JMIR Medical Informatics

Impact Factor (2023): 3.1
Volume 10 (2022), Issue 9   ISSN 2291-9694   Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Viewpoint

Original Papers

Review

Corrigenda and Addenda

Viewpoint

Using Electronic Health Records for the Learning Health System: Creation of a Diabetes Research Registry

Brian J Wells[1*], MD, PhD; Stephen M Downs[2*], MD, MS; Brian Ostasiewski[3*], BS

[1]Department of Biostatistics and Data Science, Wake Forest University School of Medicine, Winston Salem, NC, United States

[2]Department of Pediatrics, Wake Forest University School of Medicine, Winston Salem, NC, United States

[3]Center for Biomedical Informatics, Wake Forest University School of Medicine, Winston Salem, NC, United States

[*]all authors contributed equally

Corresponding Author:
Brian J Wells, MD, PhD
Department of Biostatistics and Data Science
Wake Forest University School of Medicine
1 Medical Center Blvd
Winston Salem, NC, 27157
United States
Phone: 1 336 416 5185
Email: bjwells@wakehealth.edu

Abstract

Electronic health records (EHRs) were originally developed for clinical care and billing. As such, the data are not collected, organized, and curated in a fashion that is optimized for secondary use to support the Learning Health System. Population health registries provide tools to support quality improvement. These tools are generally integrated with the live EHR, are intended to use a minimum of computing resources, and may not be appropriate for some research projects. Researchers may require different electronic phenotypes and variable definitions from those typically used for population health, and these definitions may vary from study to study. Establishing a formal registry that is mapped to the Observation Medical Outcomes Partnership common data model provides an opportunity to add custom mappings and more easily share these with other institutions. Performing preprocessing tasks such as data cleaning, calculation of risk scores, time-to-event analysis, imputation, and transforming data into a format for statistical analyses will improve efficiency and make the data easier to use for investigators. Research registries that are maintained outside the EHR also have the luxury of using significant computational resources without jeopardizing clinical care data. This paper describes a virtual Diabetes Registry at Atrium Health Wake Forest Baptist and the plan for its continued development.

(JMIR Med Inform 2022;10(9):e39746) doi:10.2196/39746

KEYWORDS

electronic health record; EHR; Learning Health System; registry; diabetes

Background

The first electronic health records (EHRs) were developed to support clinical care, but later became primarily focused on billing after the creation of diagnosis-related group (DRG) codes [1]. DRGs are intended to provide precise estimates of resource use across different hospitals. Unfortunately, the documentation necessary to support billing frequently does not result in a data content and structure ideal for the secondary use of these data for research. Safran et al [2] outlined a framework for using EHR data for secondary purposes. The use of EHR data for research purposes has increased significantly at Wake Forest and elsewhere over the past several years. However, complex outcome studies that use data at different time points are still rare. Research investigators struggle with the processing and statistical analyses of EHR-derived data due to the time-varying nature, inconsistency, inaccuracy, lack of documentation, and incompleteness of clinical data. Investigators report that the amount of time spent deciphering and *cleaning* these data make many research projects impractical. A systematic review of the use of EHR data for population health identified several common barriers for the use of these data for population health, of which missing data were most cited [3]. Handling of missing data requires an understanding of the reasons for missing data, some of which can be project-specific reasons and related to decisions about how to handle them. Simply excluding patients with missing data may reduce sample size and can lead to biased results. One common method for handling missing data in EHR projects is multiple imputation, where statistical models are used to estimate values for missing data elements [4]. Investigators may be unfamiliar with these techniques or may lack the knowledge and skills to perform the task in a robust fashion. Imputation will be one of the services provided to investigators. Prior to imputation, it is necessary to explore data to identify implausible values that may arise due to inaccurate measurements or data entry errors. Textbox 1 highlights some of the data-processing steps that may be required prior to using clinical data for statistical analyses.

Textbox 1. Common data-processing steps required to analyze clinical data.

Common data-processing steps

- Removal of extreme values
- Correction of erroneous entries
- Imputation of missing values
- Calculation of predefined variables
- Determination of active medication classes on a given date
- Calculation of dates and time to events
- Creation of a single analytic data set with a single row per patient from normalized tables

Research Registries

Research registries derived from the EHR can provide a foundation that improves the efficiency for research projects in a specific disease area. Registries can provide formal documentation of the institutional knowledge gained over time from previous investigations and input from the research community. The sharing of experiences provides an opportunity for critical evaluation of the data from investigators with different areas of expertise, leading to improved data quality and knowledge of the data necessary for interinstitutional projects. Preprocessed data, predefined variables, linkage with other institutional databases (eg, echocardiogram and pulmonary function tests), linkage with external data (eg, American Community Survey and North Carolina Death Registry), and creation of statistical functions can greatly reduce the time and cost of secondary data analyses. Data preprocessing can include data cleaning (eg, removal of extreme values and imputation of missing data), which can reduce the risk of biased results but would be inappropriate for clinical data. Prescription medications provide another opportunity for data preprocessing. For example, calculation of dosages and quantity of medications can be determined by applying regular expressions to free text prescription instructions. Research registries also provide a mechanism for pooling knowledge and resources from disparate research areas. For example, chart reviews conducted for one specific research study could provide important knowledge that benefits all users of the registry. Similarly, researchers could pool resources to purchase external data (eg, National Death Index or Centers for Medicare & Medicaid Services [CMS] data) that will benefit all. Research registries provide a repository for collecting research items not intended for the legal medical record to support activities such as creating risk prediction models and conducting epidemiologic studies. Furthermore, the research registries also provide potential populations of patients for research studies (clinical trials, pragmatic trials, implementation science, population health, and medical informatics). The increased recognition and credibility of an institution’s clinical data for research that comes with a successful registry can improve the chances for research funding.

Population Health Registries

There has been a proliferation of population health registries in EHR systems. These real-time data are necessary for clinical care, and these registries are designed to put minimal burden on the EHR system, especially given that they are using the live EHR system, which is critical for clinical care. These types of EHR-based population health registry tools (eg, Healthy Planet, Epic Systems) provide current snapshots of patients and are helpful for population health management. These operational reporting tools are fast, provide real-time data, and are incorporated into the clinical workflow. These minute-by-minute updates of clinical data are unnecessary for many types of secondary data analyses. Population health registries have motivations that may differ from research investigations. For example, population health registries support quality-based metrics such as indicators maintained by the National Quality Foundation, which may be publicly reported and are used to guide reimbursement incentives for programs such as the Medicare Shared Savings Plan. In these instances, disease phenotypes and variable definitions are pre-defined by the interested parties. In this scenario, there may be a single criterion used to define the population and associated metrics. Creating additional criteria would be counterproductive. By contrast, a research registry should provide comprehensive data on members collected over time, requires statistical analyses, and may contain multiple definitions for the same variable. These data allow evaluations at user-defined time points or time-varying analyses. Because the tool is not integrated into clinical workflows, there is an opportunity to incorporate large quantities of data into computationally intensive analyses that would otherwise be a drain on clinical systems.

Population health registries are ideally suited for clinical care and quality improvement in that they are available instantaneously on the live EHR, have standardized definitions, and use limited computing resources. By contrast, the type of research registry that we have created enables the creation of different cohorts for the same disease entities, makes use of additional computation resources that would be inappropriate for the clinical EHR and allows different variable definitions depending on the specific study. Table 1 lists additional differences between our research registry and population health registries.

Registries created from EHR data may have different goals and requirements. The table compares features of research and population health registries.

It should also be noted that EHR vendors each use their own proprietary *technical* data models that will map to ontologies such as International Classification of Diseases codes. The precise mappings are not made publicly available, which makes multicenter studies involving different EHR systems more difficult. The registry we have built is mapped to the Observational Medical Outcomes Partnership (OMOP) common data model (CDM). CDMs such as OMOP have been instrumental in creating interoperability standards in support of clinical research networks that span multiple institutions. This registry will take advantage of the data mappings available in OMOP and benefit from the automated tools developed for OMOP for identifying potential data issues. The Phenotype Knowledge Base contains a repository of electronic phenotypes to support registry construction and variable definitions [5]. These phenotypes have been successfully integrated into the OMOP data model to facilitate implementation at different research institutions [6]. We will also have the opportunity to create additional custom mappings to our OMOP instance, which can be leveraged by local researchers.

Table 1. Characteristics of research registries vs population health registries.

Research registry	Population health registry
Intermittent updates	Real-time updates
Higher computational resources	Low resource use
Complex definitions from a variety of sources and multiple definitions for similar concepts	Simple definitions defined by QI-based[a] reimbursement
Variety of external data sources	Data limited to EHR[b]
Extensive data processing	Limited data processing
Complex temporal relationships	Single point in time
Easily accessible and detailed documentation	Documentation or coding sometimes lacking or not easily accessible
Does not need to be integrated into workflow	Integration in clinic workflow is crucial
Does not require front-end EHR access.	Requires front-end EHR access with PHI[c]
Mapped to open-source common data models	Mapped to vendor-based *technical* data models

[a] QI: quality improvement.

[b] EHR: electronic health record.

[c] PHI: protected health information.

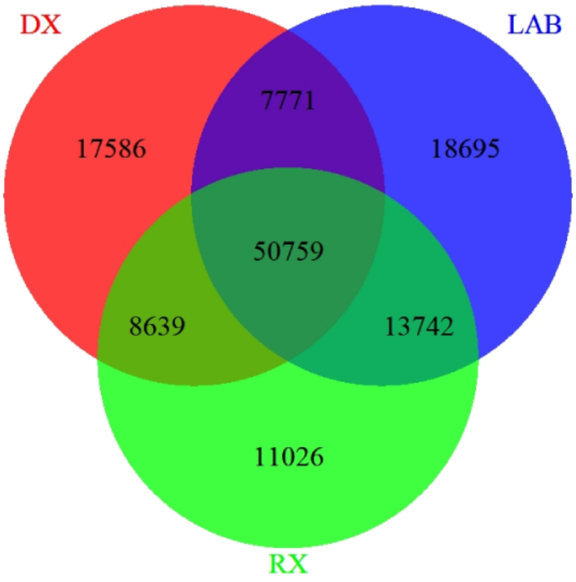
Custom Phenotypes

As mentioned previously, research projects may require variable definitions that are different from quality-based metrics, and variable definitions may vary from one project to the next. Varying variable definitions are also necessary for cohort discovery. The definition of diabetes may differ between projects. For example, a case control study needing a limited number of cases may want to have a highly specific definition for type 2 diabetes such as the one created by Kho [7]. By contrast, a study evaluating the accuracy of different electronic phenotypes may require a highly sensitive definition to capture all possible diabetes cases for manual chart review [8]. Figure 1 shows a Venn diagram illustrating the different patient populations that would be captured from our data warehouse depending on whether one uses diagnosis codes, hemoglobin A_{1c} laboratory values, or prescriptions for hypoglycemic medications.

In other instances, existing definitions may be available from agencies such as Agency for Healthcare Research and Quality or the CMS. For example, we used the CMS definition for an acute exacerbation of chronic obstructive pulmonary disease for a study looking at the impact of a chronic obstructive pulmonary disease care pathway on reducing readmissions [9].

In addition to phenotypes used for cohort discovery, research projects require definitions for covariates included in the statistical analyses. Depending on the situation, investigators may desire different definitions for comorbidities such as hypertension. Textbox 2 shows the contrast between an example of a simple definition for hypertension based on diagnoses codes vs a complex definition that might be used for a study, where maximizing the sensitivity for identifying hypertension is key.

Figure 1. Sets of patients with possible diabetes according to definitions based on diagnoses codes (DX), laboratory values (LAB), or prescriptions (RX).



Textbox 2. Example definitions of hypertension.

Research registry

- International Classification of Diseases (ICD) code for hypertension (HTN) in encounter diagnoses, past medical history, or problem list *OR*
- Minimum of 3 blood pressure (BP) readings >140/90 over 3 months in the electronic health records
  - Outpatient BP excluding urgent care clinics, emergency department, or observation visits
  - Based on last BP of encounter
  - Exclude BPs when associated temperature≥38 °C *OR*
- Active prescription for an antihypertensive agent

Population health

- ICD code for HTN in encounter diagnoses

OMOP Limitations

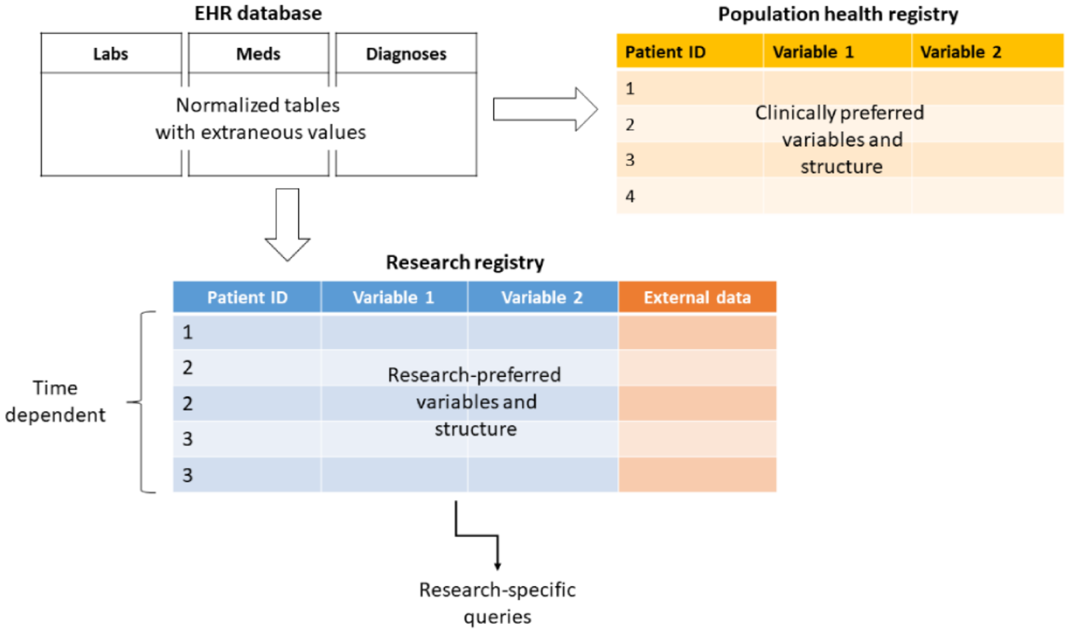
While the use of OMOP has many advantages in terms of standardization, there are still significant areas of limitations. Medications are one area where common data models are still lacking. For example, OMOP contains a single drug exposure table for prescriptions, drug administration, dispensing information, and patient-reported information. Unfortunately, dispensing information, patient-reported information, and compliance are rarely captured in structured EHR data. In addition, there are no explicitly linked medical reasons for the exposures in OMOP, and the RxNorm categorizations may not be appropriate for a specific research study. A registry cannot resolve all these issues, but the structure provides the flexibility to create and validate new phenotypes. For example, researchers can create and share relevant medication groupings, and algorithms based on specific prescription information (eg, dates of prescriptions, stop dates, number of pills, and number of refills) can be created as proxies for active medications and compliance. Similarly, associated information (eg, presence or absence of different diagnoses codes and laboratory values) can define reason for medication. These new phenotypes can be used locally and shared with the OMOP community without being formally integrated into the OMOP model.

OMOP will not be able to represent all the new phenotypes that the registry will require, making it necessary to characterize our own concepts. Some of these concepts may be derived entirely from existing OMOP concepts, but many will require the creation of our own. Like all CDMs, OMOP has limitations in its capacity to represent information inherent to the transformation from one data model (eg, EHR) to another. In addition, it will be crucial to have a formal data quality structure in place to ensure mappings are correct and routinely updated as data change. We have established a phenotype working group that includes the authors as well as additional faculty members in the Center for Biomedical Informatics.

Data Structure

The data structure of the EHR database, a typical population health registry, and a research registry can vary significantly. EHR databases are stored in database management systems using individual, partially normalized tables for each specific data domain. This structure reduces storage space and speeds data extractions. By contrast, most statistical analyses require an individual flat data table (also known as *pivot table*), where the unit of analyses is the individual rows of patients. The data sets need to include columns for both independent and dependent variables and may require calculations of follow-up time between baseline variables and the outcomes of interest. Some external variables that do not exist in the EHR may be linked with the data set. For example, we link our registry with the North Carolina state death index, allowing better ascertainment of mortality outcomes and censoring of follow-up time. Variables may also be derived from the source data (eg, highest blood pressure in the past 24 hours) and time dependent analyses necessitate multiple rows for each patient that reflect the patient's current state at a given point in time. Figure 2 provides a graphical representation of the different data structures between the EHR database, an EHR-based population health registry, and a research registry.

Figure 2. Comparing data structures of electronic health records (EHR), population health registries, and research registries.



