMSE of 0.61, which is comparable to other state-of-the-art systems1. However, we point out that RMSE is not able to showcase the full picture of prediction variability especially when targeted property values are highly fluctuated. We recommend using percentage of predictions within twofold of measured values as an alternative or additional metrics to better capture the outliers of prediction results. The same experiment on lipophilicity dataset scored 44% within twofold of measurement, which is easier to assess a model performance than RMSE. We train deep learning models using the following settings: (1) train each dataset separately using STL, (2) train 4 public datasets simultaneously using MTL, and (3) train all datasets including the newly created ADME dataset using MTL. All results are 10-fold cross-validated with scaffold stratification of training and test datasets to avoid testing on chemical scaffolds used in the training set and improve generalization.

Results and Conclusion

Our deep learning pipeline achieved competitive performance when evaluated on the publicly-available dataset when predicting lipophilicity values. Despite a 1% reduction in human clearance property, we observed improvement when the model was trained in MTL setting. Furthermore, with the inclusion of ADME stability dataset in MTL training, our pipeline achieved significant improvement on the prediction of related properties like solubility (+10%) and human clearance (+56%). Such an improvement is achieved under a relatively small dataset (834) from manual curation. Lastly, we recommend using percentage within twofold of measurement as an alternative evaluation metrics.

Table 1. Cross-validated prediction performance (percentage within twofold of measurement, higher is better)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sample Size</th>
<th>Properties</th>
<th>STL</th>
<th>MTL</th>
<th>MTL+ADME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lipophilicity</td>
<td>4,200</td>
<td>logP</td>
<td>0.44</td>
<td>0.50</td>
<td>0.48</td>
</tr>
<tr>
<td>Solubility</td>
<td>1,311</td>
<td>logS</td>
<td>0.33</td>
<td>0.36</td>
<td>0.46</td>
</tr>
<tr>
<td>PPB (plasma protein binding)</td>
<td>1,614</td>
<td>Human PPB</td>
<td>0.96</td>
<td>0.96</td>
<td>1.00</td>
</tr>
<tr>
<td>Clearance</td>
<td>1,102</td>
<td>Human clearance</td>
<td>0.32</td>
<td>0.31</td>
<td>0.87</td>
</tr>
</tbody>
</table>

References

Advancing Hallmarks of Cancer Identification from Scientific Literature

Carson Tao, Ph.D.\textsuperscript{1}, Naifeng Liu\textsuperscript{1,2}, Murali Dandu\textsuperscript{1,2}, Yoann Mamy Randriamihaja, Ph.D.\textsuperscript{1}  
\textsuperscript{1}Sumitovant Biopharma, New York, NY, USA; \textsuperscript{2}CUNY Graduate Center, New York, NY, USA; \textsuperscript{3}University of California San Diego, San Diego, CA, USA

Introduction

In oncological studies, HoC (hallmarks of cancer) refers to the characteristics from malignant cell’s behavior. Based on the taxonomy developed for the past decade, there are 10 major HoC categories that reveal the dynamic changes in the genome. The proper identification of HoC categories from the scientific literature enables downstream studies by providing categorical association to each sentence inside a PubMed abstract. While HoC classification has been addressed on the abstract-level in a recent study, the current state-of-the-art method in sentence-level classification is proposed by Baker et al.\textsuperscript{1} using SVM with extensive external features. We tackle HoC classification at sentence-level using two methods, SVM and deep learning-based, aimed at an increment to the current state-of-the-art performance.

Data and Method

We used the official HoC annotation dataset provided by Baker et al. The official dataset contains 1,852 abstracts and 17,432 sentences, out of which only 25% are associated with HoC categories (multi-label). Such an imbalanced representation makes the classification more challenging on the sentence-level when compared to its abstract-level counterpart. We prepared training and test dataset using stratified sampling and evaluated our methods in 4-fold cross-validation, which is the same as Baker et al.’s experiment. We propose two methods on sentence-level HoC classification. (1) We use TFIDF vectorizer for the representation of unigram/bigram features and dynamic class weighting to mitigate negative impact of imbalanced training samples. We train an SVM model with L1 regularization. (2) We utilize PubMedBERT language model for the generation of contextual representation and fine-tuned it with a linear layer. We train 10 multi-binary models on both methods for the output of multi-label prediction.

Results and Conclusion

Table 1 shows cross-validated system performance when compared to the previous state-of-the-art. Due to page limitation, we show the 5 most frequent categories. The micro-average F1 is computed from results of all 10 HoC categories. With class weighting in the training process, our SVM-based method outperformed Baker et al. by 4.8% using only lexical features (i.e., without external features such as verb phrases and named-entities in one-hot encoding). Furthermore, our deep learning-based approach achieved an overall 16.1% performance increment without manual feature engineering. Specifically, fine-tuned language model improved the Cellular energetics class by 18.3%, which has the least amount of training samples (213). Overall, our system achieved 0.707 micro-average F1, which to the best of our knowledge represents an increment to the current state-of-the-art performance on the sentence-level HoC classification. In our future work, we plan to utilize multi-label output layer to reduce computation complexity in training classifiers and expand our study to other HoC sub-classes branched from the major 10 HoC categories.

Table 1. Cross-validated per-category F1 score on the official HoC dataset

<table>
<thead>
<tr>
<th>5 Major HoC Categories</th>
<th>Baker et al.</th>
<th>SVM</th>
<th>DL</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sustaining proliferative signaling</td>
<td>0.473</td>
<td>0.485</td>
<td>0.609</td>
<td>+13.6%</td>
</tr>
<tr>
<td>Activating invasion and metastasis</td>
<td>0.634</td>
<td>0.681</td>
<td>0.792</td>
<td>+15.8%</td>
</tr>
<tr>
<td>Tumor promoting inflammation</td>
<td>0.501</td>
<td>0.549</td>
<td>0.697</td>
<td>+19.6%</td>
</tr>
<tr>
<td>Genomic instability and mutation</td>
<td>0.484</td>
<td>0.540</td>
<td>0.645</td>
<td>+16.1%</td>
</tr>
<tr>
<td>Resisting cell death</td>
<td>0.669</td>
<td>0.725</td>
<td>0.816</td>
<td>+14.7%</td>
</tr>
<tr>
<td><strong>Micro-Average (of 10 HoC Categories)</strong></td>
<td><strong>0.546</strong></td>
<td><strong>0.594</strong></td>
<td><strong>0.707</strong></td>
<td><strong>+16.1%</strong></td>
</tr>
</tbody>
</table>

References

Evaluating Primary Care Impacts of Auto-Primary Care Provider Assignment Using a Matched Randomized Controlled Trial

Jennifer Temple, MSE, John Schrom, MPH, Lindsay Sher, MD, Laurel Bennett, Esha Datta, PhD, Colleen Bouey, Andrew Diamond, MD, PhD
One Medical, San Francisco, CA

Introduction: Primary care leads to more high-value care and better health care experience for patients [1]. Retrospectively, patients with a regular medical care doctor have been found to have lower rates of emergency room use and hospital admissions than patients with no regular doctor [2]. Here we used a randomized controlled trial to evaluate how assigning a primary care provider (PCP) automatically impacts patient and provider behavior within the primary care setting.

Methods: Patients who had not elected to join a PCP panel but had visited a PCP within the last 18 months were selected as part of the matched randomized controlled trial. Patients were matched on relevant characteristics such as age and previous utilization and then assigned to test or control groups. A PCP was attributed to each member of the test group, based on an internal attribution algorithm using that patient’s historical visit patterns. Patients assigned to PCP panels received greater appointment access to their PCPs, but neither patients nor PCPs were explicitly notified of the assignment. For members of the control group, no additional provider access was given. Data on patient behavior was collected for the two groups for six months following assignment.

Results: Using a test group of 17,000 patients and a control group of 17,000 patients, fifteen hypotheses were tested, positing differences in patient engagement, provider engagement, continuity of care, patient-reported experience and operational burden for the groups. The Type 1 Error for the set of hypotheses was 0.05, with a Bonferroni correction of 15. The two null hypotheses addressing operational burden were rejected. A lower percentage of patient-initiated messages were routed to care providers on a supplementary care team for the test group (59% for the test group, 64% for the control group, p-value=8.1x10^-8). The average tasks per patient addressed by the supplementary care team was also lower for the test group (9% lower, p-value=8.7x10^-5). We failed to reject any of the six null hypotheses for patient engagement (Ex: H₀=The proportion of patients scheduling annual physicals was equal). We failed to reject any of the three null hypotheses around patient experience (Ex: H₀=Membership renewal rates were equal). We failed to reject the two hypotheses around access (Ex: H₀=The Bice Boxerman Continuity of Care Index was equal). And we failed to reject either of the null hypotheses around provider engagement (Ex: H₀=The mean number of messages sent from the organization to patients was the same).

Conclusions: Silently assigning patients who have previously visited care providers to provider panels has operational benefits, but does not change patient engagement, provider engagement, continuity of care, or the patient-reported experience over a six-month period when compared to a matched control population. The relationship between a patient and a PCP cannot be easily replicated with assignment.