Diabetes-Specific Registry

We chose diabetes as one of the first registries to make available in our Clinical and Translational Science Award Program given that it represents a focus area of our research enterprise. In addition, diabetes is a natural choice for a research registry given the rising incidence, chronic nature, established quality metrics, comorbidities, availability of treatments, and research funding. Research also indicates that blood sugar and associated risk factors are poorly controlled in patients with diabetes. In addition to a desire for improving the health of their patients, health care institutions have direct financial incentives for adequately treating patients with diabetes. Quality indicators approved by a successful diabetes research registry would provide an opportunity for the creation of risk prediction models that could be used to target patients at high risk as well as those who are most likely to benefit from a specific intervention. Thorough statistical evaluations of quality improvement projects and population health interventions would provide crucial feedback on the potential net benefits of these programs.

The identification of diabetes in EHR is surprisingly complex. Common methods for identifying potential cases include searches for medications, laboratory values, and diagnosis codes. Each of these approaches has its own limitations. Medications used for diabetes may also be used to treat other conditions. For example, metformin is commonly prescribed for polycystic ovarian syndrome in women. Blood glucose values may be abnormally elevated due to inadequate fasting times, which are generally not easily determined in the EHR. Diagnosis codes may be incorrectly used before patients meet formal criteria for diabetes or may be associated with the incorrect diabetes type. The issues in correctly identifying patients with diabetes highlight the importance of flexible research registries. Recognizing the potential need for different diabetes definitions, we chose to create our registry based on the concept of a highly sensitive *Wide Net* with the goal of capturing any evidence of possible diabetes in the EHR. Figure 3 provides a graphical display of this concept.

This approach mirrors the one used by the SEARCH for Diabetes in Youth evaluation of using EHRs for diabetes surveillance [8]. Approaches such as these are necessary given the infeasibility of manually reviewing all patient charts. The SEARCH work found that the simple use of diabetes codes could accurately determine EHR evidence of diabetes, and the ratio of type 1 to type 2 codes had a high sensitivity and specificity for identifying youth with type 1 diabetes. Additional work is needed to determine the accuracy of this approach in adults, and further algorithms are needed for identifying children with type 2 diabetes or other diabetes types. This registry provides a great source of data for future electronic phenotypic development and validation.

Our registry contains 128,218 patients with possible diabetes according to one or more of these 3 domains, while only 50,759 patients have evidence of possible diabetes based on all 3 variables simultaneously (Table 2). Identifying random subsets of patients who meet different combinations of these criteria provides an opportunity to glean valuable information from manual chart reviews of these patients. Annotated data sets allow for evaluation of existing and creation of new electronic phenotypes for diabetes status, type, and date of diagnoses.

Figure 3. Venn diagram showing the use of electronic algorithms combined with chart reviews to identify patients with diabetes. DM: Diabetes Mellitus; EHR: electronic health record; HbA$_{1c}$: hemoglobin A$_{1c}$; ICD: International Classification of Diseases.

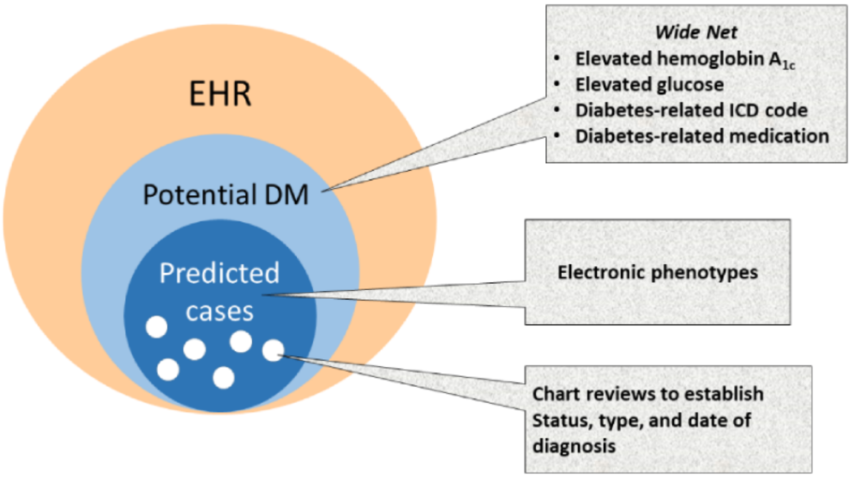

EHR
Potential DM
Predicted cases
Wide Net
• Elevated hemoglobin A1c
• Elevated glucose
• Diabetes-related ICD code
• Diabetes-related medication
Electronic phenotypes
Chart reviews to establish Status, type, and date of diagnosis

Table 2. Characteristics of patients[a] who showed evidence of possible diabetes based on diagnoses codes, laboratory values, or medications.

Characteristics	Cohort 1: diagnosis	Cohort 2: labs[b]	Cohort 3: medications
Total unique patients, n (%)	84,755 (66)	90,967 (71)	84,165 (66)
Age (years), median (IQR)	66.02 (19.43)	65.46 (20.20)	64.62 (20.98)
Sex, n (%)			
Female	43,510 (51.34)	44,008 (48.38)	43,374(51.53)
Male	41,239 (48.66)	46,950 (51.61)	40,783 (48.46)
Race, n (%)			
White	59,547 (70.26)	65,693 (72.22)	60,014 (71.30)
Black	19,120 (22.56)	19,042 (20.93)	17,905 (21.27)
Other	5794 (6.84)	5938 (6.53)	6004 (7.13)
Missing	286 (0.34)	267 (0.29)	223 (0.26)
Ever smoker, n (%)	43,414 (51.22)	48,842 (53.69)	44,133 (52.44)
Insulin (1 or more prescriptions in the past year), n (%)	25,663 (30.28)	25,943 (28.52)	26,685 (31.70)
Charlson comorbidity index, n (median)	83,699 (2)	89,692 (2)	83,094 (2)
Median household income, n (median)	66,034 (46,283)	69,253 (45,688)	64,839 (45,927)
Most recent hemoglobin A$_{1c}$, n (median)	64,959 (6.9)	72,833 (7.1)	69,933 (7.0)
Most recent eGFR[c], n (median)	73,037 (70)	88,633 (66)	80,424 (70)
Most recent LDL[d], n (median)	58,463 (88)	60,398 (88)	59,864 (89)

[a] Patients may exist in 1, 2, or all 3 of the cohorts.

[b] Random blood sugar ≥200 mg/dL or hemoglobin A$_{1c}$≥6.5%.

[c] eGFR: estimated glomerular filtration rate calculated using the Chronic Kidney Disease Epidemiology Collaboration (CKD-Epi) equation.

[d] LDL: low-density lipoprotein.

Jupyter Notebooks

Much like the interinstitutional heuristic and algorithm sharing enabled by sites supporting an OMOP CDM, there is potential for intrainstitutional collaboration and technique leveraging. Views in OMOP can be created by the honest brokers to provision only the cohort and relevant data permitted by an institutional review board application to specific authorized study personnel.

Jupyter is a free, open-source, interactive web-based computational notebook widely adopted by data scientists across thousands of enterprises, including Fortune 500 companies, international research facilities, universities, and start-ups. A Jupyter hub server allows users to centrally create and share codes, equations, visualizations, as well as text and results. It will also allow researchers to interact directly with their data views in OMOP via a programmatic language of their choice, whether it be Python (Python Software Foundation), R (The R Foundation), or even direct SQL. A library of Jupyter Notebooks with example code and outputs provided by data analysts can give researchers a rich starting base of programmatic techniques that they can modify, improve, and share back for other researchers to use in their own Jupyter Notebook analyses, greatly reducing the learning curve and lessening code redundancy and reimplementation.

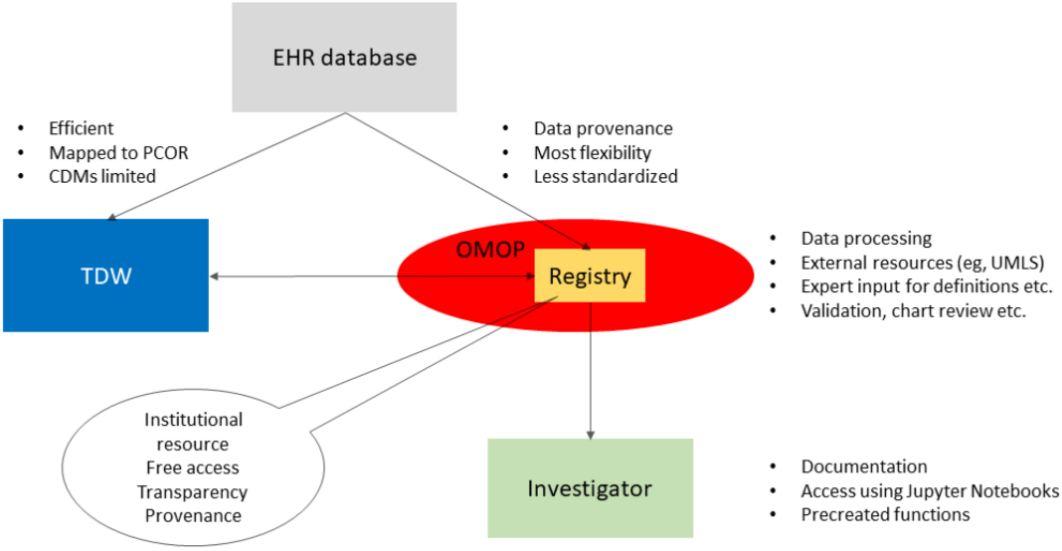
Schematic

Figure 4 shows a schematic of the overall architecture of the registry and highlights some of the guiding principles governing the registry creation.

Data processing will undoubtedly uncover errors in the clinical data (eg, implausible values), which will be cleaned for data analyses. Data cleaning will be performed at the registry or post–data extract level. We are not attempting (at least at this point) to try and change values in the source clinical data, which is a difficult process and could have clinical implications. It is our hope that the registry could be used for data quality projects that might recognize a way to improve data collection or documentation.

As mentioned previously, the registry is mapped to the OMOP CDM and linked with our existing translational data warehouse. This ensures the standardization of data within the registry while exploiting our established infrastructure. Infusion of additional data from the vendor EHR database as well as data external to our Clinical Information Systems and our institution provides flexibility and continued creation of additional phenotypes. We have created a digital phenotype working group that will prioritize electronic phenotype creation and ensure appropriate documentation. Access to the registry through Jupyter Notebooks increases transparency and simplifies the sharing of code between investigators.

Figure 4. Schematic of the overall architecture of the registry, highlighting some of the guiding principles governing the registry creation. CDM: common data model; EHR: electronic health records; OMOP: Observational Medical Outcomes Partnership Common Data Model; PCOR: patient-centered outcomes research common data model; TDW: Translational Data Warehouse in the Wake Forest Clinical and Translational Science Institute; UMLS: Unified Medical Language System.

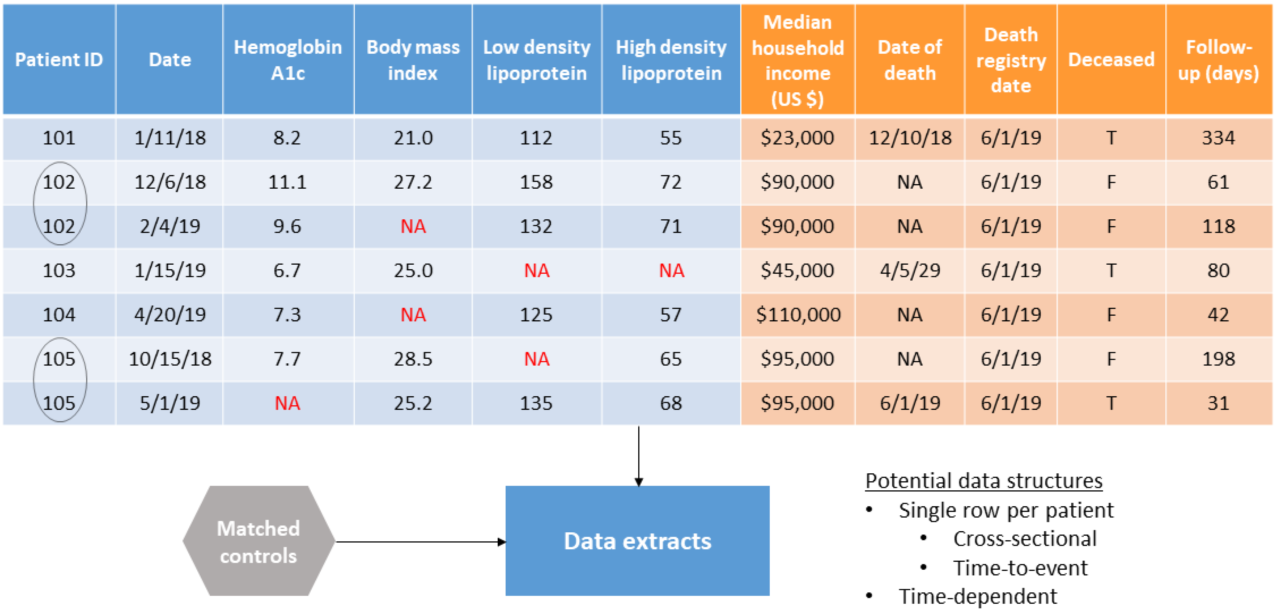


Data Extracts

Using existing R code created at Wake Forest will allow investigators to extract individual analytic tables that define patient characteristics at each given point in time per the specific study design. Figure 5 highlights how this table would appear.

Additionally, a Wake Forest Center for Biomedical Informatics–sponsored pilot grant is establishing a tool for creating randomly selected control patients to simplify the conduct of case control studies. We also have existing R code for the imputation of missing values using multiple imputation with chained equations that can be applied after the analytic data set has been created. Creation of multiply imputed data sets allows an estimation of the amount of missing information and stability of coefficient estimates [4].

Figure 5. Example analytic data set extracted from the registry in a pivot table format. F: false; NA: not applicable; T: true.



Patient ID	Date	Hemoglobin A1c	Body mass index	Low density lipoprotein	High density lipoprotein	Median household income (US $)	Date of death	Death registry date	Deceased	Follow-up (days)
101	1/11/18	8.2	21.0	112	55	$23,000	12/10/18	6/1/19	T	334
102	12/6/18	11.1	27.2	158	72	$90,000	NA	6/1/19	F	61
102	2/4/19	9.6	NA	132	71	$90,000	NA	6/1/19	F	118
103	1/15/19	6.7	25.0	NA	NA	$45,000	4/5/29	6/1/19	T	80
104	4/20/19	7.3	NA	125	57	$110,000	NA	6/1/19	F	42
105	10/15/18	7.7	28.5	NA	65	$95,000	NA	6/1/19	F	198
105	5/1/19	NA	25.2	135	68	$95,000	6/1/19	6/1/19	T	31

Future

Although the registry will be based on coded information, we recognize the growth in the data science community of graph representation of data. The ability to use Jupyter Notebooks to access data and to create and share code will allow investigators to integrate new methods such as graph theory for statistical analyses and to create data visualizations to share. We are particularly interested in examining diabetes-related treatment pathways and intend to use the concept relationship table in OMOP to define treatment pathways commonly used as well as pathways based on guidelines. The characterization of treatment pathways is ripe for graph representation.

We recognize that the data, informatics tools, and analytic techniques available for EHR-based analyses are rapidly changing. We have identified a group of clinical, informatics, and statistical professionals who can serve as registry stakeholders. Periodic meetings will allow for continuous feedback that will guide decisions on registry directions and priorities. The Wake Forest Clinical and Translational Science Institute has an established mechanism for continuous evaluation of the informatics program, of which this registry will be a part. Evaluations will include metrics on registry use, publications and grants using the registry, as well as formal (eg, surveys) and informal feedback.

Summary

Secondary use of EHR data for research is still in its infancy, and tools to aid investigators in complex epidemiological-type studies needed for the Learning Health System are lacking. Typical population health registries do not provide the flexibility, computational resources, and data complexity necessary for many research endeavors. The virtual diabetes registry described in this paper is providing our researchers with tools that we hope will enable them to conduct sophisticated statistical analyses in the most transparent and efficient way possible. The registry is being built in a way that will allow for its continuous refinement based on user experience and in a format that will enable interinstitutional collaboration.

Acknowledgments

We would like to acknowledge the Informatics Program of the Wake Forest Clinical and Translational Science Institute (WF CTSI), which is supported by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through Grant Award Number UL1TR001420.

Conflicts of Interest

None declared.

References

1. Fetter RB, Shin Y, Freeman JL, Averill RF, Thompson JD. Case mix definition by diagnosis-related groups. Med Care 1980 Feb;18(2 Suppl):iii, 1-iii,53. [Medline: 7188781]

2. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, Expert Panel. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. J Am Med Inform Assoc 2007;14(1):1-9 [FREE Full text] [doi: 10.1197/jamia.M2273] [Medline: 17077452]
3. Kruse CS, Stein A, Thomas H, Kaur H. The use of Electronic Health Records to Support Population Health: A Systematic Review of the Literature. J Med Syst 2018 Sep 29;42(11):214 [FREE Full text] [doi: 10.1007/s10916-018-1075-6] [Medline: 30269237]
4. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. EGEMS (Wash DC) 2013;1(3):1035 [FREE Full text] [doi: 10.13063/2327-9214.1035] [Medline: 25848578]
5. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc 2016 Nov;23(6):1046-1052 [FREE Full text] [doi: 10.1093/jamia/ocv202] [Medline: 27026615]
6. Hripcsak G, Shang N, Peissig PL, Rasmussen LV, Liu C, Benoit B, et al. Facilitating phenotype transfer using a common data model. J Biomed Inform 2019 Aug;96:103253 [FREE Full text] [doi: 10.1016/j.jbi.2019.103253] [Medline: 31325501]
7. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc 2012;19(2):212-218 [FREE Full text] [doi: 10.1136/amiajnl-2011-000439] [Medline: 22101970]
8. Wells BJ, Lenoir KM, Wagenknecht LE, Mayer-Davis EJ, Lawrence JM, Dabelea D, et al. Detection of Diabetes Status and Type in Youth Using Electronic Health Records: The SEARCH for Diabetes in Youth Study. Diabetes Care 2020 Oct;43(10):2418-2425 [FREE Full text] [doi: 10.2337/dc20-0063] [Medline: 32737140]
9. Ohar JA, Loh CH, Lenoir KM, Wells BJ, Peters SP. A comprehensive care plan that reduces readmissions after acute exacerbations of COPD. Respir Med 2018 Aug;141:20-25 [FREE Full text] [doi: 10.1016/j.rmed.2018.06.014] [Medline: 30053968]

Abbreviations

CDM: common data model
DRG: diagnosis-related group
EHR: electronic health record
OMOP: Observational Medical Outcomes Partnership

Edited by C Lovis; submitted 20.05.22; peer-reviewed by L Patel; comments to author 26.06.22; revised version received 24.08.22; accepted 28.08.22; published 23.09.22.

Please cite as:
Wells BJ, Downs SM, Ostasiewski B
Using Electronic Health Records for the Learning Health System: Creation of a Diabetes Research Registry
JMIR Med Inform 2022;10(9):e39746
URL: https://medinform.jmir.org/2022/9/e39746
doi:10.2196/39746
PMID:36149742

Original Paper

A Free App for Diagnosing Burnout (BurnOut App): Development Study

Jordi Godia[1], MSc; Marc Pifarré[1], MSc; Jordi Vilaplana[1], PhD; Francesc Solsona[1], PhD; Francesc Abella[2], PhD; Antoni Calvo[3], MSc; Anna Mitjans[3], MBA; Maria Pau Gonzalez-Olmedo[3†], PhD

[1]Universitat de Lleida, Lleida, Spain
[2]IRBLleida, Lleida, Spain
[3]Fundacio Galatea, Barcelona, Spain
[†]deceased

Corresponding Author:
Francesc Solsona, PhD
Universitat de Lleida
Jaume II, 69
Lleida, 25001
Spain
Phone: 34 973702735
Email: francesc.solsona@udl.cat

Abstract

Background: Health specialists take care of us, but who takes care of them? These professionals are the most vulnerable to the increasingly common syndrome known as burnout. Burnout is a syndrome conceptualized as a result of chronic workplace stress that has not been successfully managed.

Objective: This study aims to develop a useful app providing burnout self-diagnosis and tracking of burnout through a simple, intuitive, and user-friendly interface.

Methods: We present the BurnOut app, an Android app developed using the Xamarin and MVVMCross platforms, which allows users to detect critical cases of psychological discomfort by implementing the Goldberg and Copenhagen Burnout Inventory tests.

Results: The BurnOut app is robust, user-friendly, and efficient. The good performance of the app was demonstrated by comparing its features with those of similar apps in the literature.

Conclusions: The BurnOut app is very useful for health specialists or users, in general, to detect burnout early and track its evolution.

(JMIR Med Inform 2022;10(9):e30094) doi:10.2196/30094

KEYWORDS

diagnose burnout; Android app; medical informatics; health care; health professionals; mobile health; digital health; health applications; online health; mobile phone

Introduction

In this study, we deal with burnout syndrome. Burnout syndrome is becoming increasingly popular. It is not a disease but a signal of emotional distress. Significant efforts have been made to determine its causes.

In 1974, Freudenberger and Richelson [1] suggested that feelings of exhaustion, frustration, and tiredness are generated by an overload. He included the term work addiction in the explanation, also being the first to propose that this type of relationship is associated with a productive imbalance. Freudenberger and Richelson [1] later expanded this theory by saying that these feelings were because of the irrational workloads imposed by the workers themselves or the people around them.

Maslach and Jackson [2] defined burnout as a syndrome with three dimensions (the most widely accepted to date):

1. Emotional exhaustion: emotional exhaustion because of the demands of work
2. Depersonalization: indifference and apathy toward society

3. Low personal fulfillment: low feelings of success and personal fulfillment

Burnout is included in the 11th Revision of the International Classification of Diseases [3] as an occupational phenomenon. It is not classified as a medical condition and is defined in the 11th Revision of the International Classification of Diseases as follows: burnout is a syndrome conceptualized as resulting from chronic workplace stress that has not been successfully managed. Burnout refers specifically to phenomena in the occupational context and should not be applied to describe experiences in other areas of life.

A partial list of potential contributing causes includes (1) length of training, (2) mentality of delayed gratification, (3) insufficient protected research time and funding, (4) long working hours, (5) imbalance between career and family, (6) hostile workplace environment, and (7) gender- and age-related issues [4]. Burnout can have a significant negative impact on the quality of patient care by negatively influencing clinical decision-making, increasing medical errors and malpractice claims, and lowering patient satisfaction [5-7]. Burnout may also lead to high turnover, difficult relationships between providers and staff, and drug and alcohol abuse [8].

In general, those most vulnerable to distress from the syndrome are professionals in whom worker-client human interactions of an intense or lasting nature are observed [9,10]. Balch and Shanafelt [11] found that health care professionals are at a disproportionately higher risk than other workers in stressful jobs that focus on public services. Burnout is markedly more common among physicians than depression, substance abuse, or suicide [11]. Shanafelt et al [12] reported that 45% of physicians had experienced at least one symptom of burnout. Another study found that high rates of depersonalization were the greatest among early-career physicians and decreased with age [13]. Burnout may affect >60% of family practice providers [14].

A recent study [15] found that physician turnover and reduced clinical hours attributable to burnout resulted in approximately US $4.6 billion in costs each year in the United States. The rising prevalence of burnout among physicians and other health care professionals has become a major policy concern in the United States during the COVID-19 pandemic [16]. Regardless of burnout status, the results showed that all professional health care groups had high levels of anxiety. Primary care physicians had significantly higher anxiety scores than all other health care professionals. Thus, a sense of tension, anxiety, distress, and other symptoms of mental disorders [16,17] would greatly help detect burnout syndrome.

Preventive actions for burnout include checklists, tools for early detection, training programs for high-risk occupations, awareness-raising actions, and good practice guidelines [18]. Free and user-friendly apps could be good tools for tackling all the actions on the list, with guarantees of success.

In most of the research conducted on mobile apps available to detect burnout, such as MindDoc [19], Psychosomat [20], and BreathePro [21], no free tax and public algorithms were used, although they are very popular. By contrast, Lafraxo et al [22] created an app to detect burnout in a nursery where they used the Copenhagen Burnout Inventory (CBI) [23] algorithm.

The research question for this study is whether a free cloud-based mobile app provides potential patients with burnout syndrome with a diagnosis based on their needs and goals. This study presents a mobile app called BurnOut to offer potential patients with burnout syndrome a diagnosis generator based on their needs and goals using cloud-based mobile apps to help diagnose burnout using the CBI [23] and Goldberg Health Questionnaire (GHQ) [17], which are free and public algorithms. The use of these free and contrasted algorithms benefits the study as the final results can be followed up and reproduced for comparison.

In this study, we present the BurnOut app. Our proposal was focused on offering a self-operating tool to diagnose burnout. This also allows users to monitor their evolution. The most common instrument for diagnosing burnout is the licensed Maslach Burnout Inventory test, developed by Maslach and Jackson in 1981 [24]. The BurnOut app implements the CBI [23], a valid, free, and reliable alternative. It also implements a version of the GHQ with 12 questions used to detect mental disorders [17].

Methods

BurnOut App

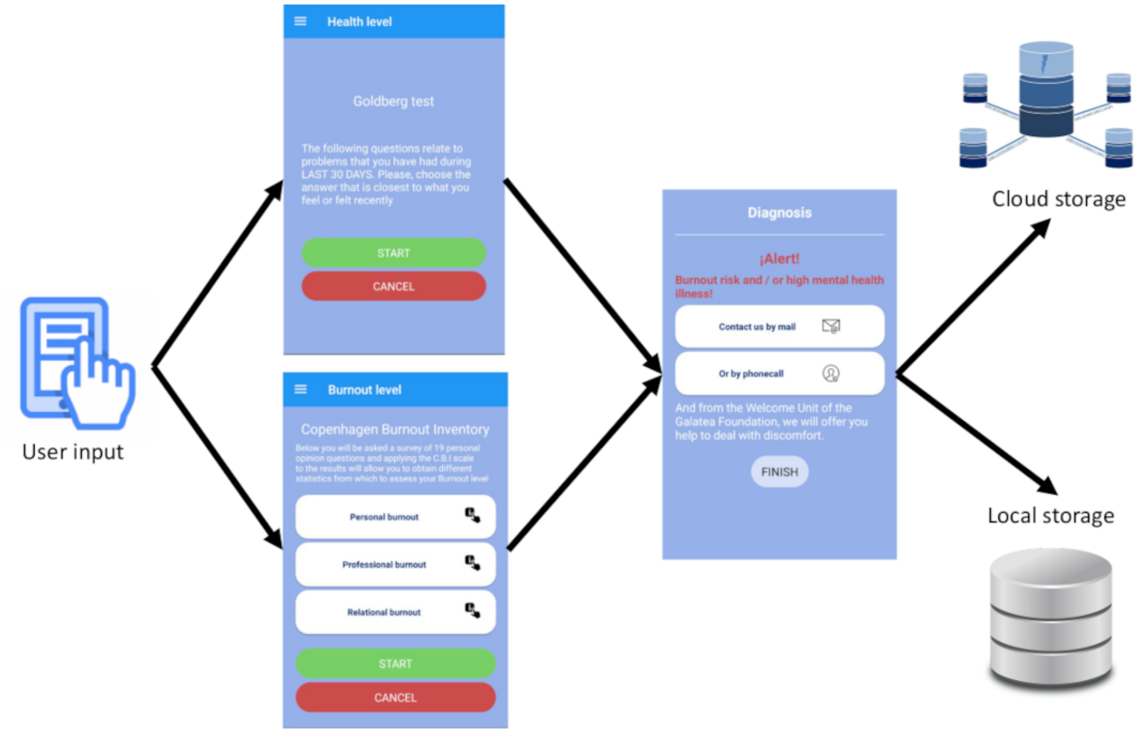
Figure 1 shows the basic operation of the BurnOut app. It represents the flowchart between the user interface (UI) of the CBI and GHQ tests of the BurnOut app, which generates the final custom diagnosis. The diagnosis is shown on the mobile display and saved jointly with the tests in the local database and cloud storage.

The BurnOut app provides 3 main features: data collection, burnout diagnosis, and user monitoring. The UI for such features is user-friendly. The app also provides language support in Spanish, Catalan, and English. The default language is the same as that used on the smartphone.

The app provides a mail contact for doubts or suggestions with the developer through an *About* section, where the user can check some informative data from the app as a legal copyright.

The Android version of the app can be downloaded from Google Play using the app link [25].

Figure 1. BurnOut app operation.



The following questions relate to problems that you have had during LAST 30 DAYS. Please, choose the answer that is closest to what you feel or felt recently

Below you will be asked a survey of 19 personal opinion questions and applying the C.B.I scale to the results will allow you to obtain different statistics from which to assess your Burnout level

Burnout risk and / or high mental health illness!

And from the Welcome Unit of the Galatea Foundation, we will offer you help to deal with discomfort.

User Testing

A sample of 40 random users (summarized in Table 1) representing different social levels, professional activities, ethnical profiles, ages, and genders was chosen from the cities of Lleida and Tarragona (Spain).

Table 1. Test participant features (N=40).

Feature and class	Participants, n (%)
Gender	
Male	17 (43)
Female	23 (58)
Age (years)	
<30	30 (75)
35-65	10 (25)
Technological profile	
Yes	13 (33)
No	27 (68)

Collecting User Data

The first time a user signs up on the app, a questionnaire with the following information must be filled in:

1. Eating habits: a few standard questions about healthy eating habits such as the frequency of meals or the number of fruits and vegetables consumed per day
2. Physical activity: how often does one walk and whether they engage in moderate or intensive physical activity for at least 10 minutes; if so, the length and frequency are requested
3. Consumption of toxic substances: yes or no questions about consumption of drugs; if yes, consumption for the past 30 days is requested

These data provide information about the user's lifestyle and factors that can cause psychosocial risk and affect their psychoemotional wellness.

Tests

Overview

BurnOut uses 2 instrumentally proven tests to evaluate users' burnout (with the CBI) and psychological discomfort (with the GHQ). The tests do not constitute a final diagnosis but help give the professional an idea about their mental health and the perception of the risk of exhaustion, which can serve as a guide for a more precise diagnosis.

The main features of the GHQ and CBI are as follows:

1. GHQ: This test has several versions; the implemented version comprises 12 questions, used by the Medical College of Barcelona. It evaluates the presence or lack of psychological discomfort (ie, distress). GHQ has a binary score (true or false). Further details are provided in the GHQ Test section.
2. CBI: It comprises 19 multiple-choice questions to detect the symptoms of burnout. It has no final score and evaluates 3 subscales: personal, work-related, and client-related burnout.

GHQ Test

The GHQ test evaluates the existence of psychological discomfort. It measures the sense of tension, depression, inability to defend oneself, anxiety-based insomnia, lack of self-confidence and self-esteem, and other symptoms of mental disorders [16,17,26].

There are 4 variants of this questionnaire, and the GHQ-12 variant used in this study is recommended for measuring psychological distress. This test contains 12 questions (Textbox 1) about the emotional or psychological problems that the user experienced in the past 30 days. Each answer comprises 4 options that are equivalent to numerical values (Table 2). The bimodal scoring method was used in the BurnOut app by the official manual. The maximum score (number of points on the test) is 12, and the possible range is 0 to 12 [27]. A score ≥4 indicates the possible presence of mental distress, and a score ≥8 indicates the presence of various symptoms of stress-related psychological disorders. To ensure the diagnosis and avoid ignoring any symptoms, a score equal to or higher than the threshold value of 3 was classified as distress in the BurnOut app.

Textbox 1. Goldberg Health Questionnaire questions.

Questions

1. Have you been able to concentrate well on what you did?
2. Have your worries made you lose a lot of sleep?
3. Have you felt that you play a useless role in life?
4. Have you felt capable of making decisions?
5. Have you felt under strain?
6. Have you ever felt that you cannot overcome your difficulties?
7. Have you been able to enjoy your activities every day?
8. Have you been able to deal adequately with your problems?
9. Have you felt unhappy and depressed?
10. Have you lost confidence in yourself?
11. Have you thought that you are useless?
12. Do you feel reasonably happy, considering the circumstances?

Table 2. Goldberg Health Questionnaire answer values.

Question number and answer options	Value
Question 1	
Better than usual	0
As usual	0
Less than usual	1
Much less than usual	1
Question 2	
Not at all	0
No more than usual	0
A little more than usual	1
Much more than usual	1
Question 3	
More useful than usual	0
As usual	0
Less than usual	1
Much less than usual	1
Question 4	
More than usual	0
As usual	0
Less than usual	1
Much less than usual	1
Question 5	
Not at all	0
No more than usual	0
A little more than usual	1
Much more than usual	1
Question 6	
Not at all	0
No more than usual	0
A little more than usual	1
Much more than usual	1
Question 7	
More than usual	0
As usual	0
Less than usual	1
Much less than usual	1
Question 8	
More capable than usual	0
As usual	0
Less capable than usual	1
Much less capable than usual	1
Question 9	
Not at all	0

Question number and answer options	Value
No more than usual	0
A little more than usual	1
Much more than usual	1
Question 10	
Not at all	0
No more than usual	0
A little more than usual	1
Much more than usual	1
Question 11	
Not at all	0
No more than usual	0
A little more than usual	1
Much more than usual	1
Question 12	
More than usual	0
Approximately the same as usual	0
Less than usual	1
Much less than usual	1

CBI Test

The CBI is one of the most widely used burnout inventories [23,28]. The CBI questionnaire explores the following three dimensions of burnout:

1. Personal burnout: degree of fatigue or emotional exhaustion experienced by a person
2. Work-related burnout: Burnout related to one's job, experienced in relation to the work without trying to establish causal relationships
3. Client-related burnout: the degree of emotional fatigue or exhaustion that someone experiences in relation to their work with other people

The psychometric qualities of this test make it a good tool for diagnosis and prevention. The CBI questionnaire is intended only to allow users to make a conservative self-assessment of their burnout status. This result is not intended as a medical diagnosis. However, it can inform them as to whether they should seek medical and psychotherapeutic assistance. Only a trained physician is qualified to advise on the initiation, modification, or discontinuation of the medication.