References
Identifying Social Isolation from Twitter and Reddit Data during COVID-19

Maryanne Thangarajah, MS*, Rahul Joseph Fernandez, MS*, Jyotishman Pathak, PhD and Braja G. Patra, PhD
Department of Population Health Science, Weill Cornell Medicine, New York, NY, USA

Introduction: The COVID-19 pandemic has restricted social interactions for many, and as a result social isolation and loneliness have escalated during this time. Prior studies have reported that poor social support is a risk factor for worse outcomes in psychiatric populations\(^1\). Social media users frequently express their feelings of social isolation publicly, although till date, there has been little work on identifying loneliness from social media data during COVID-19\(^2\). Guntuku et al.\(^2\) identified the tweets related to isolation manually and then analyzed words for mental well being, use of drug words, temporal patterns, and predictive analysis. To study the effect of social isolation during COVID-19, we identify the social media posts from Twitter and Reddit that are related to social isolation using natural language processing (NLP) techniques.

Data: We developed two text corpora related to social isolation from Twitter and Reddit data. The Reddit corpora was obtained from the ‘r/lonely’ subreddit, which stands as the primary discussion forum for loneliness on the site, and the posts were extracted using the Pushshift Reddit API (https://github.com/pushshift/api) from March 15, 2020 to May 15, 2020. We used TWINT (https://github.com/twintproject/twint) for collecting tweets between the same time frame using the social isolation lexicon\(^3\). A total of 5,668,103 tweets and 53,626 Reddit posts were collected. We annotated tweets and posts if they were related to social isolation (1) or not (0). We calculated the inter-annotator-agreement (IAA) for 100 tweets and posts using two annotators; the IAA were 0.81 and 0.87 for tweets and Reddit posts, respectively. A total of 2,000 random tweets (1:796, 0: 204) and Reddit posts (1: 842, 0:158) were annotated for the presence of social isolation or not. Some examples from tweets and posts are: I’m broken, alone, and afraid (T1, 1); Alone but never lonely (T2, 0). I’m very lonely and I don’t know if nobody cares or if they do care but they don’t show any care because they don’t get that I’m very lonely. Is it my fault? (R1, 1).

Methods: We developed a binary classifier classification based systems to identify tweets and Reddit posts into two primary groups, i.e., related to social isolation (1) or not (0). The preprocessing steps comprise the removal of stopwords, junk words and non-English text. In tweets, user names and hash tags were also removed. We tested the classification performance using several embeddings (Count vectorizer, TF-IDF, word2vec, LSA) and machine learning algorithms (Logistic Regression, Support Vector Machines, Random Forest).

Results and Discussion: The word2vec (minimum word count=6, window size=5, vector size=100) with Random Forest classifier system obtained F-scores of 0.88 and 0.73 using 10-fold cross validation for tweets and Reddit posts, respectively. Tweets are restricted to 280 characters and there is no restriction to Reddit posts. This may be a reason for better classification performance in the case of tweets. Also, Reddit posts were gathered in a specific forum (r/lonely); in contrast all tweets containing the search terms were collected. While annotating data, we observed that some social media users preferred the social isolation during COVID-19. We are planning to annotate tweets and posts with more details such as positive and negative social isolation. In addition to annotation, we plan to experiment with deep learning algorithms using a larger dataset. Further, social isolation tweets and posts can be used for studying different mental health outcomes/implications from social media.

References


* contributed equally
Evaluation of a Multi-Component Phenotype Algorithm for Systemic Lupus Erythematosus across the PCORnet, OMOP, and i2b2 Common Data Models

Steven Tran, BS1, Luke Rasmussen, MS1, Jennifer Pacheco, MS1, Philip Silberman, MA1, Martin Borsje, BS1, Prasanth Nannapaneni, BS1, Daniel Schneider, MS1, Yuan Luo, PhD1, Abel Kho, MD1, Rosalind Ramsey-Goldman, MD1, Theresa Walunas, PhD, MS1

1Northwestern University Feinberg School of Medicine, Chicago, IL

Introduction

Participation in national clinical data sharing networks requires alignment of local clinical data with common data models (CDMs) that support integration of data from multiple sources using a standard framework. As a result, many institutions are implementing multiple CDMs. The aim of this study was to evaluate the performance of a previously published computable phenotype for describing and identifying systemic lupus erythematosus (SLE) patients based on the Systemic Lupus International Collaborating Clinics (SLICC) Classification criteria when it is ported from a local data warehouse to the i2b2, OMOP, and PCORnet common data models.

Methods

We adapted the SLE phenotype to the Northwestern Medicine (NM) Enterprise Data Warehouse (EDW) and NU i2b2, OMOP, and PCORnet instances – all of which are sourced from the NM EDW. Each of the phenotype’s 17 SLICC criteria were determined by a rules-based algorithm built on diagnosis, medication, lab, and procedure codes. We assessed the phenotype over 168 clinician-confirmed SLE patients and 100 healthy controls and calculated agreement between the EDW and CDMs for overall SLE classification, individual SLICC criteria membership, and individual code occurrence level aggregates - count, first and last date, lab values (e.g., agreement in the count of each ICD-10 code between each CDM and the EDW), using Cohen’s kappa (Cκ) and intraclass correlation coefficients (ICC).

Results

For overall SLE classification, agreement of the OMOP and PCORnet datamarts with the EDW was high (Cκ 0.928, 0.802, respectively) while that of i2b2 was low (Cκ 0.328). For the panel of SLICC criteria, OMOP had high agreement for 14/17 SLICC criteria (Cκ 0.792-1.000). PCORnet had high agreement for 13/17 criteria (Cκ 0.708-1.000). i2b2 had high agreement for 10/17 criteria (Cκ 0.813-1.000) with the exception of lab-based criteria. Figure 1 depicts agreement of individual code counts across the CDMs compared to the EDW. OMOP had mostly moderate to high agreement for diagnosis and lab codes (majority of ICCs 0.5-1.0), and high agreement for procedure codes. PCORnet had agreement spread across the spectrum (ICC 0.0-1.0) for diagnosis, lab, and procedure codes. i2b2 had very high agreement for diagnosis and procedure codes (majority with ICC of 1.0), and agreement at both extremes for lab codes. All 3 CDMs had very poor agreement for medication codes (majority with ICC of 0.0).

Discussion and Conclusion

While agreement between the EDW and CDMs at the level of individual codes was relatively poor, agreement at the levels of overall SLE classification and individual SLICC criteria were reasonably high for OMOP and PCORnet, suggesting that inconsistencies at a micro, code level become less apparent at the macro, phenotype level. NU’s i2b2 instance performed very poorly in SLE classification due it not containing lab units as an artifact of its implementation for feasibility testing. This indicates that the CDM purpose and the strategy for mapping primary data to the CDM is an important consideration for understanding the results.

Figure 1. Agreement of individual code counts with reference to the EDW using intraclass correlation (2-way mixed, single score). ICC=1 indicates perfect agreement, ICC=0 indicates no agreement.
Clinical Trial Matching for Lung Cancer Patients

Cetin Urtis PhD¹, Souliya Chittarath¹, Ryan Fischbach¹, Charles Elliott²,
Jimmy Ruiz MD¹, Umit Topaloglu PhD¹

¹Wake Forest School of Medicine, Winston-Salem NC 27157
²Google Inc., Mountain View, CA 94043

Introduction

Despite the potential benefits of oncology clinical trials, only ~5% of adult cancer patients participate, while 55% of patients report a willingness to enroll in trials [1]. Clinical trial recruitment is labor intensive and time-consuming, and serve as a barrier to clinical trial accrual. In this study, we tested natural language processing (NLP) and machine learning (ML) models in order to automate the clinical trial screening process and enhance trial enrollment.

Method

Upon IRB approval, we obtained 4000 non-small cell lung cancer (NSCLC) patients’ clinical notes data from the electronic health records (EHR). We have used clinical trial eligibility criteria (EC) of the Docetaxel + Plinabulin Compared to Docetaxel + Placebo in Patients with Advanced NSCLC (DUBLIN-3) (NCT02504489) study. The trial has 12 inclusion criteria and 12 exclusion criteria. There are only 15 labelled patients where 9 of them are eligible and 6 of them are illegible for the trial. Each patient record and EC were pre-processed and transformed into sentences. We consider two methods for embedding: The pretrained NLP model Bio_ClinicalBERT [2] and Poincare embedding [3] of SNOMED-CT. We calculate cosine similarities for each patient and EC embedded sentences. For each EC, we formed the control distributions by considering the top 10 percentile of scores. Using test patients and EC sentences, we assessed the model accuracy by using a statistical hypothesis testing.

Results and Discussion

In our test, Poincare model outperforms Bio_ClinicalBERT (Table 1). We believe that our preliminary results are promising and both models can be improved. As future steps, BioClinicalBERT is fine-tuned for predicting hospital readmission tasks [2], fine-tuning it for clinical trials tasks may improve results. Additionally, Poincare embeddings may be enriched with other sets of clinical corpora, which would provide better hierarchical representations of clinical text.