The CBI comprises a survey of 19 questions that evaluate 3 dimensions that affect burnout symptoms (personal burnout [Textbox 2], work-related burnout [Textbox 3], and client-related burnout [Textbox 4]).

The questions in the CBI application did not appear in the same order, as shown here. The questions were mixed with those on other topics. This is recommended to avoid stereotypical response patterns. These dimensions comprise 3 independent subscales to determine the risk of burnout according to the combination of the scores.

These questions can have 2 different packs of 5 answers with a numerical value associated with them (Table 3). These packs are indistinctly used. There is only one exception in the 13th question of the work-related dimension—"Do you have enough energy for family and friends during leisure time?"— for which the score is reversed. The total score in the dimension is the average of the scores obtained in this dimension. Depending on this value, each dimension is categorized into one of the three burnout ranked levels: low, moderate, and high (Table 4).

The CBI final diagnosis depends on the scores obtained for each dimension (Table 5).

Textbox 2. Copenhagen Burnout Inventory questions: personal burnout dimension.

Personal burnout questions

1. How often do you feel tired?
2. How often are you physically exhausted?
3. How often are you emotionally exhausted?
4. How often do you think: “I can’t take it anymore”?
5. How often do you feel worn out?
6. How often do you feel weak and susceptible to illness?

Textbox 3. Copenhagen Burnout Inventory questions: work-related burnout dimension.

Work-related burnout questions

1. Is your work emotionally exhausting?
2. Do you feel burned out because of your work?
3. Does your work frustrate you?
4. Do you feel worn out at the end of the working day?
5. Are you exhausted in the morning at the thought of another day at work?
6. Do you feel that every working hour is tiring for you?
7. Do you have enough energy for family and friends during leisure time?

Textbox 4. Copenhagen Burnout Inventory questions: client-related burnout dimension.

Client-related burnout questions

1. Do you find it hard to work with clients?
2. Do you find it frustrating to work with clients?
3. Does it drain your energy to work with clients?
4. Do you give more than you get back when you work with clients?
5. Are you tired of working with clients?
6. Do you wonder how long you will continue working with clients?

Table 3. Copenhagen Burnout Inventory answer values.

Answer pack 1	Answer pack 2	Value
Always	To a very high degree	100
Often	To a high degree	75
Sometimes	Somewhat	50
Seldom	To a low degree	25
Never or almost never	To a very low degree	0

Table 4. Copenhagen Burnout Inventory dimension valuations.

Dimension	Level		
	Low	Moderate	High
Personal burnout	<50	50-74	75-100
Work-related burnout	<50	50-74	75-100
Client-related burnout	<50	50-74	75-100

Table 5. Copenhagen Burnout Inventory risk.

Copenhagen Burnout Inventory score	Cases
Low	Low-risk in all 3 dimensions; 2 low-risk dimensions and 1 moderate-risk dimension
Moderate	3 moderate-risk dimensions; 2 moderate-risk dimensions, 1 low-risk dimension, and 1 high-risk dimension
High	2 high-risk dimensions and 3 high-risk dimensions

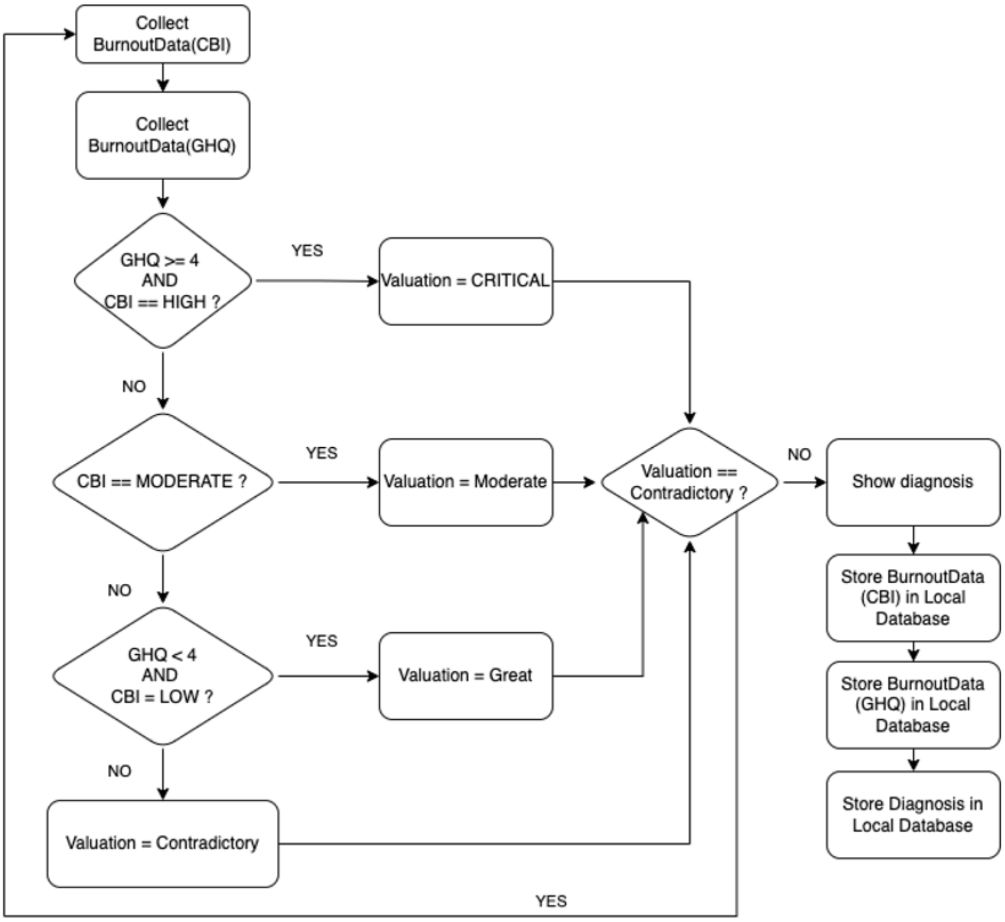
Diagnosis

Figure 2 explains how the BurnOut app obtains the diagnosis as a combination of GHQ (see the *GHQ Test* section) and CBI (see the *CBI Test* section) tests. There were four possible diagnoses:

1. Critical: high risk on the GHQ (the outcome is true) and high burnout risk on the CBI
2. Moderate: moderate burnout risk on the CBI (independently of the GHQ results)
3. Great: no risk on the GHQ and low burnout risk on the CBI
4. Contradictory: the app recommends repeating the tests

The resulting diagnosis is visualized and registered in the local database, as well as in the GHQ and CBI tests. The recommendation is to self-administer the tests and repeat them every 3 months.

Figure 2. Diagnosis procedure. CBI: Copenhagen Burnout Inventory; GHQ: Goldberg Health Questionnaire.


Collect BurnoutData(CBI)
Collect BurnoutData(GHQ)
GHQ >= 4 AND CBI == HIGH ?
YES
Valuation = CRITICAL
NO
CBI == MODERATE ?
YES
Valuation = Moderate
Valuation == Contradictory ?
NO
Show diagnosis
Store BurnoutData (CBI) in Local Database
Store BurnoutData (GHQ) in Local Database
Store Diagnosis in Local Database
NO
GHQ < 4 AND CBI = LOW ?
YES
Valuation = Great
NO
Valuation = Contradictory
YES

Monitoring

There are 3 ways of monitoring the evolution of the user's health condition.

Burnout Stats Chart

The user can check the diagnosis of the past 4 CBI tests in a column chart.

Last Burnout Test Stats

A bar chart that hosts the last time the user passed the CBI test, indicating which state has improved or worsened to have a clearer view of the evolution and strengths and weaknesses. To increase the user's feeling of doing well, when the dimensions are 0 after the last time they passed the test, instead of the chart, a message of congratulations appears to increase the user's satisfaction, highlighting that the user is on the right path to wellness and encourages them to continue with the process.

Mental Health Indicator

If psychological discomfort is obtained in the GHQ test, a highlighted *Alert* message will appear to warn the user that they are in a critical state. If the state of mental health from the GHQ test diagnosis is defined as wellness, a message of congratulations will appear to increase the user's satisfaction. This will encourage them to continue the process. In other words, it will increase adherence to using the BurnOut app.

BurnOut App Implementation

We attempted to follow guidelines to ensure a pleasant design and an easy-to-use app. The reason is that iOS mobile personal health record apps have better usability scores than Android apps, as stated in the study by Zapata et al [29], after analyzing a wide range of apps.

Currently, there are many alternative technologies to develop apps for mobile devices. Developing the app twice, in the native code for Android and iOS, implies extra production costs. Furthermore, the use of open-source tools for creating cross-platform apps using HTML, Cascading Style Sheets, and JavaScript with a framework such as *PhoneGap* could decrease performance and lead to a lack of advanced device-specific features provided by the latest application programming interfaces (APIs).

The BurnOut app was developed using an efficient approach—*Xamarin*, a Microsoft cross-platform framework that offers native performance and API access to provide a native user experience.

Xamarin allows the use of Model-View-ViewModel (MVVM) pattern, which implies making a clean separation between the logic and UI. This implies that Android and iOS will use the same code on the logic side, called core (shared logic code), leading to faster development by reducing duplicated code. When modifying the code, it is not essential to change the logic on both platforms—only from the specific *View*. Then, specific platform *Views* must be developed using native code for both Android and iOS. The framework provides methods to bind the UI with the core, and it will be automatically updated when a core property is changed.

Owing to the MVVM pattern, there is a clean separation among the View, which is the UI shown to the user on the screen; the Model, which contains the data of the current view; and the View-Model, which handles the communication between the view and the model.

The business and validation logic is performed just once as they are included in the core. This method reduces the coding as the core is shared between the platforms.

Thus, the BurnOut app is written using native API access and native performance and has easy maintainability while providing faster development.

However, the Xamarin framework and MVVM patterns have some drawbacks. As it is designed for cross-platform apps, unnecessary libraries are loaded at the app start-up, resulting in a noticeably slow process and an increase in the required memory size.

Local Database

The BurnOut app uses 2 storage systems: a NoSQL local database, which is used by both Android and iOS, and *Akavache*, an asynchronous, persistent key-value storage system.

Cloud Synchronization

The BurnOut app provides cloud synchronization among multiple devices, achieved by a server that implements a representational state transfer service with Spring. Representational state transfer has quickly become the de facto standard for building services on the web as it is easy to build and consume. In addition, it provides application security (encryption and authentication). Caching is built into the protocol. Service routing through the domain name system is a resilient and well-known system that is ubiquitously supported.

Cloud storage is implemented following the same technique. It stores the key-value registries on a server. Cloud data are asynchronously synchronized in the background so that the synchronization process is almost unnoticeable to the user. It performs incremental backup; therefore, it only uploads new or modified data. When users sign in, the local database is updated according to the data stored in the cloud. When a user signs out, the local database is erased as it would no longer be used and would already be saved in the cloud. This avoids inconsistencies between the local database and the cloud.

Results

Overview

In this section, the robustness, usability, and efficiency of the BurnOut app are tested.

To rigorously evaluate the application robustness, we recorded all the crashes that occur when using the app to control where and when they occur.

To measure usability, people had to perform several actions within the app, such as registering, introducing data, testing diagnoses, and monitoring tools. They also rated the user experience of the BurnOut app between 1 and 5. In addition, changes made based on their observations helped make the BurnOut app more user-friendly.

The efficiency of the BurnOut app is compared with that of Android and iOS devices in Table 6. The efficiency parameters measured were start-up, diagnosis generation, and log-in. Start-up is defined as the elapsed time for an app to start. Diagnosis generation is the elapsed time the app spends generating a personalized diagnosis plan for the user's needs. Log-in is defined as the time required for the app to log the user in and set up their health status for monitoring and fetching the data from the cloud or database. These times were the averages of 3 different measurements. Times <1 second guarantee that the user's train of thought remains uninterrupted [30].

Table 6. Devices used to test the Burnout app efficiency.

Operating system and device	Version
Android	
Xiaomi Redmi Note 8	Android 10
Xiaomi Redmi Note 5	Android 9
Samsung Galaxy S4	Android 5.0.2
Samsung Galaxy J3	Android 5.1
LG G2	Android 4.4
iOS	
iPhone X	iOS 14
iPhone 8	iOS 11
iPhone 6 Plus	iOS 10.3
iPhone 6	iOS 10.3
iPhone 5	iOS 10.3

Robustness

A total of 17 crashes were detected during a testing period of 15 days of using BurnOut. Most of them occurred when loading data from the local database asynchronously (3/17, 18%), loading data from the cloud database (11/17, 65%; considering that it is still under implementation), and logging in (4/17, 24%). Furthermore, 11% (2/17) of them were because of simple programming errors, such as null pointer exception, memory allocation, and communication between app and server, and others were produced by database exceptions because of bad queries and mistreatment of asynchronous behavior.

Usability

The first users who evaluated the BurnOut app were positive toward it and, in general, thought that its usability was good. Some encountered issues that were mentioned in the study by McIlroy et al [31] regarding problems when reviewing mobile apps.

To measure the usability of the BurnOut app, a short survey of the users mentioned in the *User Testing* was section conducted using the industry-standard System Usability Scale (SUS) [32]. It comprises a 10-item questionnaire with 5 response options ranging from strongly agree to strongly disagree. It enables the evaluation of a wide variety of products and services, including hardware, software, mobile devices, websites, and applications. The SUS was chosen as its characteristics fit our interests perfectly: it is a very easy scale to administer to participants, it can be used on small sample sizes with reliable results, and it is valid (ie, it can effectively differentiate between usable and unusable systems).

The participants ranked each question from 1 to 5 based on how much they agreed with the statement they were reading. A score of 5 meant they agreed completely, and a score of 1 meant they disagreed vehemently. All the testers answered the survey. The group had representative proportions of age, sex, and physical condition compared with the complete treatment group. Table 7 presents the obtained SUS results.

The rationale behind the calculation is very intuitive. The total score is 100, and each question weighs 10 points.

As odd-numbered questions are all in a positive tone, if the response is *strongly agree*, they are given the maximum score, which is 10 for each question. If the response is *strongly disagree*, they are given the minimum score, which is 0. By subtracting 1 from each of the odd-numbered questions, we ensure that the minimum is 0. Then, by multiplying by 2.5, we ensure that the maximum is 10 for each question.

In contrast, if the response is *strongly agree* for the even-numbered questions in a negative tone, they are given the minimum score, which is 0 for each question. If the response is *strongly disagree*, they are given the minimum score, which is 0. Thus, by subtracting the points for each question from 5, we ensure that the minimum is 0. Then, by multiplying by 2.5, we ensure the maximum is 10 for each of the questions.

Once all the results have been obtained, we calculate the average value for each test, thus obtaining the final score. In this case, we obtained a score of 95.8. These results are incredibly satisfying as, according to the general guideline for the interpretation of an SUS score (Table 8), we achieved an excellent rating.

The results obtained were consistent with the sensations shared by the users after performing the tests. All of them considered that the web part was very intuitive and did not require much time to learn how to get the most out of it, whereas all of them highlighted the simplicity-usefulness relationship of the bot.

Table 7. System Usability Scale survey with 40 randomly selected users in Lleida and Tarragona.

Questions	1	2	3	4	5
I think that I would like to use this system frequently	0	0	0	7	33
I found the system unnecessarily complex	40	0	0	0	0
I thought the system was easy to use	0	0	0	11	29
I think that I would need the support of a technical person to be able to use this system	30	10	0	0	0
I found the various functions in this system were well integrated	0	0	0	16	24
I thought there was too much inconsistency in this system	40	0	0	0	0
I would imagine that most people would learn to use this system very quickly	0	0	0	3	37
I found the system very cumbersome to use	34	6	0	0	0
I felt very confident using the system	0	0	2	8	30
I needed to learn a lot of things before I could get going with this system	35	5	0	0	0

Table 8. Interpretation of System Usability Scale score.

System Usability Scale score	Rating
>80.3	Excellent
68-80.3	Good
68	Okay
51-68	Poor
<51	Awful

Efficiency

First, the start-up efficiency was evaluated (Multimedia Appendix 1). The first thing to point out here is that using Xamarin.Forms alongside the MVVMCross framework instead of using native languages added a noticeable delay when starting up the app on both Android and iOS. This is because it usually loads several libraries.

Multimedia Appendix 2 shows the measured time taken to obtain a diagnosis on both platforms. A performance analogous to that obtained by the start-up was obtained. In general, the Android and iOS outcomes were very similar.

Multimedia Appendix 3 shows the diagnosis loading, although retrieving it from the local database once it is cached. The BurnOut app performs slightly faster on iOS than on Android. In absolute terms, this represents a difference of 0.48 seconds.

In addition, the time difference between the best and the worst result when generating the diagnosis was 0.656 seconds and 0.652 seconds on Android and iOS, respectively. Overall, we can assure that the iOS version was slightly faster than the Android version.

When comparing diagnosis loading from the local database on the user progress section, Android was slightly faster than iOS by just 0.434 milliseconds; however, on older devices, the difference was very significant. For example, iPhone 5 was 1.6 times faster than LG G2 when performing the same task.

As a concluding remark, BurnOut app operations are fluent with high response times because of the implementation of asynchronous tasks to compensate for the slowness of the Xamarin framework. According to the results, we can asseverate that the performance difference between Android and iOS is very low.

Discussion

Principal Findings

The BurnOut app is a useful, user-friendly app that provides the most accurate possible diagnosis approach, focusing on the psychosocial risks that cause burnout syndrome according to the CBI and monitoring user evolution over time in a cross-platform system with interesting extras such as mental health evaluation through the GHQ.

The BurnOut app offers the main functionalities that a potential patient of burnout syndrome may need, as shown in the *Results* section.

The BurnOut app is robust and user-friendly as the users who took the test had an SUS score of 95.8 out of 100, which qualifies as excellent.

Regarding start-up efficiency, the app was noticeably slower than the app using native tools (not shown in this paper). This, through the use of Xamarin, is because of some extra libraries being added, which took some time to load; however, this is strictly necessary as it allows us to use MVVMCross framework to save a lot of time by sharing some code and the ability to fix any bug once instead of fixing it on each platform. In general, the Android and iOS outcomes regarding start-up when using Xamarin were very similar.

In contrast, it is not easy to compare the performance with the other apps listed in Table 9, as each app offers different features. None of them generates a custom diagnosis based on CBI and GHQ results, which is one of the main advantages of our app.

In addition, the BurnOut app uses a local database and will soon use cloud storage. This means that data would never be lost because of mobile malfunctions as the data would be constantly synchronized and stored in the cloud. Upload and download operations would take at most 3 to 5 seconds. In addition, synchronization, as well as local storage in the database, would be performed asynchronously in the background to avoid poor user experience.

Some of the apps shown in Table 9 do not have synchronization in the cloud and therefore do not communicate with any server. Thus, waiting times were almost eliminated. However, time penalties in the BurnOut app are imperceptible to the users, and we can assure them that they will never lose information as it is saved in a server cloud.

Table 9. BurnOut app market popularity.

App	Downloads (thousands)	Ratings	Average rating	Pro version
BurnOut app	N/A[a]	N/A	N/A	No
MindDoc	>3,000,000	37,341	4.4	Yes
Psychosomat	>10,000	117	4.3	Yes
Breathe Pro	>100,000	662	4.5	Yes

[a] N/A: not applicable.

Comparison With Prior Work

The available tools to measure burnout among health care professionals have various strengths and limitations. Most health care systems will be able to find a validated instrument or instruments that meet their particular needs and situations [33]. Table 10 summarizes 7 common tests in terms of their overall strengths and limitations. The most commonly used instrument to measure burnout among health care professionals is the Maslach Burnout Inventory [30]. It comprises 3 domains: emotional exhaustion, depersonalization, and a low sense of personal accomplishment. Other instruments available to measure burnout include the Oldenburg Burnout Inventory [34] and CBI [23]. The Oldenburg Burnout Inventory has 3 domains: physical, cognitive, and affective exhaustion and disengagement from work. The CBI has 3 domains: personal (physical and psychological fatigue and exhaustion), work (physical and psychological fatigue and exhaustion related to work), and client-related (or a similar term such as patient and student) burnout. Some health systems and investigators use the Physician Worklife Survey single item (“Overall, based on your definition of burnout, how would you rate your level of burnout?”) to measure burnout symptoms [35].

A wide range of preventive actions was reported by Eurofound’s correspondents, from awareness-raising activities such as information campaigns to training, consultation with health professionals, sharing examples of good practices, and the provision of tools to conduct risk assessments on stress and early detection of burnout [18,36].

The US National Academy of Medicine published a valid and reliable list of instruments for measuring burnout [37]. Each tool has its advantages and drawbacks, and some are more appropriate for specific populations or settings.

Let us examine the burnout app market. *Moodpath*, *Psychosomat*, and *BreathePro* are the most popular Android apps. *MindDoc* is also available for iOS.

Moodpath provides support for depression, psychotherapy, and mental health. It is based on a 2-week depression screening. It functions as a mood diary. It provides a list of the symptoms detected in the diary. It provides the user’s mental health assessment and helps understand the psychology behind the user’s mood. *Psychosomat* provides support for depression, including on the basis of burnout syndrome. Support is provided before, during, and after psychotherapy, as well as in self-discovery processes. Owing to the user-defined criteria, it is also suitable for bipolar disorders. *BreathePro* is a kind of breathing training program. It helps avoid burnout and stress in users and measures stress resistance through an iPhone camera.

Jung et al [38] found that customer ratings were more critical to the survival of free apps, and there was also a benefit from entering the markets early. Various app store analyses were presented in the study by Martin et al [39]. A strong correlation between rating and downloads (popularity) and the fact that free apps have higher ratings than nonfree apps were the most interesting findings. Tian et al [40] studied 1492 high- and low-rated apps from Google Play. They concluded that the size of the app, number of promotional images on the store page, and target Software Development Kit version are the features that most accurately differentiate apps with high ratings from those with low ratings.

Table 9 presents a comparison between the burnout apps according to downloads, ratings (how many users have evaluated that app), average rating (0-5), and the availability of its Pro version. Table 11 shows a comparison between the main features provided by the burnout apps.

All the apps except the BurnOut app provide similar features and share the same weakness. All of them focus on offering the monitoring of symptoms.

Another important concluding remark of this analysis is that overall, the analyzed apps do not offer any kind of interactivity or feedback between the patients and the clinicians.

The BurnOut app is the only app that supports burnout test diagnosis. It uses the CBI and GHQ tests. CBI and GHQ are royalty free; hence, their use is free. This is a significant advantage when the goal is to develop a free app.

Table 10. Strengths and limitations of burnout measures [33].

Test	Strengths	Limitations
MBI[a] 22 items	Strong psychometrics; robust data showing scores correlated with outcomes of interest; can detect meaningful effect sixes from interventions	Cost; length; moderately complex to analyze; may not be sensitive to change over a short time frame
MBI 2 items	Strong psychometrics; short; robust data showing scores correlated with outcomes of interest	Cost; may not be sensitive to change
CBI[b] 16 items	May be used by all professionals; free	Length; moderately complex to analyze; limited data showing scores correlated with outcomes of interest among health care professionals in the United States
OBI[c] 19 items	May be used by all professionals; free	Length; moderately complex to analyze; limited data showing scores correlate with outcomes of interest
PWS[d] 1 item	Short; free; simple to analyze; may be used by all professionals	Length; limited data showing scores correlated with outcomes of interest; too brief to have strong psychometrics; limited emotional exhaustion domain of burnout

[a]MBI: Maslach Burnout Inventory.

[b]CBI: Copenhagen Burnout Inventory.

[c]OBI: Oldenburg Burnout Inventory.

[d]PWS: Physician Worklife Survey.

Table 11. Burnout app features.

App	User-friendly interface[a]	External widget support	Cloud synchronization	Depression support	Burnout diagnosis	Mental health diagnosis	Customized assessment (based on results)	Personal contact (human support)
BurnOut	Yes	No	Yes	No	Yes	Yes	Yes	No
MindDoc	Yes	No	Yes	Yes	Yes	Yes	Yes	No
Psychosomat	Yes	No	No	Yes	No	No	No	No
Breathe Pro	Yes	No	No	No	Yes	No	No	No

[a]According to the Google guidelines stated in Google Design Guidelines.

Limitations

The BurnOut app was not evaluated in a clinical trial with a balanced cohort comprising an intervention group of real diagnosed patients with burnout and another control group of people without this syndrome. The difficulty is to collect reliable results. Potential patients should access the app and obtain a diagnosis. This makes it very difficult to obtain satisfactory results.

Conclusions

The BurnOut app provides potential patients with burnout syndrome with a diagnosis generator based on their needs and goals. It is also important to understand the contextual use of the app and how it affects communication services [41]. This study analyzes the most popular operating systems and device approaches and provides an overview of the possible frameworks for building a multiplatform app. In addition, the BurnOut app integrates native iOS and Android services to provide a powerful but native feel and look.

Both the iOS and Android versions of the BurnOut app were implemented using Xamarin, a Microsoft framework. It was proven that the BurnOut app is robust. The Android and iOS versions were compared in terms of several key points. The efficiency on both platforms was tested, resulting in very similar performance.

The user experience was enhanced by using native APIs. This hides the fact that a multiplatform framework was used to build the app.

Overall, the usability (with a 95.8 SUS score), efficiency, and robustness of the app were good enough, as the people who were surveyed had no problems when using the app, and most of them felt confident (Table 7). We know that users are willing to use personal health records [42], and it has been demonstrated that it can help users reach their goals successfully while keeping track of their health data and assisting them to follow a personalized plan that will fulfill their needs perfectly and help improve their wellness.

Future trends could focus on the implementation of notifications to keep users motivated during long working days and encourage them. These notification messages and tips and the frequency with which they are sent would depend on the user's registered results and health state. In addition, it could be used to remind them to continue using the app and encourage them to continue with the process.

As the View-Model logic is the same across platforms such as Android, iOS, or the web, an interesting new feature would be porting the app to the web to make it easy for the users to access all their web-based data.

Acknowledgments

This work was supported by project PID2020-113614RB-C22, funded by MCIN/AEI/10.13039/501100011033. JV is a Serra Húnter fellow.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Start-up efficiency (in seconds).
[PNG File , 19 KB - medinform_v10i9e30094_app1.png]

Multimedia Appendix 2
Diagnosis generation efficiency (in seconds).
[PNG File , 18 KB - medinform_v10i9e30094_app2.png]

Multimedia Appendix 3
Local stats loading for tracking user progress efficiency (in seconds).
[PNG File , 20 KB - medinform_v10i9e30094_app3.png]

References

1. Freudenberger HJ, Richelson G. Burnout: The High Cost of High Achievement. Norwell, MA, USA: Anchor Press; 1980.
2. Maslach C, Jackson SE. The measurement of experienced burnout. J Organiz Behav 1981 Apr;2(2):99-113. [doi: 10.1002/job.4030020205]
3. International Classification of Diseases 11th Revision. World Health Organization. 2018. URL: https://icd.who.int/en [accessed 2022-08-09]
4. Balch CM, Freischlag JA, Shanafelt TD. Stress and burnout among surgeons: understanding and managing the syndrome and avoiding the adverse consequences. Arch Surg 2009 Apr;144(4):371-376. [doi: 10.1001/archsurg.2008.575] [Medline: 19380652]
5. Linzer M, Manwell LB, Williams ES, Bobula JA, Brown RL, Varkey AB, MEMO (Minimizing Error, Maximizing Outcome) Investigators. Working conditions in primary care: physician reactions and care quality. Ann Intern Med 2009 Jul 07;151(1):28-W9. [doi: 10.7326/0003-4819-151-1-200907070-00006] [Medline: 19581644]
6. Shanafelt TD, Balch CM, Bechamps G, Russell T, Dyrbye L, Satele D, et al. Burnout and medical errors among American surgeons. Ann Surg 2010 Jun;251(6):995-1000. [doi: 10.1097/SLA.0b013e3181bfdab3] [Medline: 19934755]
7. Kumar S. Burnout and doctors: prevalence, prevention and intervention. Healthcare (Basel) 2016 Jun 30;4(3):37 [FREE Full text] [doi: 10.3390/healthcare4030037] [Medline: 27417625]
8. Byers J, Gale L. Physician burnout: what can be done? HealthCareDive. 2016 Nov 7. URL: https://www.healthcaredive.com/news/physician-burnout-strategies-regulatory-burden/428523/ [accessed 2022-08-09]
9. Maslach C, Schaufeli WB, Leiter MP. Job burnout. Annu Rev Psychol 2001;52:397-422. [doi: 10.1146/annurev.psych.52.1.397] [Medline: 11148311]
10. Schaufeli WB, Bakker AB, Hoogduin K, Schaap C, Kladler A. on the clinical validity of the maslach burnout inventory and the burnout measure. Psychol Health 2001 Sep;16(5):565-582. [doi: 10.1080/08870440108405527] [Medline: 22804499]
11. Balch CM, Shanafelt T. Combating stress and burnout in surgical practice: a review. Adv Surg 2010;44:29-47. [doi: 10.1016/j.yasu.2010.05.018] [Medline: 20919512]
12. Shanafelt TD, Boone S, Tan L, Dyrbye LN, Sotile W, Satele D, et al. Burnout and satisfaction with work-life balance among US physicians relative to the general US population. Arch Intern Med 2012 Oct 08;172(18):1377-1385. [doi: 10.1001/archinternmed.2012.3199] [Medline: 22911330]
13. Dyrbye LN, Varkey P, Boone SL, Satele DV, Sloan JA, Shanafelt TD. Physician satisfaction and burnout at different career stages. Mayo Clin Proc 2013 Dec;88(12):1358-1367. [doi: 10.1016/j.mayocp.2013.07.016] [Medline: 24290109]
14. Lacy BE, Chan JL. Physician burnout: the hidden health care crisis. Clin Gastroenterol Hepatol 2018 Mar;16(3):311-317. [doi: 10.1016/j.cgh.2017.06.043] [Medline: 28669661]

15. Han S, Shanafelt TD, Sinsky CA, Awad KM, Dyrbye LN, Fiscus LC, et al. Estimating the attributable cost of physician burnout in the United States. Ann Intern Med 2019 Jun 04;170(11):784-790. [doi: 10.7326/M18-1422] [Medline: 31132791]
16. Goldberg DG, Soylu TG, Grady VM, Kitsantas P, Grady JD, Nichols LM. Indicators of workplace burnout among physicians, advanced practice clinicians, and staff in small to medium-sized primary care practices. J Am Board Fam Med 2020;33(3):378-385 [FREE Full text] [doi: 10.3122/jabfm.2020.03.190260] [Medline: 32430369]
17. Goldberg DP. The Detection of Psychiatric Illness by Questionnaire: A Technique for the Identification and Assessment of Non-psychotic Psychiatric Illness. Oxford, UK: Oxford University Press; 1972.
18. Aumayr-Pintar C, Cerf C, Parent-Thirion A. Burnout in the workplace: a review of data and policy responses in the EU. Eurofound. Luxembourg, Luxembourg: Publications Office of the European Union; 2018 Sep 10. URL: https://www.eurofound.europa.eu/publications/report/2018/burnout-in-the-workplace-a-review-of-data-and-policy-responses-in-the-eu [accessed 2022-08-10]
19. Mental Health Starts with You. MindDoc. URL: https://minddoc.com/us/en [accessed 2022-08-09]
20. Brecht C. Psychosomat. URL: https://christian-brecht.de/psychosomat/ [accessed 2022-08-09]
21. Breathe Pro. URL: https://baixarapk.gratis/en/app/1105390591/breathe-pro [accessed 2022-08-09]
22. Lafraxo MA, Ouadoud M, El Madhi Y, Rehali M, Soulaymani A. Burnout syndrome prevention measures among nursing staff: implementing a mobile application based on MIT's app inventor tool using the scratch programming code. Int J Onl Eng 2021 Apr 06;17(04):81-95. [doi: 10.3991/ijoe.v17i04.20393]
23. Kristensen TS, Borritz M, Villadsen E, Christensen KB. The Copenhagen Burnout Inventory: a new tool for the assessment of burnout. Work Stress 2005 Jul;19(3):192-207. [doi: 10.1080/02678370500297720]
24. Bria M, Spânu F, Băban A, Dumitraşcu DL. Maslach Burnout Inventory – General Survey: factorial validity and invariance among Romanian healthcare professionals. Burnout Res 2014 Dec;1(3):103-111. [doi: 10.1016/j.burn.2014.09.001]
25. BurnOut App. Google Play Store. URL: https://play.google.com/store/apps/details?id=com.jordi.medical.AwesomeApp [accessed 2022-08-15]
26. Goldberg DP, Williams P. A User's Guide to the General health Questionnaire. Slough, UK: NFER-Nelson; 1988.
27. Kulic L, Galjak M, Jovanovic J. Correlation of sociodemographic variables and interpersonal sources of stress at work among miners. Vojnosanitetski Pregled 2019;76(11):1169-1177. [doi: 10.2298/vsp170708030k]
28. Suri K, Hegde SK, Sadanand S, Randhawa S, Bambrah HS, Turlapati S. Psychological distress and burnout among counsellors working in health information helplines. Int J Community Med Public Health 2021;8(1):304-310. [doi: 10.18203/2394-6040.ijcmph20205712]
29. Cruz Zapata B, Hernández Niñirola A, Idri A, Fernández-Alemán JL, Toval A. Mobile PHRs compliance with Android and iOS usability guidelines. J Med Syst 2014 Aug;38(8):81. [doi: 10.1007/s10916-014-0081-6] [Medline: 24957397]
30. Maslach C, Jackson SE, Leiter MP. Maslach Burnout Inventory. 4th edition. Palo Alto, CA, USA: Mind Garden; 2016.
31. McIlroy S, Ali N, Khalid H, E. Hassan AE. Analyzing and automatically labelling the types of user issues that are raised in mobile app reviews. Empir Software Eng 2016 Jun;21(3):1067-1106. [doi: 10.1007/s10664-015-9375-7]
32. Brooke J. SUS: a retrospective. J Usability Stud 2013 Feb;8(2):29-40.
33. Dyrbye LN, Meyers D, Ripp J, Dalal N, Bird SB, Sen S. A pragmatic approach for organizations to measure health care professional well-being. National Academy of Medicine. 2018 Oct 1. URL: https://nam.edu/a-pragmatic-approach-for-organizations-to-measure-health-care-professional-well-being/ [accessed 2022-08-09]
34. Demerouti E, Bakker AB. The Oldenburg Burnout Inventory: A Good Alternative to Measure Burnout and Engagement. Handbook of Stress and Burnout in Health Care. 2008. URL: https://www.academia.edu/2796247/The_Oldenburg_Burnout_Inventory_A_good_alternative_to_measure_burnout_and_engagement [accessed 2022-08-15]
35. Dolan ED, Mohr D, Lempa M, Joos S, Fihn SD, Nelson KM, et al. Using a single item to measure burnout in primary care staff: a psychometric evaluation. J Gen Intern Med 2015 May;30(5):582-587 [FREE Full text] [doi: 10.1007/s11606-014-3112-6] [Medline: 25451989]
36. Arora M, Diwan AD, Harris IA. Prevalence and factors of burnout among Australian orthopaedic trainees: a cross-sectional study. J Orthop Surg (Hong Kong) 2014 Dec;22(3):374-377 [FREE Full text] [doi: 10.1177/230949901402200322] [Medline: 25550022]
37. Valid and Reliable Survey Instruments to Measure Burnout, Well-Being, and Other Work-Related Dimensions. US National Academy of Medicine. URL: https://dotymed.files.wordpress.com/2016/10/valid-and-reliable-survey-instruments-to-measure-burnout-well-being-and-other-work-related-dimensi.pdf [accessed 2022-08-09]
38. Jung EY, Baek C, Lee JD. Product survival analysis for the App Store. Mark Lett 2012 Sep 26;23(4):929-941. [doi: 10.1007/s11002-012-9207-0]
39. Martin W, Sarro F, Jia Y, Zhang Y, Harman M. A survey of App Store analysis for software engineering. IIEEE Trans Software Eng 2017 Sep 1;43(9):817-847. [doi: 10.1109/tse.2016.2630689]
40. Tian Y, Nagappan M, Lo D, Hassan AE. What are the characteristics of high-rated apps? A case study on free Android Applications. In: Proceedings of the 2015 IEEE International Conference on Software Maintenance and Evolution. 2015 Presented at: ICSME '15; September 29-October 1, 2015; Bremen, Germany p. 301-310. [doi: 10.1109/icsm.2015.7332476]
41. Karikoski J, Soikkeli T. Contextual usage patterns in smartphone communication services. Pers Ubiquit Comput 2013 Mar;17(3):491-502. [doi: 10.1007/s00779-011-0503-0]