Table 1: Our results for both embedding approach

<table>
<thead>
<tr>
<th></th>
<th>Bio_ClinicalBERT</th>
<th>Poincare Embedding</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Eligible Pts</td>
<td>0.60</td>
<td>0.63</td>
</tr>
<tr>
<td>Illegible Pts</td>
<td>0.59</td>
<td>0.60</td>
</tr>
</tbody>
</table>

References

1. Joseph M Unger, PhD, Dawn L Hershman, MD, Cathee Till, MS, Lori M Minasian, MD, Raymond U Osarogiagbon, MD, Mark E Fleury, PhD, Riha Vaidya, PhD, “When Offered to Participate”: A Systematic Review and Meta-Analysis of Patient Agreement to Participate in Cancer Clinical Trials, *JNCI: Journal of the National Cancer Institute*, Volume 113, Issue 3, March 2021, Pages 244 - 257.

736
RNASee: A Rules-Based Tool for APOBEC-Editing Site Identification

Melissa Van Norden¹, Zackary Falls, PhD¹, Sapan Mandloi, PhD¹,², Ram Samudrala, PhD¹,
Peter L. Elkin, MD¹,²,³

¹Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, NY, USA; ²Department of Veterans Affairs, VA Western New York Healthcare System, Buffalo, NY, USA; ³Faculty of Engineering, University of Southern Denmark, Denmark

Introduction

Human protein biodiversity is thought to primarily result from genetic, epigenetic, transcriptional, and post-translational differences, but RNA editing may serve as another source of diversity. APOBEC3A and G, which edit C>U in RNA and single-stranded DNA, have previously been studied as potential causes of oncogenic driver and passenger mutations, and it is possible these enzymes may also play a role in other human disease conditions.¹

Other studies have attempted to quantify the optimality of nucleic acid sequences as APOBEC3A/G substrates,¹ but this is the first study, to our knowledge, of a well-benchmarked program designed to predict RNA editing sites from consensus sequences.

Results

RNASee is a rules-based program developed to find probable sites for APOBEC-mediated RNA C>U editing. It assesses cytidines preceded by a pyrimidine in a stem-loop structure as potential editing sites. These sites are scored based on stem strength and sequential similarity to known editing sites. When run on the human genome, RNASee previously predicted APOBEC editing at 1293 sites of variation related to human disease.²

RNASee was benchmarked on a set of 3086 known APOBEC editing sites across 2343 genes. Known editing sites were extracted from Table S1 of Asaoka et al.³ Currently, RNASee is 59.8% sensitive and 98.5% specific for known APOBEC-mediated RNA editing sites when only the top 10 results per gene are considered. Although recall increases when more sites are returned, specificity falls; at the top 25 cutoff, RNASee is 71.5% sensitive, but only 96.1% specific. A random forest model trained on the same set of sites and genes was 69.5% sensitive, but its specificity was lower than RNASee at 95.0%. When the random forest and top 10 RNASee models were combined using set intersection (only sites returned by both models were considered positive), specificity increased to 99.7%, while sensitivity fell to 45.9%.

Conclusion

RNASee is useful for finding probable APOBEC-mediated RNA-editing sites, editing of which may cause variation in that transcript’s protein product. Further study and refinement of the RNASee platform may help to identify potential genes and variant sites that are altered by APOBEC enzymes for future research into APOBEC3A and G’s influence on human health.

References

Improving the Identification of Substance Use from Clinical Notes with BERT

Stuart J. Waller BS, Cosmin A. Bejan PhD
Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA

Introduction

Social determinants of health (SDoH) – the social and economic conditions that influence health risks and outcomes – have a significant effect on overall well-being across a lifespan. Addressing the underlying factors associated with SDoH is essential towards improving health and reducing health disparities. Despite recent efforts to better integrate SDoH into electronic health records (EHRs), there’s no standardized framework to automatically capture this information. Motivated by the recent success of the Bidirectional Encoder Representations from Transformers (BERT) language model in various natural language processing (NLP) applications, we evaluated 5 BERT models on extracting the status of 3 categories of substance use (tobacco, alcohol, and drug) from clinical text.

Methods

Our dataset contained 2,220 de-identified clinical notes from the Vanderbilt University Medical Center’s EHR data. It was manually annotated with drug use information, extending our previous annotations on tobacco and alcohol. The annotations were performed at mention-level and consisted of 6 categories for tobacco use (Current Smoker, Past Smoker, Never Smoker, Unknown Smoker, Smoker, and Secondary Smoker), 5 similar categories for alcohol use, and 5 similar categories for drug use. We associated each note with an Ever/Never binary category: Never and Unknown were mapped to Never User (NU) while the rest of the categories constituted Ever User (EU). Tobacco had 420 EU and 1,800 NU, alcohol had 663 EU and 1,557 NU, and drug had 49 EU and 2,171 NU.

Leveraging the Ever/Never categories, we trained a binary classification model for each substance – as our goal was to identify patients with substance use phenotypes across the entire EHR. Each note was preprocessed using an NLP pipeline: text was set to lowercase, and punctuation and symbols were removed. Because BERT models allow input sequences of at most 512 tokens, we also employed a keyword-based truncation approach (our average untruncated note was 1,022 tokens). For each substance, we identified a list of relevant keywords and truncated each note around the first detected keyword.

In addition to vanilla BERT-Large, we selected 4 state-of-the-art BERT models with the largest available sizes. We used PubMedBERT and BioELECTRA, both pretrained on biomedical domain text. ELECTRA employs a different pretraining technique and ConvBERT utilizes a different model architecture. We also tried ClinicalBERT, – a model pretrained on EHR data, – but the performance was significantly worse. Each of our 5 models was fine-tuned on the training set of each substance’s truncated dataset and evaluated on the corresponding test set. We reported the binary classification performances on the test set of each substance’s truncated dataset using precision (P), recall (R), and F1 score (F1).

Results

ELECTRA-Large achieved the best performance on tobacco across the board (95.6% P, 93.0% R, 94.1% F1) as well as the highest recall (98.1%) and F1 (94.3%) for alcohol (Table 1). ConvBERT-Base achieved a comparable F1 (94.1%) and the highest precision value (92.7%) for alcohol. All 5 models yielded similar drug scores with the exception of BERT-Large’s distinguishably high F1 (94.1%). PubMedBERT was the only model that didn’t produce a higher F1 for alcohol than tobacco.

Conclusions

Our results suggest that state-of-the-art language models are effective for extracting SDoH information from EHR data. Future work includes identifying more specific substance use phenotypes (e.g., Type, Amount, and Frequency) and constructing a longitudinal profile of these phenotypes at patient-level.

Table 1. Evaluation of 5 BERT models on substance use extraction.

<table>
<thead>
<tr>
<th>Model</th>
<th>Tobacco Use</th>
<th>Alcohol Use</th>
<th>Drug Use</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
</tr>
<tr>
<td>BERT-L</td>
<td>94.0</td>
<td>91.4</td>
<td>92.7</td>
</tr>
<tr>
<td>ELECTRA-L</td>
<td>95.6</td>
<td>93.0</td>
<td>94.1</td>
</tr>
<tr>
<td>ConvBERT-B</td>
<td>95.3</td>
<td>91.4</td>
<td>92.9</td>
</tr>
<tr>
<td>PubMedBERT</td>
<td>95.0</td>
<td>92.0</td>
<td>93.2</td>
</tr>
<tr>
<td>BioELECTRA</td>
<td>94.8</td>
<td>90.3</td>
<td>92.4</td>
</tr>
</tbody>
</table>
Machine Learning Model for De-identifying Health-Related Qualitative Research Documents
Ruichong Wang1, Aditi Gupta, PhD1, Jessica Mozersky, PhD2,
James Dubois, DSc, PhD2, Albert M. Lai, PhD1
1Institute for Informatics and 2Bioethics Research Center, Washington University in St. Louis, St. Louis, MO

Introduction
Qualitative research methods such as interviews or focus groups generate unstructured, non-numeric textual data from study subjects. Sharing health research data is critical for accelerating the translation of research into actionable knowledge that can impact health care services and outcomes. Current de-identification tools are focused on clinical notes and are inadequate for health-related qualitative research data, which are structured differently (e.g., question and answer in interviews, timestamps, and stutters) and contain many more non-HIPAA Safe Harbor (HSH) identifiers. In this paper, we present a de-identification model for extracting HSH (name, dates, etc.) and non-HSH identifiers (organization, age, etc.) from qualitative research data using machine learning (ML) algorithms.