42. Fernández-Alemán JL, Seva-Llor CL, Toval A, Ouhbi S, Fernández-Luque L. Free Web-based personal health records: an analysis of functionality. J Med Syst 2013 Dec;37(6):9990. [doi: 10.1007/s10916-013-9990-z] [Medline: 24221916]

Abbreviations

API: application programming interface
CBI: Copenhagen Burnout Inventory
GHQ: Goldberg Health Questionnaire
MVVM: Model-View-ViewModel
SUS: System Usability Scale
UI: user interface

Edited by C Lovis; submitted 30.04.21; peer-reviewed by A Hidki, M Sung, B Chaudhry; comments to author 10.07.21; revised version received 19.05.22; accepted 15.07.22; published 06.09.22.

Please cite as:
Godia J, Pifarré M, Vilaplana J, Solsona F, Abella F, Calvo A, Mitjans A, Gonzalez-Olmedo MP
A Free App for Diagnosing Burnout (BurnOut App): Development Study
JMIR Med Inform 2022;10(9):e30094
URL: https://medinform.jmir.org/2022/9/e30094
doi:10.2196/30094
PMID:36066932

Original Paper

An Assessment of Mentions of Adverse Drug Events on Social Media With Natural Language Processing: Model Development and Analysis

Deahan Yu[1], MHI; V G Vinod Vydiswaran[1,2], PhD

[1]School of Information, University of Michigan, Ann Arbor, MI, United States

[2]Department of Learning Health Sciences, Medical School, University of Michigan, Ann Arbor, MI, United States

Corresponding Author:
V G Vinod Vydiswaran, PhD
Department of Learning Health Sciences
Medical School
University of Michigan
1161F NIB
300 N Ingalls St
Ann Arbor, MI, 48109
United States
Phone: 1 734 647 1207
Fax: 1 734 647 3914
Email: vgvinodv@umich.edu

Abstract

Background: Adverse reactions to drugs attract significant concern in both clinical practice and public health monitoring. Multiple measures have been put into place to increase postmarketing surveillance of the adverse effects of drugs and to improve drug safety. These measures include implementing spontaneous reporting systems and developing automated natural language processing systems based on data from electronic health records and social media to collect evidence of adverse drug events that can be further investigated as possible adverse reactions.

Objective: While using social media for collecting evidence of adverse drug events has potential, it is not clear whether social media are a reliable source for this information. Our work aims to (1) develop natural language processing approaches to identify adverse drug events on social media and (2) assess the reliability of social media data to identify adverse drug events.

Methods: We propose a collocated long short-term memory network model with attentive pooling and aggregated, contextual representation generated by a pretrained model. We applied this model on large-scale Twitter data to identify adverse drug event–related tweets. We conducted a qualitative content analysis of these tweets to validate the reliability of social media data as a means to collect such information.

Results: The model outperformed a variant without contextual representation during both the validation and evaluation phases. Through the content analysis of adverse drug event tweets, we observed that adverse drug event–related discussions had 7 themes. Mental health–related, sleep-related, and pain-related adverse drug event discussions were most frequent. We also contrast known adverse drug reactions to those mentioned in tweets.

Conclusions: We observed a distinct improvement in the model when it used contextual information. However, our results reveal weak generalizability of the current systems to unseen data. Additional research is needed to fully utilize social media data and improve the robustness and reliability of natural language processing systems. The content analysis, on the other hand, showed that Twitter covered a sufficiently wide range of adverse drug events, as well as known adverse reactions, for the drugs mentioned in tweets. Our work demonstrates that social media can be a reliable data source for collecting adverse drug event mentions.

(JMIR Med Inform 2022;10(9):e38140) doi:10.2196/38140

KEYWORDS

natural language processing; machine learning; adverse drug event; pharmacovigilance; social media; drug; clinical; public health; health monitoring; surveillance; drug effects; drug safety

Introduction

Background

Adverse reactions to drugs are among the most significant concerns in both clinical practice and public health monitoring, but they do not have a consistent definition in the literature. According to Edwards and Aronson [1], side effects of a particular drug are defined as “unintended effects related to the pharmacological properties occurring at normal dose” of the drug. Unintended effects can be either harmful or beneficial. For example, β-blockers are mainly used for hypertension, but they can also relieve chest pain (or angina) in patients [1]. According to a World Health Organization (WHO) report [2], adverse reactions are defined as “any response to a drug that is noxious, is unintended, and occurs at doses normally used in humans.” A similar definition of adverse reaction was used by Asscher et al [3] and Pirmohamed et al [4], except that their definitions included the condition that the drug was used in its proper clinical application. In other words, the WHO definition allows for an improper use of a drug with a normal dose, while Asscher et al and Pirmohamed et al do not include such cases. A definition of adverse reactions by Karch and Lasagna [5] is similar but includes the effects of intentional overdoses and drug abuse. Although various definitions of adverse reactions are used, the most common component of these definitions is unintended consequences caused by or suspected to be due to the use of a drug [1-5].

Adverse events, on the other hand, are defined as “untoward occurrences following exposure to a drug but not necessarily caused by the drug” [1,3]. While the terms “adverse event” and “adverse reaction” are similar, they cannot be used interchangeably, because there is no causality assumption in the definition of adverse events, while there is a causality assumption in the definition of adverse reactions. Adverse reactions are reported to be among the top 10 leading causes of death [6,7]. To increase postmarketing surveillance of drugs and improve drug safety, multiple measures have been put into place. These include implementing spontaneous reporting systems, such as the US Food and Drug Administration Adverse Events Reporting System (FAERS) [2,7,8].

On the other hand, researchers have also looked at developing automated systems that use electronic health records and social media data [9-11] to collect experiences of adverse events that can be further investigated as possible adverse reactions. Recently, deep neural network–based models have been developed to detect adverse events in tweets [12-14]. Long short-term memory (LSTM) networks and pretrained language models, such as bidirectional encoder representations from transformers (BERT) [15] and generative pretraining language models [16], have been chosen as models for this application [12-14]. However, there is still room for improvement in the implementation of such systems [9-11]. Various neural network systems have been presented by other researchers, but no system to date incorporates both recurrent-based networks (eg, LSTM) and attention-based networks (eg, BERT). Capturing both sequentially processed output and contextually processed output could help the model better learn the data and the task. Lastly, machine learning and deep learning models have shown their effectiveness at detecting adverse event mentions in social media data [17], but it is still uncertain whether social media are valid as a data source for the purpose of adverse event detection.

Goal of This Study

In this paper, we use the term “adverse drug event” (ADE) rather than “adverse event.” We formulated the task of identifying ADE mentions from tweets as a classification task, that is, labeling tweets based on whether or not they contain a mention of an ADE. We propose a neural network–based framework that incorporates augmented medical representation and contextual representation to build a robust classification model. Our work aims to develop a natural language processing (NLP) system that identifies ADE mentions based on social media texts and to assess the reliability of social media data, especially Twitter, as a means to collect that information. Our research questions are as follows: “Could contextual representation from a pretrained language model help enhance a model for classifying ADE tweets?” and “Could social media be a reliable data source to collect mentions of ADEs?”

We conducted a comprehensive experimental analysis to validate the effectiveness of the model. In addition, we performed a systematic evaluation study to determine the reliability of Twitter as a data source for collecting mentions of ADEs. Our work makes the following empirical contributions: (1) we demonstrate that incorporating contextual representations with augmented medical representations significantly improves the performance of the adverse event detection task compared to not incorporating contextual representations, (2) we show that the current automated systems to identify mentions of ADEs in tweets are not sufficiently generalizable, and (3) we observe that Twitter covers a sufficiently wide range of ADEs relatively well, including known ADEs, and conclude that social media can be a reliable data source for collecting ADE mentions.

Related Work

Before a drug is released to market, an initial description of related ADEs is obtained through randomized controlled trials [18]. These trials may provide an initial description that is not fully complete [19]. Due to the incompleteness of the initial list of ADEs, pharmacovigilance plays a significant role in the postmarketing phase and is necessary to collect any new information on ADEs. Social media, including Twitter, have been explored as platforms for pharmacovigilance, such as by collecting mentions of ADEs through NLP [11,17,20-22]. The text of tweets is relatively short but still conveys information about patient experiences that are often self-disclosed. For a tweet to be considered ADE-related, the tweet must not only mention at least one adverse event, but must also mention a drug by name. Notably, a tweet cannot be considered ADE-related if there is no mention of drugs.

Data sets of labeled tweets for identifying mentions of ADEs have been developed to benchmark NLP systems in shared competitions [11-13,22-25]. These annotated data sets have allowed researchers to develop automated systems and compare them against each other. Early systems for identifying mentions of ADEs in tweets were based on curated lexicons, heuristic rules, pattern matching, or supervised machine learning approaches [23,24]. Various dictionary-based features, such as ADE lexicons, drug names, and medical concepts were explored, along with linguistic and sentiment analysis. Recently, neural network–based models have been a popular choice due to their outstanding performance [11].

Methods

Data Sources

We used 3 Twitter-based data sets to develop and evaluate our models—1 for training and 2 for evaluation. The training set and the first evaluation set were obtained from a shared task for automatic classification of English-language tweets that report adverse effects, organized as part of the 2020 Social Media Mining for Health (SMM4H) workshop [11]. According to the organizers of the shared task, the tweets were collected via Twitter's public streaming API. Generic and trade names for drugs, along with their common misspellings, were used as keywords to collect data. After the collection, the tweets were annotated independently by 2 annotators, with a Cohen κ for interannotator agreement of 0.82. Tweets with disagreements were reannotated until the pair reached consensus. The second evaluation set was obtained from a publicly available reference data set called WEB-RADR (web-recognizing adverse drug reactions), developed by Dietrich et al [22]. These tweets were collected in a similar fashion. Annotations were done by 2 teams of 9 annotators each; however, no measures for interannotator agreement are reported.

Table 1 summarizes the statistics of the 3 data sets. After preprocessing and removing duplicate tweets, there were 24,700 tweets in the training set. Of these, approximately 9% (2362) of them were labeled as ADE tweets, that is, tweets containing 1 or more mentions of adverse events along with at least 1 mention of a drug. The remaining 91% (22,338) of tweets were labeled as non-ADE tweets, meaning that these tweets did not contain any mention of an adverse reaction but contained a drug mention. Of 24,700 tweets, 20,098 (81.4%) were used to train the models while the other 4602 (18.6%) were used for validation. The distribution of ADE versus non-ADE tweets was more skewed in the SMM4H evaluation set. Of the 4759 tweets in the evaluation set, only 194 (4.1%) were labeled as ADE tweets and 4565 (95.9%) were labeled as non-ADE tweets.

We also evaluated our models on WEB-RADR [22], which we used as a second, independent data set. The original data set consists of 57,473 tweets, with 1056 tweets (1.8%) labeled as ADE tweets and 56,417 (98.2%) as non-ADE tweets. However, from the original data set, we were able to successfully collect only 34,369 (59.8%) tweets, possibly due to suspended accounts or deleted tweets. Of these, 645 (1.9%) were labeled as ADE tweets, while the remaining 33,724 (98.1%) were non-ADE tweets.

All tweets were preprocessed to separate punctuation marks, remove special characters and URLs, replace user mentions beginning with @, and replace text emoticons with a normalized token. No specific text cleaning packages were used.

Table 1. Statistics for the training and evaluation data sets.

Data set	Tweets, N	ADE[a] tweets, n	Non-ADE tweets, n	Unique drugs, n	Drugs in tweets but not in library, n
SMM4H[b] training	24,700	2362	22,338	1020	31
SMM4H evaluation	4759	194	4565	688	129
WEB-RADR[c] evaluation	34,369	645	33,724	685	25,646

[a]ADE: adverse drug event.

[b]SMM4H: Social Media Mining for Health.

[c]WEB-RADR: web-recognizing adverse drug reactions.

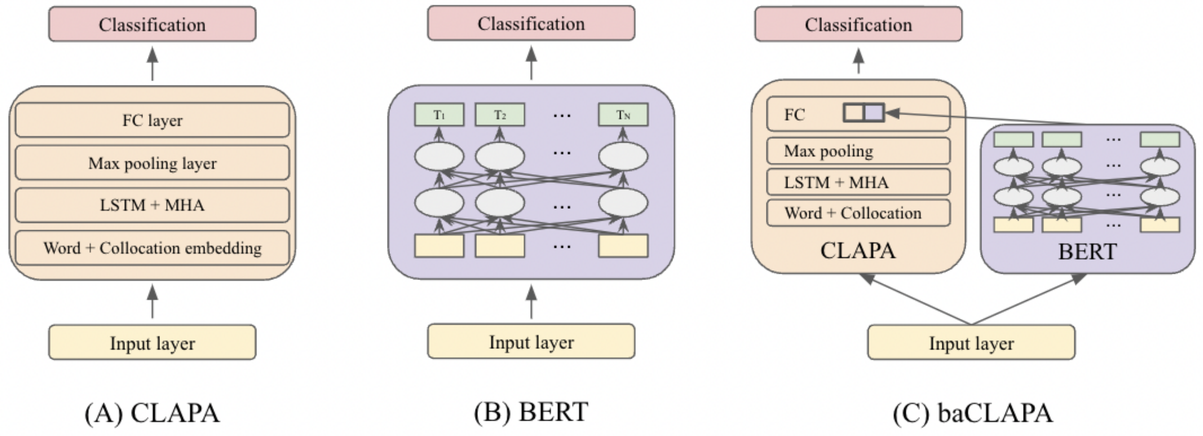
NLP System Development

Model Selection

In recent years, pretrained language models have been widely deployed as base models for numerous NLP tasks that can be fine-tuned to a data set for a particular downstream task, often referred to as transfer learning. Despite relatively simple training, such transfer learning approaches have been shown to be powerful tools for many NLP tasks, including ADE classification. Transfer learning makes downstream tasks successful because these language models are trained on a large corpus; hence, they gain strong representational power.