Methodology
We performed an in-depth analysis of two different types of qualitative research datasets. We used a pre-existing gold-standard dataset of ~280,000 words (70 files) to train the ML models [1]. The first dataset consisted of narrative texts written by patients and caregivers in Narrative Inquiry in Bioethics. The second dataset consisted of qualitative interviews with researchers regarding their attitudes towards qualitative data sharing. We trained our ML models using 55/70 files as the training set and used the remaining 15 files as the validation set. Since the de-identification task was a highly imbalanced binary classification task, with only about 1.5% of the total words as identifiers, we used the Synthetic Minority Oversampling Technique (SMOTE) to oversample the data and trained the model on the balanced training set. We included features such as parts-of-speech (POS) tags, stop words, number of non-lower-case letters, total word length, and Word2Vec representations. After our first iteration of training, we found that POS tags, number of non-lower-case letters, and word length were the most important features, so we incorporated information about neighboring words for these features into revised classifier models. This improved the F1 score by 14%. We compared the performance of 5 different machine learning classification algorithms. We also included other state-of-the-art models, such as Scrubber2 from the National Library of Medicine and Philter1 from UCSF for comparison.

Results
Our Random Forest classifier model had the best performance with an F1-score of 0.87 (Table 1). We analyzed the false positives and false negatives generated by all models. Our models mislabeled capitalized abbreviations such as PLOS and VSED, and certain appellations (Doctor, Mr., Mrs.) as identifiers. Our models missed some race and numbers-related words but performed far better than existing tools in terms of precision and recall. In existing tools, Scrubber and Philter, many capitalized non-PHI words (Church, Officer), color-related words (pale, green), weekday words (Monday, Friday), and conversation-filler words (Umm, Uh) were marked as identifiers. As for false negatives, both Scrubber and Philter missed a lot of date-related words (Thanksgiving, 2013, 20th), numbers, proper nouns (Food and Drug Administration), and non-HSH locations (USA, Australia).

Discussion
We developed a novel machine learning model for the de-identification of health-related qualitative research documents. Our model’s precision, recall, and F1 score out-performed existing systems that mostly use rule-based algorithms and mark only HSH identifiers, as they have been developed for de-identification of clinical notes.

References
Extraction of COVID-19 Vaccine Adverse Events Using Natural Language Processing

Liwei Wang, MD, PhD\(^1\), Huan He, PhD\(^1\), Siddhi Mathur, MD\(^{2,3}\), Zehan Li, MS\(^2\), Akintunde Akinseye, MSPH\(^2\), Mukesh Hamal, PhD\(^2\), Jessica Patricoski\(^2\), Alisa Zayas, RN\(^2\), Andrew Wen, MS\(^1\), Sijia Liu, PhD\(^1\), Sunyang Fu, MHI\(^1\), and Hongfang Liu, PhD\(^1\)

\(^1\)Department of AI and Informatics Research, Mayo Clinic, Rochester, MN, USA; \(^2\)Department of Biomedical Informatics and Data Science, Johns Hopkins University, Baltimore, MD, USA; \(^3\)Department of Obstetrics and Gynecology, University of Toronto, Ontario, Canada

Introduction: Since the Food and Drug Administration issued Emergency Use Authorizations for COVID-19 Vaccine in December 2020, there has been an increase in the number of vaccine adverse events (VAEs) that have been reported in the Vaccine Adverse Event Reporting System (VAERS). The VAERS is a spontaneous reporting system for post-licensure vaccine safety monitoring. As the most important adverse events described in the free-text report narrative have been coded using the Medical Dictionary for Regulatory Activities (MedDRA), the data are empowered with the potential to serve as a distant supervision resource for natural language processing (NLP) tool development for extracting adverse events from Electronic Health Records (EHR), given the convenience of annotation using VAERS and the data availability constraints from EHR. As a preliminary study, we aim to develop rule-based NLP algorithms and deep learning models for VAE extraction from VAERS that can be transferred to EHR in the future.

Methods: We first collected datasets associated with the COVID-19 vaccines from the VAERS. We then counted VAE frequencies and selected the 8 high-frequency VAEs (Table 1) for our analyses. To build a benchmark for NLP algorithms, two reviewers annotated these VAEs in 450 reports separately and their inter-annotator agreement (IAA) based on F1 was calculated to ensure the quality of annotation. Among the 450 reports, 350 were used as a training set to develop rule-based algorithms and fine-tune a deep learning model, and the remaining 100 were used as a testing set. The rule-building process using MedTagger as NLP platform involved pattern construction, false-positive and false-negative results analysis, and iteratively refining over a total of 4 rounds. Deep learning model fine-tuning was based on the pretrained Bio_ClinicalBERT model, which was initialized from BioBERT and trained on all Medical Information Mart for Intensive Care (MIMIC) notes. During fine tuning, the max sequence length was set to 512, the batch size was set to 16 and the total number of training epochs was set to 100. We then evaluated the rule-based NLP algorithms and the fine-tuned Bio_ClinicalBERT model on the testing set using standard metrics, i.e., precision, recall, and F-measure based on loose match at document level.

Results and Discussion: Overall IAA for all concepts was 0.82 based on loose match with the highest score for chill (0.94) and lowest for myalgia (0.74), and 0.65 based on exact span match with the highest score for pyrexia (0.91) and lowest for pain (0.45). Our evaluation showed that Bio_ClinicalBERT’s performance exceeded that of the rule-based algorithms slightly (Table 1). Due to the ambiguity between pain and myalgia (muscular pain), the rule-based algorithm showed relatively low performance for both concepts, while Bio_ClinicalBERT model showed low performance for myalgia. We consider the two NLP methods could be combined to improve VAE extraction performance. In the future work, we will add more concepts to rules and Bio_ClinicalBERT Model, fine-tune, combine and test them on real world data from EHR.

<table>
<thead>
<tr>
<th>Concepts</th>
<th>Rule-based algorithm</th>
<th>Bio_ClinicalBERT Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>Chill</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Dizziness</td>
<td>0.9</td>
<td>1</td>
</tr>
<tr>
<td>Fatigue</td>
<td>0.9</td>
<td>1</td>
</tr>
<tr>
<td>Headache</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Pain</td>
<td>0.583</td>
<td>1</td>
</tr>
<tr>
<td>Nausea</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Myalgia</td>
<td>0.636</td>
<td>1</td>
</tr>
<tr>
<td>Pyrexia</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Overall</td>
<td>0.851</td>
<td>1</td>
</tr>
</tbody>
</table>

References
The Translational Drug Combination Knowledgebase (TDCKB): An Informatics Bridge for Cancer Drug Combination Research

Lei Wang, Ph.D.1, Lai Wei, Ph.D.1, Shijun Zhang, MS1, Williams Carson, MD2,3, James L. Chen, MD1,4, Dwight Owen, MD4, Megan Gregory, Ph.D.1, Lang Li, Ph.D.1

1Department of Biomedical Informatics; 2Division of Surgical Oncology, The Arthur G. James Cancer Hospital and Richard J. Solove Research Institute; 3Department of Surgery, Comprehensive Cancer Center; 4Division of Medical Oncology, Department of Internal Medicine, Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA.

Introduction

Utilizing drug combination to target diverse mechanisms has been a promising trend in cancer therapy1. Any combinations should be first evaluated in Phase I clinical trial for safety and tolerance. Powerful statistical designs2 as well as accurate prior knowledge on single drug or drug combination will assist researchers in drug combination Phase I trial design.

There are several primary challenges for current drug combination knowledgebase: structured PK and toxicity data are distributed in different databases3,4; important data elements, such as maximum tolerated dose (MTD) and dosing limiting toxicity (DLT) are missing in any current databases, but in the scientific literature. In addition, oncologists and biostatisticians are not fully aware that drug databases may serve as significant data resources.

TDCKB is the first knowledge base of drug combinations that integrates all drug toxicity and PK data. Most importantly, it synthesizes drug toxicity and PK data and evidence to permit direct generation of new knowledge to assist the design of Phase I trials of cancer drug combinations.

Methods

PK and toxicity data were curated from texts in PubMed, AACR/ASCO abstracts, and Drug Label. The structured data will be integrated computationally from the other publicly available data sources.

A robust data model to store extracted knowledge for single-drug and drug-combination was designed and utilized in TDCKB. We leveraged standard health care terminologies like Drug Bank, Rx-Norm, SNOMED-CT and MedDRA for normalization. As a molecule-centric database, each generic drug name or chemical name will be associated with a unique drug name identifier as the primary key. Meanwhile, a unique drug-combination key will be generated based on the individual drug key.

We implemented User-Centered Design (UCD) process in designing TDCKB user interface (UI). The UI layout will be simple, consistent, purposeful, comprehensive, easy to navigate and informative with readable or scannable typography. The queried critical information for clinical trial design was displayed in the main tab and the complete knowledge with the resource description can be accessed in the detail tab. Drop-down and auto-complete features are also available.

Results

Currently TDCKB integrates pharmacological, PK, toxicity, and drug interactions data from public and internal resources. TDCKB current provides more than 10,000 pieces of pharmacological and pharmaceutical data for approved and investigational drugs including brand name, indication, mechanism of action, and ATC codes. For PK parameters, fraction of metabolized by CYP enzymes for 45 drugs, 1,074 pieces of CYP-based in vitro inhibition constant, and maximum concentration and unbound fraction for 414 drugs can be queried in TDCKB. TDCKB also integrates 88,500 pieces of DDI evidences from PubMed including 1,480 pieces of cancer DDI toxicity evidence, 1,048,000 DDI statements from DrugBank, and over 100,000,000 potential associations between single drug or combination and adverse event from FAERS. Most importantly, TDCKB now provide MTD and DLT data regarding more 98 cancer drug or combinations. This MTD and DLT data was annotated with a fine-grained annotation including patient characteristics, treatment plan, original study design, DLT definition and occurrence, and MTD values.