In our previous work, we proposed a collocated LSTM model with attentive pooling and aggregated representation (CLAPA) that utilized neighborhood information to build a better representation of medical concepts [26]. The model focused on enhancing medical concepts by incorporating neighborhood information through a collocation graph. While CLAPA enriched the representation of medical concepts, it had relatively weak representation of other context information, such as semantics. The capability of a pretrained model to provide a robust representation of context information may help assist CLAPA to learn better. With this motivation, we extended CLAPA to BERT-augmented CLAPA (baCLAPA), which incorporated BERT's logits with CLAPA's trained representation. BERT was chosen because it was the most competitive model among pretrained models reported in the 2019 SMM4H task [13]. The 3 models compared in this task are illustrated in Figure 1 and summarized below.

Figure 1. Schematic diagram of the 3 models that highlights how each model is configured. A: CLAPA; B: BERT; C: baCLAPA. baCLAPA: bidirectional encoder representations from transformers–assisted collocated long short-term memory with attentive pooling and aggregated representation; BERT: bidirectional encoder representations from transformers; CLAPA: collocated long short-term memory with attentive pooling and aggregated representation; FC: fully connected; LSTM: long short-term memory; MHA: multi-head attention.



(A) CLAPA (B) BERT (C) baCLAPA

CLAPA Model

CLAPA [26], illustrated in Figure 1A, uses collocation information to improve the representation of medical concepts. CLAPA requires three main components: (1) medical concepts, (2) a collocation graph, and (3) a model architecture.

First, for medical concepts, the generic names and brand names of medications were collected from MedlinePlus [27]. A few generic medical words or brand names, such as "Amen" and "Heather," were removed to reduce noise. Then, the list of medical concepts was expanded by adding medical words from tweets in the training set that were missing in the drug list obtained from MedlinePlus. As a result, a total of 4888 medical concepts were collected, including 4747 drug names from MedlinePlus and 141 drug names from the SMM4H training set.

Second, for the collocation graph, each unique word in the training set was assigned as a node, and edges were added between node pairs if the corresponding pair of words were adjacent to each other. After the graph was constructed, the graph was reduced by retaining only the closest 15 neighbor nodes per medical concept, following an empirical analysis of neighborhood size [26].

Third, for the model architecture, LSTM networks with 4 layers and 300 input sizes were implemented, followed by 3 multi-head attention layers and max pooling and pooling layers. FastText pretrained embedding [28] was used for word embedding. All hyperparameters were jointly trained with a learning rate of 0.001 and a cross-entropy loss function.

BERT Model

As another baseline model, we instantiated a BERT model [15], illustrated in Figure 1B. The bert-base-uncased model was used for classification and was tuned based on the recommendation for hyperparameter settings [15]. The BERT model was fine-tuned on the training set without any further modification on hyperparameters. Two tokens, [CLS] and [SEP], were added to the beginning and end of the input representation. Each sentence was tokenized through BertTokenizer and fed into the BERT model. Our BERT model contained the same parameters as the base model, with 12 layers, 768 input sizes, and 12 multi-head attentions. The hyperparameters of the classification layer were jointly trained with a learning rate of $5e^{-5}$.

baCLAPA Model

Our proposed baCLAPA model is illustrated in Figure 1C. The model consists of 2 parallel stacks—a CLAPA model and a BERT model. The input sentence feeds into both the CLAPA and BERT models. Each network independently learns input embeddings. Once each model produces the final hidden states, the states are reduced to representations with a size of 2, which are commonly referred to as logits. The raw output representation of BERT is then incorporated into CLAPA, either as large as the final hidden states or as small as logits, mapped to a 2-dimensional vector space for a binary classification. In the task presented in this paper, BERT's logits were used to assist CLAPA because logits provide a brief but comprehensive representation of how networks have learned from inputs. Thus, BERT's logits were concatenated with CLAPA's logits to generate predictions. In Figure 1C, 2 bold boxes inside the fully connected layer show how BERT's logits and CLAPA's logits are concatenated. Formally, this can be written as follows:

where refers to the last fully-connected layer in CLAPA, and are logits from CLAPA and BERT, and is the final logit. Once CLAPA and BERT produce their logits, BERT's logits are passed to CLAPA. Then, a concatenation of their logits is fed into , which produces . Finally, the final logits are fed into Softmax for binary classification.

Baselines

Two additional models were used as baselines. First, we used an SVM model with a linear kernel, with other hyperparameters set to default values. The input representation included a term frequency–inverse document frequency weighted representation with trigram features. As a second baseline, we used a random model with weighted distribution.

Validation Study to Determine the Reliability of Social Media as a Data Source

Study Questions

To validate the reliability of the social media data as a means to collect ADEs, we analyzed tweets that were collected by our baCLAPA model. This study aimed to answer two questions about social media data: (1) what kinds of ADEs are mentioned on Twitter? and (2) of the ADE mentions for each known drug on Twitter, how many also mentioned known adverse reactions listed in an authoritative source? Answering the first question would reveal how various kinds of ADEs are covered on social media, and answering the second would reveal how many relevant ADEs are mentioned on social media. The known adverse events were collected from MedlinePlus, an authoritative, popular, and credible website run by the US National Library of Medicine.

Obtaining ADE Tweets: Data Source and ADE Classification

The Twitter data used for this study were obtained from a paper by Vydiswaran et al [29]. The data were collected via the Twitter API, user timelines, and the Decahose stream, which is a 10% random sample of the real-time Twitter stream. First, the Twitter API and user timelines were used to collect all tweets from users near the Detroit metropolitan area. Then, the data set was expanded through the Decahose stream. In total, the data set contained 28.8 million tweets. More details about the data collection can be found in the paper by Vydiswaran et al [29].

First, the 28.8 million tweets were filtered through our drug list, which consisted of 4888 drug names. This step allowed us to sort out tweets containing at least one drug keyword. This let us identify 34,536 of 28.8 million tweets as drug-mentioning tweets. Then, our baCLAPA model was applied to those tweets and identified 1544 ADE tweets.

Qualitative Content Analysis of Tweets

We conducted a qualitative content analysis [30] to answer the two questions mentioned above: (1) how many different types of ADEs are covered on Twitter? and (2) how many ADE tweets about a particular drug identify an adverse event that is a known adverse reaction for that drug on MedlinePlus? We first extracted 139 unique drugs mentioned in the 1544 tweets. Then, we conducted a qualitative content analysis to derive themes for the ADEs within the 1544 tweets. During the qualitative coding process, we found that the drug word “caffeine” mostly referred to coffee and the word “vitamin” was too general to determine which vitamin supplement was taken. Therefore, tweets containing only these drug words were dropped—462 tweets for “caffeine” and 141 tweets for “vitamin”. A total of 941 ADE tweets were thus qualitatively analyzed. These tweets were manually coded to identify themes for ADEs until the themes were saturated. The themes were reviewed by a domain expert after the analysis was completed.

Once we identified the themes, we collected information about known adverse reactions for each drug through MedlinePlus and compared them against themes identified by the content analysis. For example, when analyzing ADE tweets about ibuprofen, we identified two themes: nausea and sweating. When reviewing information about ibuprofen on MedlinePlus, we only found relevant mentions of ibuprofen potentially causing nausea, and did not find any sweating-related adverse reactions. Thus, ibuprofen was paired with the nausea-related ADE theme as a known adverse reaction but not with the sweat-related ADE theme. This way, we linked all ADE tweets and known adverse reactions to a particular drug to each ADE theme.

Results

Experimental Results of the NLP System

We first present the performance of the models on the validation set. This allows us to compare the overall performance of the models, including the baselines. Both CLAPA and baCLAPA were evaluated on the SMM4H evaluation set [31]. We further evaluated the models on another data set, the WEB-RADR evaluation set, to validate whether the extended models performed better than the original models on various data sets.

As shown in Table 2, the random and SVM baseline models did not outperform the neural network–based models, but the recall score for the SVM model was the second highest. Of all models, baCLAPA performed the best for all performance metrics: precision, recall, and F1. On average, it performed about 0.026 F1 points better than CLAPA on the validation set.

To further evaluate our method, we picked the best CLAPA and baCLAPA models from the 10 validation runs. Their performance on the validation set is shown in the first 2 result rows of Table 3. On both evaluation data sets, baCLAPA outperformed CLAPA on the F1 metric. Precision and recall values are not available for CLAPA on the SMM4H evaluation set because it was used only for the best (baCLAPA) model [31]. While baCLAPA performed better for F1 score than CLAPA on the SMM4H evaluation set by 0.07, the improvement was relatively small on the WEB-RADR evaluation set. Most of this improvement was attributable to the significantly higher recall. CLAPA outperformed baCLAPA on the precision measure on WEB-RADR.

Table 2. Average performance of 10 runs on the validation set. Italics represent the best model for each performance metric.

Model	Precision (SD)	Recall (SD)	F1 score (SD)
Random	0.099 (0.01)	0.103 (0.01)	0.101 (0.01)
SVM[a]	0.386 (0)	0.638 (0)	0.481 (0)
CLAPA[b]	0.581 (0.03)	0.623 (0.03)	0.599 (0.01)
BERT[c]	0.54 (0.03)	0.602 (0.04)	0.567 (0.01)
baCLAPA[d]	*0.603* (0.02)	*0.652* (0.03)	*0.625* (0.007)

[a] SVM: support vector machine.

[b] CLAPA: collocated long short-term memory with attentive pooling and aggregated representation.

[c] BERT: bidirectional encoder representations from transformers.

[d] baCLAPA: bidirectional encoder representations from transformers–assisted collocated long short-term memory with attentive pooling and aggregated representation.

Table 3. Evaluation of collocated long short-term memory with attentive pooling and aggregated representation (CLAPA) and bidirectional encoder representations from transformers–assisted CLAPA (baCLAPA) on 2 evaluation sets. Italics represent the best model for each performance metric.

Data set and model	Precision	Recall	F1 score
Validation			
CLAPA[a]	0.563	0.649	0.603
baCLAPA[b]	0.589	0.676	*0.629*
SMM4H[c] evaluation			
CLAPA	—[d]	—	0.44
baCLAPA	0.48	0.54	*0.51*
WEB-RADR[e] evaluation			
CLAPA	0.356	0.386	0.371
baCLAPA	0.334	0.479	*0.394*

[a] CLAPA: collocated long short-term memory with attentive pooling and aggregated representation.

[b] baCLAPA: bidirectional encoder representations from transformers–assisted collocated long short-term memory with attentive pooling and aggregated representation.

[c] SMM4H: Social Media Mining for Health.

[d] Not available.

[e] WEB-RADR: web-recognizing adverse drug reactions.

Qualitative Content Analysis of ADE Tweets

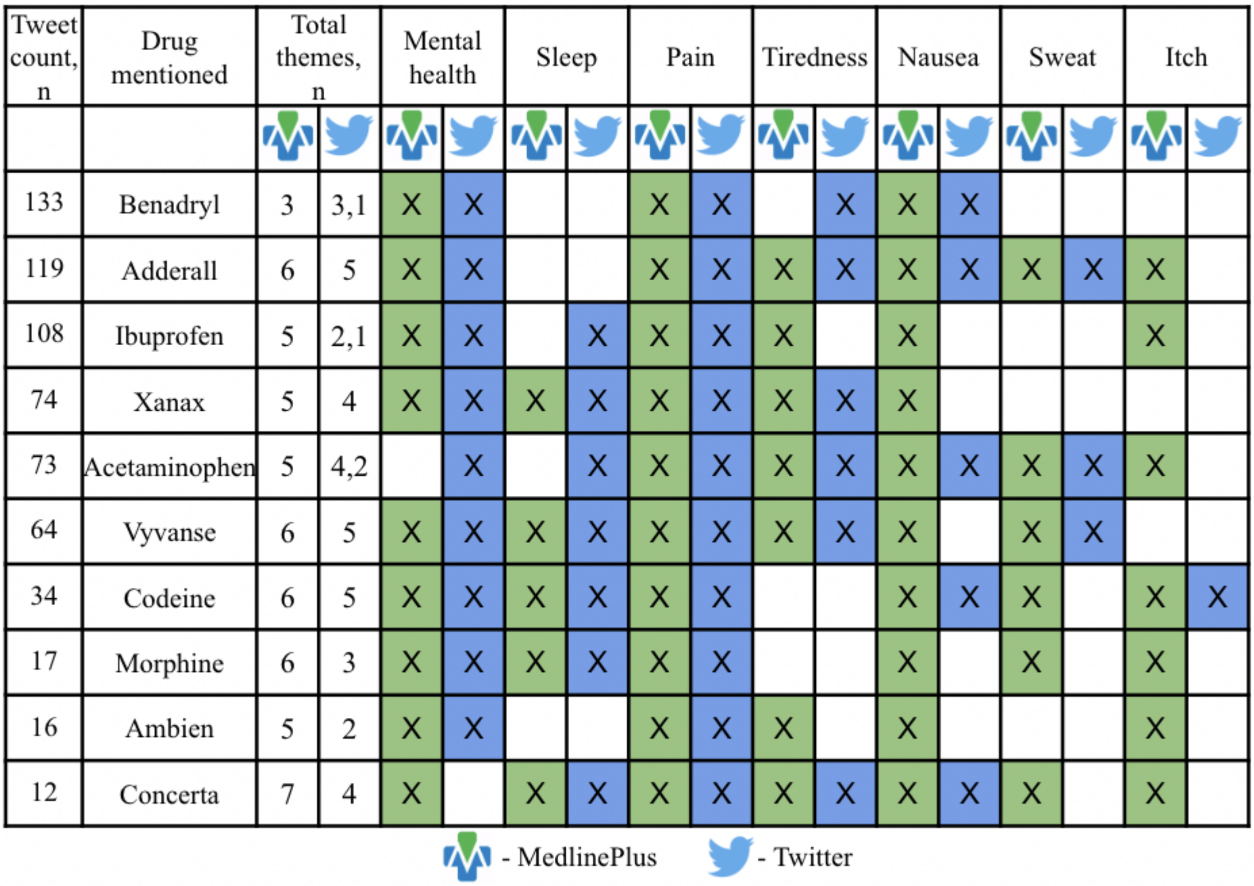
Table 4 summarizes the top 7 ADE themes, shows the frequency of tweets for each theme, and provides paraphrased examples. The major thematic areas include mental health–related ADEs and sleep-related ADEs. Tweeters also frequently shared their experience of pain-related ADEs. The remaining themes were discussed less frequently in our data set.

Each row in Figure 2 represents a drug. In the 108 tweets for “ibuprofen,” mentions of 3 drugs are grouped together: Advil (n=40), ibuprofen (n=36), and Motrin (n=32); the 73 tweets for “acetaminophen” group together mentions of 2 drugs: Tylenol (n=71) and acetaminophen (n=2). The third column indicates the number of themes with known adverse reactions from MedlinePlus as well as the number of themes with ADE mentions on Twitter. The fourth column can have 2 different numbers, separated by a comma: the first is the number of themes that overlap with known adverse reactions, while the second, if present, is the number of themes that do not overlap with known adverse reactions. For example, Benadryl has 3 themes with known adverse reactions, all of which were captured in tweets, and 1 theme that was not listed in MedlinePlus but only mentioned in tweets. For Adderall, 6 themes contained known adverse reactions; 5 of these had related tweets.

Table 4. Top 7 adverse drug event themes with frequencies and examples (N=941).

Adverse drug event theme	Tweets, n (%)	Paraphrased examples
Mental health	204 (21.7)	Feeling emotionally unstable, depressed, or high
Sleep	201 (21.4)	Feeling sleepy, being knocked out by a drug, wanting to sleep, not being able to sleep, being able to stay awake at night
Pain	151 (16)	Experiencing other pains or aches, such as headache or stomachache
Tiredness	27 (2.9)	Feeling extremely tired
Nausea	21 (2.2)	Feeling nausea or a need to vomit
Sweating	20 (2.1)	Experiencing sweating
Itchiness	16 (1.7)	Feeling itchy

Figure 2. The top 10 drugs with known adverse reactions found in MedlinePlus versus adverse drug events found in tweets. X: drug with at least one known adverse reaction or adverse drug event related to a particular theme. Values before commas indicate themes mentioned in tweets as well as MedlinePlus, while values after commas indicate values indicated only in tweets.



Tweet count, n	Drug mentioned	Total themes, n		Mental health		Sleep		Pain		Tiredness		Nausea		Sweat		Itch	
		🟩	🐦	🟩	🐦	🟩	🐦	🟩	🐦	🟩	🐦	🟩	🐦	🟩	🐦	🟩	🐦
133	Benadryl	3	3,1	X	X			X	X		X	X	X				
119	Adderall	6	5	X	X			X	X	X	X	X	X	X	X	X	
108	Ibuprofen	5	2,1	X	X		X	X	X	X		X				X	
74	Xanax	5	4	X	X	X	X	X	X	X	X	X					
73	Acetaminophen	5	4,2		X		X	X	X	X	X	X	X	X	X	X	
64	Vyvanse	6	5	X	X	X	X	X	X	X	X	X		X	X		
34	Codeine	6	5	X	X	X	X	X	X			X	X	X		X	X
17	Morphine	6	3	X	X	X	X	X	X			X		X		X	
16	Ambien	5	2	X	X			X	X	X		X				X	
12	Concerta	7	4	X		X	X	X	X	X	X	X	X	X		X	

🟩 - MedlinePlus 🐦 - Twitter

Discussion

Principal Results of the NLP System

By running our models on the validation set shown in Table 2, we confirmed that the performance of CLAPA was almost the same as that of previously published models, with an F1 score of 0.5998 [26]. The performance of BERT was also similar to that of BERT-based models reported in an overview of the SMM4H 2019 shared task [13]. This confirmation ensured that our results did not include any noise due to unexpected performance of the models. Our evaluation results demonstrate that baCLAPA outperformed CLAPA on both evaluation sets. However, we made two observations: (1) there was a significant gap between the performance of each model on the SMM4H evaluation set, and (2) there was a significant decrease in performance on both evaluation sets compared to the validation set. More detailed discussion of these observations follows.