Conclusion

The TDCKB offers comprehensive and rich knowledge for individual drug and combination. It also provides user-friendly browsing, searching, and filtering functions. TDCKB is available at https://tdckb.bmi.osumc.edu.

References

Integrating Clinical Study and EHR Data for a National Post-Acute Sequelae of COVID-19 (PASC) Initiative

Nich Wattanasin, MS1, Lori Phillips, MS1, Michael Mendis1, 
Victor M. Castro, MS1, Vivian Gainer, MS1, Shawn N. Murphy, MD PhD1,2,3
1Mass General Brigham, Somerville, MA; 2Harvard Medical School, Boston, MA; 3Massachusetts General Hospital, Boston, MA

Introduction

In order to understand the magnitude of the public health impact of Post-Acute Sequelae of COVID-19 (PASC), also known as Long COVID, the NIH has introduced the ‘REsearching COVID to Enhance Recovery’ (RECOVER) initiative. The goal of the nationwide project is to distinguish the clinical spectrum and underlying biology of the SARS-CoV-2 infection over time, including the recovery process and epidemiology of PASC1. To achieve this understanding, it is crucial to be able to integrate clinical study data via electronic data capture (EDC) with electronic health record (EHR) data. At Mass General Brigham, we have developed a web-based ontology browser that enables researchers to easily view, edit, and integrate ontologies that represent data found in both EDC and EHR systems. Specifically, as part of the RECOVER plan, the novel tool allows us to accelerate the extract, transform, and load (ETL) process of REDCap data into a centralized scalable query platform such as Integrating Informatics Bench to Bedside (i2b2), allowing investigators to construct integrated PASC-related queries across disparate data sources such as clinical study data, EHR data, and biospecimen data.

Methods & Discussion

For the RECOVER initiative, cohort participants of the general COVID-19 population are recruited from across the US in a stratified approach ensuring adequate diversity of age, sex, race, and ethnicity. Data for the consented patients is collected through a centralized REDCap, a widely used electronic case report form (eCRF) software. In order to create a cohesive research environment for investigators studying PASC, data captured in REDCap is integrated with other types of relevant data, such as EHR data, biospecimen data, and mobile health data, in a centralized high throughput i2b2 instance2. To achieve this level of data integration, we represent all the query-able data elements in a unified ontology using the newly developed web-based ontology browser.

Already built into i2b2 is a method to export the data dictionary from REDCap which creates a very simple, flat ontology for REDCap data elements. At this point, one would typically edit the computer-generated i2b2 ontology by hand into a more human-readable hierarchy. This manual work can now be achieved using the new ontology browser, allowing one to rename, delete, copy, or move data elements using an intuitive point-and-click user interface (Figure, right).

Results & Conclusion

The development of a new web-based ontology tool helped to accelerate the process of creating a cohesive research environment with integrated data to study PASC. The poster will cover the tool in greater detail including the overall data flow, how-to-use overview, and technical details of deployment. The new tool will also be released as open source to the general i2b2 community as there are several use-cases for it beyond the RECOVER project.

References

Informing Training Needs for the Revised Certified Clinical Data Manager (CCDM™) Exam: Analysis of Results from the Previous Exam

Tremaine Williams, Ed.D.1, Carsten Schmidtke, Ph.D. 2, Kevin Roessger, Ph.D. 3, Vicki Dieffenderfer, Ph.D. 2, Maryam Garza, Ph.D. 1, Meredith Zozus, Ph.D. 4

1 Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA; 2 Department of Human Resource & Workforce Development, University of Arkansas, Fayetteville, AR, USA; 3 Department of Adult and Lifelong Learning, University of Arkansas, Fayetteville, AR, USA; 4 Joe R. and Teresa Lozano Long School of Medicine, University of Texas Health Science Center San Antonio, San Antonio, TX, USA

Introduction

Clinical data managers hold the responsibility for processing the data on which research conclusions and regulatory decisions are based1,2. Thus, it is important that clinical data managers apply effective data management practices. The use of practice standards such as the Good Clinical Data Management Practices (GCDMP) increases confidence in data: emphasizing that the study conclusions likely hold much more weight when utilizing standard practices2.

Methods

A quantitative, descriptive study and application of classic test theory3-4 was undertaken to analyze past data (2011-2017) from the Certified Clinical Data Manager (CCDM™) Exam to identify potential training needs. Each exam question is formally linked to an individual competency and each individual competency is associated with a competency domain5. Using classical test theory, the analysis compared corrected point-biserial correlation values (assessed question-level and domain-level reliability with respect to overall exam performance) and descriptive p-values (percent of data managers who correctly answered an exam question) to assess question difficulty12. Training needs were defined as questions that were 1) very difficult for data managers and 2) psychometrically reliable3-4. Exam questions with descriptive p-values of less than 0.3 represent items were very difficult13. Point-biserial correlation values above 0.15 indicated questions that were psychometrically reliable.

Results

Data across 952 exam attempts (123,760 responses; 130 exam questions). Descriptives included a mean exam score of 91.90, a median of 93, a mode of 92, and a standard deviation of 13.94. The exam had a range of 75 (max. 121; min. 46) and an overall of Cronbach’s alpha of 0.88. Six exam questions yielded descriptive p-values of less than 0.3 and point-biserial correlation values above 0.15. These reflected training needs in four of the profession’s competency domains: design; data processing; programming; and coordination and management tasks.

Conclusion

Overall, the exam tested six of the profession’s eight competency domains, excluding training tasks and personnel management tasks5. Analysis of past CCDM™ Exam results using classic test theory identified training needs reflective of exam takers. Training in the identified areas could, on average, benefit CCDM™ Exam takers and improve their ability to apply effective data management practices. While this may not be reflective of individual or organizational needs, recommendations for assessing individual and organizational training needs are provided.

References

Detecting and Modelling Structured and Random Noise in Reported COVID-19 Cases

Christopher R. Wilson, MS\textsuperscript{1,2}, Adam J. Batten, MS\textsuperscript{1,3,4}, Anushka U. Abeysekara, PhD\textsuperscript{1,2}, Evan J. Carey, PhD\textsuperscript{1,5}

\textsuperscript{1}Department of Veterans Affairs Health Administration; \textsuperscript{2}University of Utah, \textsuperscript{3}UCSF, \textsuperscript{4}A/B Analytics, \textsuperscript{5}Colorado School of Public Health

Abstract

Reported COVID-19 case data has multiple, complex types of noise that make direct utilization difficult. As a result, many research groups preprocess the incoming data to remove undesirable, time-varying periodic variation. However, we were motivated to develop an extension of seasonal decomposition to measure the complex quasi-periodicity, in addition to the residual white noise and batched reporting noise. We then used the information gained from this approach to generate statistically similar synthetic noise profiles.

Introduction

COVID-19 daily case count data exhibits several types of unique noise, chief among them is time-varying periodic variation. State of the art methods for detecting periodicity generally allow for a single periodicity\textsuperscript{1} or a slowly changing periodicity\textsuperscript{2}. Additionally, such methods do not account other idiosyncratic noise types, such as batched data permutations where one or more days or underreported, then concatenated on an adjacent report day. Our method is capable of detecting arbitrarily short periodicity durations with a statistical measure of confidence.

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{sample_periodicity.png}
\caption{Sample time-varying periodicity detected in COVID-19 daily reported case data from Utah. Blue sections denote a period of 7, red sections denote periods that failed a statistical measure of periodicity.}
\end{figure}

Once such information is captured, we can then use the output of this model to then create new, realistic noise profiles to be added to simulated data for downstream use via a Hidden Markov Model. Our method retrieves the correct periodicity on synthetically noised data at an accuracy above 90\% for random noise to periodic noise ratios less than 5:1.

Conclusion

Our approach adds a valuable tool to groups wishing to train models with disease data when data is sparse using statistically similar noise profiles to observed data. It additionally allows for fine-grained investigations of periodicity in such data. The code for our method is implemented in python and is available on a public github. A complete description of the method will be published in a refereed journal with feedback from our poster if accepted.

References

Data Augmentation with Word Embeddings for Medical Concept Extraction

Wenzhe Yang, Youngjun Kim, Timmy O’Connell, Arielle Redfern, Xiang Zhou, Rong Chen
Sema4, Stamford, CT, 06902, USA

Introduction: Data augmentation has the potential to effectively increase the amount of labeled data when the system struggles to achieve good performance with limited labeled data. Relatively little research has been done on data augmentation in clinical natural language processing (NLP). Kang et al.1 introduced a UMLS-based (Unified Medical Language System) data augmentation method (called UMLS-EDA). This knowledge-based method finds UMLS semantic concepts in text and replaces them with synonyms. Extending this approach, we propose an automated data augmentation method that selects synonyms based on word similarity derived from word embeddings. Our method, called Word Embedding-based Easy Data Augmentation (WE-EDA), can employ any word vector representations that can be pre-trained with large amounts of unlabeled text in the target domain. Our NLP task involves extracting cancer treatments and the dates and durations of these treatments. We tackle this task as a sequential labeling problem to assign a class label to each word in a sentence. We hypothesize that data augmentation using word embeddings pretrained with clinical text can improve medical concept extraction performance for cancer treatments recorded in an unstructured manner. Our comparative assessment showed that the information extraction model trained with the additional labeled data outperformed the model with the original labeled data.

Methods: We created a new text collection representing prostate cancer patients treated at Mount Sinai. Our corpus consists of 120 training documents and 80 test documents manually annotated by medical experts. The average length of each document is 2,042 words. The inter-annotator agreement between two annotators was measured using F1 score, and their IAA was 91.6%. We implemented sequential labeling classifiers using a Bi-LSTM+CRF architecture that included a bidirectional LSTM (Bi-LSTM) network and CRF (conditional random fields) decoding. We used a combination of static word embeddings and contextualized word embeddings. We pretrained fastText word embeddings with the MIMIC-III clinical dataset (version 1.4). We used the uncased BlueBERT-Large embeddings2 pretrained on PubMed abstracts and MIMIC clinical notes. We experimented with and compared three approaches: 1) using the original training set, 2) adding new training examples generated by UMLS-EDA1, and 3) adding new examples from WE-EDA, a novel contribution to this study. We performed the UMLS-EDA method with hyperparameters following Kang et al.’s presets1. The WE-EDA method replaced words with synonyms extracted from MIMIC-based fastText embeddings. We used cosine similarity to identify the 10 nearest neighbors and then randomly selected one word for replacement. Both augmentation methods doubled the size of the training data.

Results: Table shows results for each concept type and micro-averaged F1 scores. For each concept type, the best results appear in boldface. For strict matching, both the concept type and text span must exactly match the reference annotation. Lenient matching considers the system-detected concept correct if it overlaps with the reference annotation. Both augmentation methods produced higher F1 scores than the model with the original training data. The WE-EDA method allowed for more accurate classification with a 1.3% F1 gain. We used a paired t-test to measure statistical significance. The F1 score performance of the WE-EDA method is significantly better than the model with the original labeled data at the p < .05 significance level, but not significantly better than the UMLS-EDA method.

Conclusion: This study showed that our word embedding-based data augmentation can improve medical concept extraction when the amount of labeled data is small. Our future research involves finding word similarities from contextualized word embeddings.

References

Table. F1 score of each method on the test set

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>Original</th>
<th>UMLS-EDA</th>
<th>WE-EDA</th>
<th>Original</th>
<th>UMLS-EDA</th>
<th>WE-EDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>1,083</td>
<td>89.8</td>
<td>89.7</td>
<td>90.7</td>
<td>91.7</td>
<td>92.4</td>
<td>92.6</td>
</tr>
<tr>
<td>Date</td>
<td>378</td>
<td>79.8</td>
<td><strong>82.2</strong></td>
<td>81.2</td>
<td>83.7</td>
<td><strong>84.9</strong></td>
<td>84.4</td>
</tr>
<tr>
<td>Duration</td>
<td>39</td>
<td>55.7</td>
<td>62.7</td>
<td><strong>63.2</strong></td>
<td>84.2</td>
<td>83.5</td>
<td>82.7</td>
</tr>
<tr>
<td>Overall</td>
<td>1,500</td>
<td>86.3</td>
<td>87.1</td>
<td><strong>87.6</strong></td>
<td>89.5</td>
<td><strong>90.3</strong></td>
<td><strong>90.3</strong></td>
</tr>
</tbody>
</table>

Notes: Boldface indicates the best performance for each concept type.
Understanding Clinician Workflows to Design AI Risk Prediction Models

Olivia Zhang¹, Shruthi Chari, MS¹,², Fernando Suarez, MD³,
Daniel M. Gruen, PhD², Prasant Acharya², Oshani Seneviratne, PhD²,
Pablo Meyer, PhD¹, Deborah L. McGuinness, PhD², Prithwish Chakraborty, PhD¹
¹Center for Computational Health, IBM Research, Yorktown Heights, NY;
²Rensselaer Polytechnic Institute (RPI), Troy, NY; ³IBM Watson Health, Cambridge, MA

Overview: Incorporation of artificial intelligence (AI) models and explainable AI (XAI) methods into clinical decision support tools (CDSS) necessitates a deep understanding of the end-user. Clinicians are one of the most important user groups for applying AI tools to solve healthcare-related problems. Focusing on risk prediction, a setting of prime importance in CDSS, we share an understanding of the clinician’s workflow, unmet needs, and desired interactions through an interview-guided process that leverages past user-centric research principles. As a case study, we consider risk of comorbidities among Type-2 Diabetes (T2D) patients. We also share some early results about the design setup.

Method: To understand the clinical use-case and determine the context of AI tools, we followed a sequence of user-centric research principles. Specifically, we (1) defined the scope of our tool’s capabilities, (2) identified the end-user/target persona who would most benefit from our tool, and (3) scoped the most relevant contexts and usage scenarios for our tool. We achieved steps (2-3) by conducting semi-structured interviews with a clinician with both patient-facing and clinical-research expertise. We received feedback around both general workflow needs and needs related to our specific use-case. Questions included: What is the PCP’s current workflow? What are the pain points? What scenario(s) might benefit from a risk prediction tool? How would insights from specific features inform a treatment plan?

Findings and Next Steps: Through our interviews, we were able to (a) identify common pain points for clinicians across disciplines, (b) walkthrough typical contexts of T2D patient care by clinician type and patient characteristics, and (c) understand how risk predictions can help clinicians improve patient care planning. Following these findings, we prototyped an XAI-enabled dashboard to predict the risk of comorbidities among T2D patients and conducted structured interviews with a clinician panel for the ‘new patient scenario’ (see Fig 1). We found that clinicians appreciate connections to patient history to interpret patients’ risk scores, strengthening our findings from use-case interviews that scenario-based AI support is needed. Overall, the findings indicate the importance of grounding such work in clinical scenarios. During our poster presentation, we will share detailed scenario insights and relevant themes from clinician panel interviews. In summary, we find that AI tools should be designed to be adaptable to various usage scenarios, and when possible, to interact with the clinician’s evolving mental models in a circular process of support.

Acknowledgements: This work is supported by IBM Research AI through the AI Horizons Network.

References
Quality Evaluation of Image-Based Survival Data from Clinical Trials

Yili Zhang, PhD, Samir Gupta, PhD, and Subha Madhavan, PhD
Innovation Center for Biomedical Informatics,
Georgetown University Medical Center, Washington, DC

Introduction
With the accessibility of medical publications in digital versions increasing, rich information provided by it becomes a new source for secondary research. Researchers have made efforts on deriving unstructured text in publications but there has been no study about how to readout inherent raw data in medical statistical images, which can be helpful in further medical research such as data synthesis. As an important method to present time-to-event survivor information, Kaplan-Meier(KM) analysis and plots have been widely adopted in clinical trials, to indicate the effects of intervention used. Composed by raw event data of patients in clinical trials, how does KM plot perform in displaying these data to readers is a question. Thus, this research will 1) develop an algorithm to extract data of patients at risk from KM plots and 2) evaluate the quality of KM plots based on the accuracy of data presented in these plots.

Methods
To examine the most up-to-date KM plots, clinical trial presentations at the American Association for Cancer Research Annual Conference (AACR) 2021 were collected. KM plots were cropped from each presentation PDF file and selected by following criteria: 1) KM plots should come with numbers below the figures denoting the number of patients ‘at risk’ in each group, which will be the ground truth to be tested; 2) characters in plots are human-readable; 3) plots only have a single-color background; 4) censoring marks on KM curves should be vertical tick-marks. Next, X-Y axes and intervals of the coordinates were identified and the region the X-Y coordinates were extracted. For the extracted region, irrelevant elements for the purposes of this study such as description tables, notations, and dash lines were detected and removed to obtain only KM curves. Each KM curve was the main target to be examined in this study. After that, the location of each tick-mark was derived for censored event information. Furthermore, based on the monotone decreasing specialty of the KM plot, every drop location and drop percentage on the KM curve was derived for death event information. Then the numbers of patients at risk were calculated at each time interval, which will be measured with the ground truth presented below the figure. The normalized root mean square error (nRMSE) was used as the metric for evaluation, with the difference between the maximum and minimum values being the normalization factor. Finally, KM plots with nRMSE lower than 30% were considered as high quality and vice versa.

Results
In total, 22 KM plots were selected for data extraction, which have 40 KM curves. An example of the process in main steps is shown in Figure 1. The overall nRMSE is 0.45±0.29 with 95% confidence interval. There are 9 high-quality KM plots with 18 KM curves, which comes with nRMSE 0.18±0.01. However, the remaining 13 KM plots with 22 curves are of low quality, with nRMSE as high as 0.66±0.20. There are various reasons for low-quality plots: 6 plots have uneven line width, 3 plots have chaotic tick-marks, 2 plots are in low resolution, and KM curves in 1 plot were covered by dash lines, which all create difficulty for the algorithm to read accurate censoring and death events.

Conclusion
This study developed an algorithm to process KM plots on publications and read the row data in it. In high-quality plots, the algorithm has a good performance in comprehending the patients at risk at each time interval. However, this study also acknowledges that low-quality images often fail to present accurate raw data to readers. Images of high quality are preferred for secondary analysis.

References

Figure 1: (a) X-Y axes detected and marked in red, (b) Not-curve elements detected by bounding marked in blue box, (c) Censored events detected and marked in black.
CDS - Causal Inference with Deep Survival Model and Time-varying Covariates
Jie Zhu, MRes¹, Blanca Gallego, PhD¹
¹Centre for Big Data Research in Health (CBDRH), UNSW, Sydney, NSW, 2052, Australia

Abstract

Causal inference in longitudinal observational health data often requires the accurate estimation of treatment effects on time-to-event outcomes in the presence of time-varying covariates. To tackle this sequential treatment effect estimation problem, we have developed a causal dynamic survival (CDS) model.

Introduction

CDS estimates potential hazard functions under time-varying binary treatments using an ensemble of recurrent deep subnetworks. Following the potential outcomes framework, treatment effects are estimated as the difference in survival probabilities between the given treatment and control regimes. The ensemble training is used to capture the uncertainty of network estimation from varying random seeds, and a dedicated propensity score layer is used to adjust the selection bias presented in the data.

Results

Using simulated survival datasets, the CDS model showed good causal effect estimation performance across scenarios of varying sample dimensions, event rate, confounding, and overlapping. However, increasing the sample size was not effective in reducing the bias from a high level of confounding. In two large clinical cohort studies, our model identified the expected conditional average treatment effects and detected individual treatment effect heterogeneity over time.

Discussion

The use of a propensity score layer and potential outcome subnetworks helps correcting for selection bias. However, the proposed model is limited in its ability to correct the bias from unmeasured confounding, and more extensive testing of CDS under extreme scenarios such as low overlapping and the presence of unmeasured confounders is desired and left for future work.

Conclusion

CDS fills the gap in causal inference using deep learning techniques in survival analysis. It considers time-varying confounders and treatment options. Its treatment effect estimation can be easily compared with the conventional literature, which uses relative measures of treatment effect. We expect CDS will be particularly useful for identifying and quantifying treatment effect heterogeneity over time under the ever-complex observational health care environment.
Prediction of Lung Cancer from Machine Learning Models based on Colorectal and Ovarian Cancer Screening Trial (PLCO) and National Lung Screening Trial (NLST)

Nansu Zong 1, Jie Na 1, Yuan Luo 2, Guoqian Jiang 1
1 Department of Artificial Intelligence and Informatics Research, Mayo Clinic, Rochester, MN, 2 Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL

Introduction
Many clinic trials, such as Colorectal and Ovarian Cancer Screening Trial (PLCO) and National Lung Screening Trial (NLST), have correlated reduction in the mortality for lung cancer patients with early screenings, thereby shaping recommended lung cancer screening policies (e.g., United States Preventive Services Task Force (USPSTF)) are given to determining screening eligibility. The current screening policy used in USPSTF is based on age and smoking years, while other individual’s risk factors (e.g., ethnicity and family history of lung cancer) are not yet considered. While there are computational risk prediction models developed, those models are mostly considered as statistical models [1]. Machine learning methods have the potential to improve model performance. Here, we have adopted four machine learning models and a deep learning model to compare them with a baseline statistical model—specifically logistic-regression, over the two datasets, PLCO and NLST.

Data and Method
Two randomized controlled screening trials, PLCO and NLST, were used in this study. For PLCO, smoking participants in the CXR arm (n = 40,600) and control arm (n = 40,072) were included. For NLST, we used data from participants in CT arm (n = 26,722) and chest radiography (CXR) arm (n = 26,730). A logistic-regression model named PLCO M2012 [2] trained with the PLCO control group of smokers was used as the baseline model. We adopted four machine learning models, logistic-regression (LR), Random Forest (RF), J48, and Naive Bayes (NB), and one deep neural network (four fully connected dense layers) as our predictive models in comparison.

Experiments and Results
We selected 22 variables from the variable that achieved the best performance in nine predictive models [3] and ran six modes over the three tasks based on the PLCO control arm, PLCO CXR, and NLST all arms. As Figure 1 shows, PLCOM2012 still achieved the best results in all three tasks (80.80% for PLCO control, 80.40% for PLCO CXR, and 69.40% for NLST all arms).

Conclusion and Discussion
This study investigated the existing machine learning models and deep learning models in lung cancer prediction based on cancer trials. While machine learning and deep learning models did not outperform the PLCO M2012, the gap between the best models is negligible (77.84% for PLCO control, 78.96% for PLCO CXR, and 68.70% for NLST all arms). Our future work will adapt a more advance deep learning model to improve performance.

Acknowledgments
This work was supported by funding from the NIH NIGMS (K99GM135488) and NIH FHIRCat (R01 EB030529).

Reference
The Open-Source EMERSE (Electronic Medical Record Search Engine) Tool

Presenters:

Mark Beno, M.S.M., Case Western Reserve University, Cleveland, Ohio
Sunah Song, Ph.D., Case Western Reserve University, Cleveland, Ohio
Paola Saroufim, PharmD, MPH, MBA, Case Western Reserve University, Cleveland, OH

Abstract:

The Cleveland Institute for Computational Biology (CICB) has deployed the open-source EMERSE (Electronic Medical Record Search Engine) tool that is facilitating previously unfeasible chart review research projects and clinical cohort discovery. Developed, maintained, and enhanced by Dr. David Hanauer’s team at the University of Michigan, EMERSE is installed at 6 academic hospitals covering over 9 million patients with 400+ million indexed documents searchable against a clinically curated synonyms collection of 1.5+ million words and phrases.

Specific Purpose of EMERSE:

The EMERSE tool creates a fully indexed and searchable repository of all the unstructured clinical notes that are traditionally only accessible via painstaking review within an EHR system that is limited to reviewing one patient at a time. EMERSE allows the clinical researcher to quickly search across all notes either at the patient-level or across the entire patient population using a collection of clinically curated set of over 1.5 million words and phrases. The power of this methodology will prove to be a game changer in research studies involving chart review, rare diseases, and clinical trial cohort determination where a structured search can now be passed to EMERSE for additional and efficient cohort validation.

Degree of deployment of EMERSE:

Developed by Dr. David Hanauer in 2005, the EMERSE tool is now deployed (over the last 3 years) at 6 academic medical centers/hospitals: University Hospitals of Cleveland (via the Cleveland Institute for Computational Biology at Case Western Reserve University), University of Michigan, University of North Carolina, University of Cincinnati, University of Kentucky, and the University of California at San Francisco. The tool is being implemented at Columbia Cancer Center, MD Anderson Cancer Center, and the Dana Farber Cancer Center. The University of Virginia, UC Irvine, Moffitt, and the City of Hope are currently in the pre-implementation phase of EMERSE. EMERSE has been cited in over 447 papers and has supported multiple PhD dissertation research projects. Finally, EMERSE has a network option that will allow current and future implementation partners to perform federated searches across the soon to be activated EMERSE network.
MedTator: A Lightweight Interactive Multi-Docu...
Scalable Understanding of Coding Differences in a National Clinical Research Network Through Multi-site Visualizations

Jeffrey G. Klann, PhD\textsuperscript{a}; Griffin M. Weber, MD, PhD\textsuperscript{a}; Michael Mendis\textsuperscript{a}; Michele Morris\textsuperscript{a}; Shawn N. Murphy, MD, PhD\textsuperscript{a,b}

\textsuperscript{a}Harvard Medical School; \textsuperscript{b}Research Information Science and Computing, Massachusetts General Brigham; \textsuperscript{c}Laboratory of Computer Science, Massachusetts General Hospital; \textsuperscript{d}Beth Israel Deaconess Medical Center – All in Boston, MA; \textsuperscript{e}University of Pittsburgh School of Medicine, Pittsburgh, PA

Introduction

Clinical research data networks are proliferating, with the promise that they will enable rapid discovery and feedback into the healthcare system. However, healthcare data is complex and non-uniform across sites, systems, and time. In order to effectively utilize the data in clinical data repositories and data research networks, methods must be developed to help researchers understand how the data are arranged in these different settings. Large clinical data networks do not currently provide integrated, nimble, and actionable mechanisms to understand granular cross-site data quality.

Informatics for Integrating Biology and the Bedside (i2b2) is a clinical data warehousing platform used at over 200 locations worldwide, including in large data research networks like ACT and large portions of PCORnet. Key to i2b2’s design is its ontology system, which supports hierarchies to encode the granular knowledge of biomedicine. These hierarchies can be customized as needed, but hierarchies can also be standardized across systems, e.g. in a data network, in order to harmonize the types of data that are shared.

Despite standardization with ontologies, coding choices can vary dramatically across sites, because e.g. a laboratory test like hemoglobin A1C could be encoded with any of several valid LOINC codes. Modern ontologies contain millions of terms (the ACT ontology has 2.5 million), and sites tend to use only a small portion (but there is variance across sites).

Therefore it becomes important to analyze the distribution and study similarities/differences in ontology usage at sites when i2b2 sites are linked in data networks. This could be accomplished by counting the number of patients with each ontology term (and all its sub-terms). Such counts can be displayed in the i2b2 query tool, and these counts could be collected and aggregated across sites, to examine code distributions across an entire network. One can imagine such an approach being incorporated into the standard ETL process at each site’s data refresh, providing a regularly-updated and highly-granular topography of a network’s data.

Description of System

To actualize this, we developed several tools that work in unison. First, we developed a high-performance method for each site to count the number of patients with each medical concept, designed entirely in SQL, leveraging the set operations and indexing that make relational databases so efficient. The SQL tool allows an i2b2 site to compute all of its one-item “cohorts” at one time, a scale with which SHRINE (the i2b2 network system) was never designed to handle. This method also offers enhanced patient privacy protection by adding Gaussian noise and a low-patient threshold to the counts, following the same design as the SHRINE network tool.

Second, we developed an aggregation tool in Python that collects the reports and creates an i2b2-style ‘fact table’ of counts in a SQLite embedded database. The resulting rapidly queryable, fact-oriented data-density table is linked to the i2b2 ontology table through the ontology item key. The tool also automatically calculates statistics on the collected data, including network averages and standard deviations. Third, we developed reports using Python, Plotly, and Jupyter notebook that utilize these reports to discover site outliers, missingness, and differences in data coding across sites. In the future, we hope to enhance our tools to explore variations in mapping (e.g., choice of specific codes for a given laboratory test).

State of the System

In a recently-completed pilot, we gathered data from seven sites in the ACT network, which we used to develop the tooling and reports. We are now engaging more ACT sites and enhancing our reports. The counting scripts are nearly complete and are already available in the i2b2 Github. They will be officially released at the end of 2021 with i2b2 version 1.7.13. The aggregation script and preliminary versions of the reporting tools are also complete, and the final versions will also be made available on Github before the Summit in 2022.
Clinical Informatics as a Driver of Innovation and Quality of Care in Pediatric Integrated Behavioral Health

Jelena G. MacLeod, MD, MHS, Kammarache N. Asuzu, MD, MHS, Brian Padilla, PhD, Maryellen Flaherty-Hewitt, MD, Linda C. Mayes, MD

1Yale Child Study Center, New Haven, Connecticut; 2Fair Haven Community Health Care, New Haven, Connecticut; 3Yale Department of Pediatrics, New Haven, Connecticut

Introduction

Behavioral health diagnoses in children are common, impairing, and frequently under-identified. The COVID-19 pandemic has precipitated additional increases in youth anxiety, depression, suicidality, substance use, and eating disorders. Integrated behavioral health in pediatric primary care settings offers an important bridge to behavioral health services, and an opportunity for early identification, innovation, teaching, and discovery using an interdisciplinary approach. It is more critical than ever to optimize integrated behavioral health capacity and quality of care while minimizing additional burden on primary care providers, behavioral health clinicians, and the broader system of care. Clinical informatics is emerging as a novel solution to enhance delivery and outcomes of pediatric behavioral health services. Using customizable Electronic Health Records (EHRs) such as Epic allows for more streamlined behavioral health screening, decision-support for pediatricians, workflow enhancement across pediatric settings, and generation of individual and population-level healthcare data. This systems demonstration will cover several innovative EHR-based behavioral health interventions used at our pediatric integrated behavioral health clinic at the New Haven Primary Care Consortium.

Systems Demonstration Content

1. Clinical decision support for pediatric attendings and residents, ranging from EHR-embedded decision-tree pathways to pre-populated referrals and psychiatric medication prescriptions. Our clinic uses an Epic-based integrated behavioral health screening pathway to guide pediatrician decision-making and auto-populate appropriate referrals to child and adolescent psychiatry, health psychology, or brief psychotherapy. We recently launched an additional depression-prescribing pathway that empowers pediatric residents and attendings to safely prescribe anti-depressants while triaging cases that require more intensive support from child and adolescent psychiatry. Future EHR-embedded pathways include suicidality assessment, prescribing guidelines for managing anxiety disorders, and management of behavioral disruptions in pediatric settings.

2. Cross-systems concordance of electronic referrals to integrated behavioral health services from safety-net pediatric settings including the pediatric emergency room. We are focusing on referral automation, consistency, and measurement as pediatrics settings consult our integrated behavioral health team. In primary care pediatrics, referrals are embedded into decision-tree pathways and auto-populated into the pediatrician’s EHR workflow. The same electronic referral is newly available in the pediatrics emergency room as well, to promote continuity of care and throughput for children presenting to the emergency room for psychiatric evaluations.

3. Future directions: harnessing registry-based population health management for triage, risk-stratification, and data analytics using EHR medical and behavioral health data. The Epic EHR system has built-in population management tools that have been optimized for medical conditions but were not designed for patients with behavioral health concerns. This discrepancy creates barriers towards connecting patient behavioral health and medical data to inform resource allocation, enhance cost-effectiveness, and pave the way to predictive analytics and risk stratification of both psychiatric and medical outcomes in youth presenting with behavioral health concerns. To address this gap, we are in the design stages of an integrated behavioral health registry with full EHR functionality.

References

Disease Progression Modeling Workbench 360

Parthasarathy Suryanarayanan¹, Prithwish Chakraborty¹, Piyush Madan¹, Kibichii Bore², William Ogallo², Rachita Chandra¹, Mohamed Ghalwash¹, Italo Buleje¹, Sekou Remy², Sheryans Sethi¹, Shilpa Mahatma¹, Pablo Meyer¹, Jianying Hu¹
¹Center for Computational Health, IBM Research, USA; ²IBM Research, Kenya

Motivation: Disease Progression Modeling (DPM) pertains to characterizing the progression of diseases and their comorbidities over time, including disease staging, patient trajectory analytics, prediction, and time-to-event estimations for key disease-related events. DPM has a wide range of applications for providers, payers, and pharmaceutical companies. However, the complexity of building effective DPM models can be a road-block for their rapid experimentation and adoption. While the standardization of data models and tooling for data analysis/cohort selection alleviate some of these complexities of collaborative research and development, there are unmet needs in terms of experiment tracking, model reproducibility and prototyping.

Methods: In this demo, we introduce Disease Progression Modeling Workbench 360 (DPM360) open source project, an easy-to-install system that accelerates research and development of DPM models. It manages the entire modeling life cycle, from data analysis (e.g., cohort identification) to machine learning algorithm development and prototyping. DPM360 augments the advantages of data model standardization and tooling (OMOP-CDM, Athena, ATLAS) provided by the widely-adopted OHDSI initiative with a powerful machine learning training framework, and a mechanism for rapid prototyping through automatic deployment of models as containerized services to a cloud environment. DPM360 consists of 4 key components: (1) Lightsaber: an extensible training framework which provides blueprints for the development of disease progression models. It consists of pipelines for training, hyperparameter tuning, model calibration, and evaluation. Lightsaber comes with a reusable library of state-of-the-art machine and deep learning algorithms for DPM (e.g., LSTM for in-hospital mortality predictions). It integrates naturally with the OHDSI stack to extract patient data using the cohort definitions in the OMOP data store. These raw datasets are further transformed to machine learning features using the OMOP common data model. (2) Tracking provenance of all aspects of model building is essential for trust and reproducibility - experiments run using Lightsaber are automatically tracked in a Model Registry including model parameters, problem specific metrics, and model binaries allowing data scientists to identify algorithms and parameters that result in the best model performance. (3) The Service Builder component automatically converts models registered in Model Registry into microservices and installs them in the target cloud environment (Kubernetes or OpenShift cluster). These microservices can be tested with a Swagger based interface. (4) The Installer component installs the fully functional DPM360, including OHDSI tools, Model Registry and Service Builder into a cloud cluster using Helm charts. DPM360 is cloud agnostic and built ground-up as collection of re-usable and pluggable components (see Fig. 1). A detailed description of the target use cases, usage examples, architecture, and roadmap is available in our public GitHub repository.

Impact and Real-world Usage: DPM360 enables researchers and informaticians to focus on model specification and analysis while also automatically assisting them by enabling reuse, adoption of best practices, and empowering them to distribute models as services with little effort. The Lightsaber component has been successfully used for rapid experimentation and analysis of real-world data. DPM360 is also being made available in the secure N3C enclave and has been accepted for demonstration in OHDSI collaborator showcase. We plan to continue developing DPM360 and report on its impact on promoting collaborative research and rapid commercialization of disease progression models.

References