First, while the gap in F1 scores on the WEB-RADR evaluation set seems similar to the gap with the validation set, there was a significant gap between the F1 score of the 2 models on the SMM4H evaluation set. CLAPA's F1 score was 0.44, while baCLAPA achieved an F1 score of 0.51. We believe this is because CLAPA utilizes a training set to enhance medical concept representation. That is, the model heavily relies on the training set, which may result in overfitting. BERT might help diminish this problem because of its generalizability as a language model, that is, it computes word embeddings based on the full context of a sentence given a large text corpus. Thus, incorporating BERT would help CLAPA not just to learn the context better but also not overfit the model on the training set. We plan to investigate this observation further once gold labels are released by the data set developers, or if we observe a similar result in other data sets.

Second, the performance of both CLAPA and baCLAPA was significantly lower on the evaluation sets than on the validation set. This may be partly explained by the number of tweets in which none of the drugs from the drug list were found. In addition to the total number of tweets for each data set, Table 1 also shows the number of unique drugs found using our drug list, and the number of tweets that did not have any drug names from the list. Our drug list contained 4888 drug names, including generic names and brand names. Since the collection was initially built through MedlinePlus and expanded through the training set, it covers almost all tweets, with the exception of 31 tweets that contained very specific typos, such as “vioxe” or “viox” instead of “vioxx” (the correct spelling), and were excluded from the data set. In the training set, a total of 1020 unique drugs were identified from our list. However, the number of unique drugs was lower in the evaluation sets: 688 in the SMM4H evaluation set and 685 in the WEB-RADR evaluation set. The number of drugs found in a new data set is expected to be relatively low, because the list is incomplete: our list did not cover all drug names or common typos. However, the number of tweets that did not contain any drug words was a significant portion of the WEB-RADR evaluation set. The 25,646 tweets affected by this in the WEB-RADR data set would have been considered as non–drug-relevant tweets by the models, whereas the 129 tweets in the SMM4H evaluation set would have been considered as non-relevant tweets. When the models are uncertain whether or not a tweet is drug-relevant, which depends heavily on a drug list, the prediction task may suffer.

To summarize, baCLAPA achieved an F1 score of 0.51 on the SMM4H evaluation set and 0.394 on the WEB-RADR evaluation set. BaCLAPA outperformed CLAPA on both evaluation sets, which illustrates the effectiveness of the method. We observed a gap between the performance of the models on the SMM4H evaluation set and an overall decrease in evaluation performance. This trend seems to be valid for many current ADE systems, since the average evaluation score was significantly lower than the validation score in past SMM4H tasks [11,13,25]. This shows that although the suggested improvements in baCLAPA appear to perform well, they may not generalize as well on unseen data sets for the ADE classification tasks, as also observed by Gattepaille et al [14]. Further research is necessary to evaluate the generalizability of neural network–based models on the ADE classification task.

Principal Results for the Content Analysis of ADE Tweets

Our content analysis presents the ADE themes and a comparison between the known adverse reactions and ADE mentions to answer two questions: (1) what kinds of ADE are mentioned on Twitter? and (2) of ADEs mentioned for each known drug on Twitter, how many are also known adverse reactions listed on MedlinePlus?

Question 1: What Kinds of ADE Are Mentioned on Twitter?

Table 4 illustrates that there were 7 primary ADE-related themes within the 941 tweets available for the qualitative content analysis. Other themes that were found but not included in the table because of their infrequency include those related to jitters, body weight, skin, sexual health, digestion, and seizure. Similarly, other ADE themes, such as those related to vision and breathing, may also be found for specific drugs.

Question 2: Of ADEs Mentioned for Each Known Drug on Twitter, How Many Are Also Known Adverse Reactions Listed on MedlinePlus?

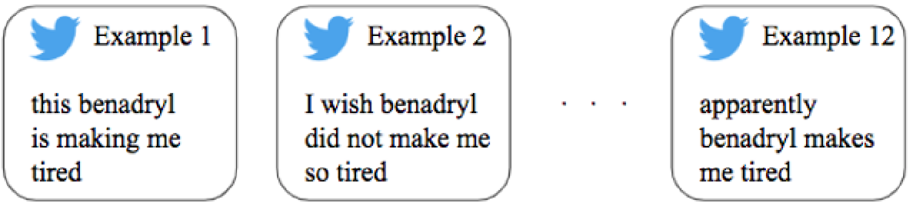
Figure 2 shows the top 10 drugs and their associated ADE themes found in MedlinePlus compared to mentions on Twitter. Based on these 10 drugs, the Twitter data covered an average of 69.6% of the known adverse reactions on MedlinePlus. When we set the number of tweets to be 30 or more, the average coverage increased to 78.4%. Based on the tweet counts, we conclude that Twitter data can adequately identify known adverse reactions for most drugs. However, this depends on the number of tweets extracted for each drug. For example, when we extracted fewer than 20 tweets, the model identified less than half of the known adverse reaction themes. Setting an appropriate minimum threshold may be a critical step for such exploratory analyses.

Finally, social media analysis can help highlight potentially new adverse reactions from drugs. For example, Figure 3 shows tweets that pair the Benadryl and tiredness-related ADE themes, which has not been reported as a known adverse reaction in MedlinePlus, but is expressed in these tweets. Looking at specific examples of tweets can help further elaborate on these as-yet-unreported pairings. These examples could be directly updated with a reporting system such as FAERS.

In-depth analysis of social media to detect ADE mentions could also show how laypersons report ADEs in their own language. Learning such expressions could help fill a vocabulary gap between patients and health professionals and enable better communication when prescribing a drug and analyzing patient-reported outcomes. Lastly, we observe that Figure 2 presents 12 new possible pairings. These occurrences could signal the need for a potential testing of ADE hypotheses derived from in-depth social media analysis.

Through this study, we have found that Twitter covers a sufficiently wide range of ADEs given a set of drugs and also covers known adverse reactions relatively well, especially when a sufficient number of drug-related tweets are analyzed. Therefore, this study demonstrates that social media can be a reliable data source for collecting ADE mentions.

Figure 3. Paraphrased examples of adverse drug event themes related to Benadryl and tiredness.



Example 1

this benadryl is making me tired

Example 2

I wish benadryl did not make me so tired

. . .

Example 12

apparently benadryl makes me tired

Limitations

Our NLP system and study have some limitations. First, we did not focus on any causality relationships between a drug and an ADE. Although our qualitative analysis may signal the need for hypothesis testing, validating such claims of causality is beyond the scope of this work. Second, one of the long-term goals for this line of research is to build an automated system to collect actual ADE mentions from social media. While the classification model helps filter out large-scale data, it does not provide the actual extent of such mentions, which prevents obtaining further information, such as pairs of drug–ADE mentions, from the filtered data. To extract such mentions from tweets, we plan to work on developing an ADE extraction model. Lastly, our system cannot yet be fully deployed in practice. Our experimental results suggest that further research and development is necessary to fine-tune the models for better generalizability.

The approach presented in this paper serves as an analytical tool to identify potential adverse events in data from Twitter and other social media. It highlights both a way to validate some of the known ADEs and uncover additional potential ADEs. However, it does not fully demonstrate the relevance of social media as an independent and comprehensive source for identifying ADEs. Since there are no “gold standard” labeled data sets on possible adverse events related to a particular drug, none of the existing approaches present a comprehensive solution to the challenge of identifying all known and unknown adverse events related to a particular drug.

Further, our analysis is also biased because of the demographics of Twitter users and the differential coverage of drugs and their adverse events on Twitter. Twitter users are typically younger and more technically savvy [32]. This is especially relevant for studies of population health, since individuals from a lower socioeconomic status, underrepresented minorities, older adults, and individuals with chronic conditions are less likely to tweet [29]. Similarly, there could have been bias in the coverage of drugs and their adverse events. Although the analysis was ultimately based on 28.8 million tweets, the data were collected for the purpose of a community-based study from the Detroit metropolitan area. Tweeters in this area may discuss a particular drug more or less often than those in other communities or regions. Thus, the representation of drug usage in our data may be different from the representation of tweets collected, regardless of geographic location, making our analysis unrepresentative of overall drug usage and the types of drugs mentioned on Twitter. Rather, our analysis is limited to a certain set of drugs and their ADE mentions. However, the methodology and analysis could be repeated for other drugs.

Conclusion

In this paper, we present a neural network–based model, baCLAPA, which incorporates a representation generated by BERT with one by CLAPA. Our experimental results demonstrate that baCLAPA outperformed CLAPA. The weak performance on unseen data signals that there is still room for improvement for the ADE classification task. Our validation study suggests that Twitter data not only include a sufficiently wide range of ADE mentions but also cover most known adverse reactions for drugs found in the relevant tweets.

Even though our work does not show any causal relationships between the drugs and ADEs mentioned, it provides possible directions to advance ADE-related work. For example, our qualitative analysis of ADE tweets could provide a basis for potential analyses and applications. It also implies that social media data can provide meaningful measurements once we have an all-purpose NLP system for collecting ADE mentions, including not just classification but also extraction. Our work demonstrates that social media can be a reliable data source for this purpose. While recent studies have developed and improved such systems, our work suggests that ADE classification systems need further research to study their robustness and reliability.

Conflicts of Interest

None declared.

References

1. Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. Lancet 2000 Oct 07;356(9237):1255-1259. [doi: 10.1016/S0140-6736(00)02799-9] [Medline: 11072960]

2. International drug monitoring: the role of national centres, report of a WHO meeting held in Geneva from 20 to 25 September 1971. World Health Organization. 1972. URL: https://apps.who.int/iris/handle/10665/40968 [accessed 2022-09-12]

3. Asscher AW, Parr GD, Whitmarsh VB. Towards the safer use of medicines. BMJ 1995 Oct 14;311(7011):1003-1006 [FREE Full text] [doi: 10.1136/bmj.311.7011.1003] [Medline: 7580589]

4. Pirmohamed M, Breckenridge AM, Kitteringham NR, Park BK. Adverse drug reactions. BMJ 1998 Apr 25;316(7140):1295-1298 [FREE Full text] [doi: 10.1136/bmj.316.7140.1295] [Medline: 9554902]
5. Karch FE, Lasagna L. Adverse drug reactions. A critical review. JAMA 1975 Dec 22;234(12):1236-1241. [Medline: 1242749]
6. Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. JAMA 1998 Apr 15;279(15):1200-1205. [doi: 10.1001/jama.279.15.1200] [Medline: 9555760]
7. Ahmad SR. Adverse drug event monitoring at the Food and Drug Administration. J Gen Intern Med 2003 Jan;18(1):57-60 [FREE Full text] [doi: 10.1046/j.1525-1497.2003.20130.x] [Medline: 12534765]
8. FDA Adverse Event Reporting System (FAERS) Public Dashboard. US Food and Drug Administration. URL: https://www.fda.gov/drugs/questions-and-answers-fdas-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-public-dashboard [accessed 2022-03-20]
9. Jagannatha A, Liu F, Liu W, Yu H. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). Drug Saf 2019 Jan;42(1):99-111 [FREE Full text] [doi: 10.1007/s40264-018-0762-z] [Medline: 30649735]
10. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inform Assoc 2020 Jan 01;27(1):3-12 [FREE Full text] [doi: 10.1093/jamia/ocz166] [Medline: 31584655]
11. Klein A, Alimova I, Flores I, Magge A, Miftahutdinov Z, Minard A, et al. Overview of the Fifth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at COLING 2020. In: Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task. 2020 Presented at: The 28th International Conference on Computational Linguistics; Dec 8, 2020; Barcelona, Spain (Online).
12. Weissenbacher D, Sarker A, Paul M, Gonzalez-Hernandez G. Overview of the Third Social Media Mining for Health (SMM4H) Shared Tasks at EMNLP 2018. In: Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop and Shared Task. 2018 Presented at: 2018 Conference on Empirical Methods in Natural Language Processing; Oct 31, 2018; Brussels, Belgium. [doi: 10.18653/v1/w18-5904]
13. Weissenbacher D, Sarker A, Magge A, Daughton A, O'Connor K, Paul M, et al. Overview of the Fourth Social Media Mining for Health (SMM4H) Shared Task at ACL 2019. In: Proceedings of the 2019 ACL Workshop SMM4H: The 4th Social Media Mining for Health Applications Workshop & Shared Task. 2019 Presented at: The 57th Annual Meeting of the Association for Computational Linguistics; Jul 28, 2019; Florence, Italy. [doi: 10.18653/v1/W19-3203]
14. Gattepaille LM, Hedfors Vidlin S, Bergvall T, Pierce CE, Ellenius J. Prospective evaluation of adverse event recognition systems in Twitter: results from the Web-RADR project. Drug Saf 2020 Aug;43(8):797-808 [FREE Full text] [doi: 10.1007/s40264-020-00942-3] [Medline: 32410156]
15. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics; Jun 2, 2019; Minneapolis, MN. [doi: 10.18653/v1/n18-2]
16. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. OpenAI Assets. URL: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf [accessed 2022-09-12]
17. Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, et al. Utilizing social media data for pharmacovigilance: A review. J Biomed Inform 2015 Apr;54:202-212 [FREE Full text] [doi: 10.1016/j.jbi.2015.02.004] [Medline: 25720841]
18. Sultana J, Cutroneo P, Trifirò G. Clinical and economic burden of adverse drug reactions. J Pharmacol Pharmacother 2013 Dec;4(Suppl 1):S73-S77 [FREE Full text] [doi: 10.4103/0976-500X.120957] [Medline: 24347988]
19. Rossi AC, Knapp DE, Anello C, O'Neill RT, Graham CF, Mendelis PS, et al. Discovery of adverse drug reactions. A comparison of selected phase IV studies with spontaneous reporting methods. JAMA 1983;249(16):2226-2228. [doi: 10.1001/jama.249.16.2226] [Medline: 6834622]
20. Karimi S, Metke-Jimenez A, Kemp M, Wang C. Cadec: A corpus of adverse drug event annotations. J Biomed Inform 2015 Jun;55:73-81 [FREE Full text] [doi: 10.1016/j.jbi.2015.03.010] [Medline: 25817970]
21. Zolnoori M, Fung KW, Patrick TB, Fontelo P, Kharrazi H, Faiola A, et al. The PsyTAR dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications. Data Brief 2019 Jun;24:103838 [FREE Full text] [doi: 10.1016/j.dib.2019.103838] [Medline: 31065579]
22. Dietrich J, Gattepaille LM, Grum BA, Jiri L, Lerch M, Sartori D, et al. Adverse Events in Twitter-Development of a Benchmark Reference Dataset: Results from IMI WEB-RADR. Drug Saf 2020 May;43(5):467-478 [FREE Full text] [doi: 10.1007/s40264-020-00912-9] [Medline: 31997289]
23. Sarker A, Nikfarjam A, Gonzalez-Hernandez G. Social Media Mining Shared Task Workshop. In: Biocomputing 2016: Proceedings of the Pacific Symposium. 2016 Presented at: The Pacific Symposium on Biocomputing; Jan 4-7, 2016; Waimea, HI p. 581-592. [doi: 10.1142/9789814749411_0054]
24. Sarker A, Gonzalez-Hernandez G. Overview of the Second Social Media Mining for Health (SMM4H) Shared Tasks at AMIA 2017. In: Proceedings of the Second Social Media Mining for Health Research and Applications Workshop. 2017

Presented at: American Medical Informatics Association Annual Symposium; Nov 6, 2017; Washington, DC. [doi: 10.18653/v1/w18-5904]

25. Magge A, Klein A, Miranda-Escalada A, Al-garadi M, Alimova I, Miftahutdinov Z, et al. Overview of the Sixth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at NAACL 2021. In: Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task. 2021 Jun 10 Presented at: 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics; Jun 6, 2021; Mexico City, Mexico (Online). [doi: 10.18653/v1/2021.smm4h-1.4]
26. Zhao X, Yu D, Vydiswaran V. Identifying adverse drug events mentions in tweets using attentive, collocated, and aggregated medical representation. In: Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task. 2019 Presented at: The 57th Annual Meeting of the Association for Computational Linguistics; Jul 28, 2019; Florence, Italy. [doi: 10.18653/v1/w19-3209]
27. MedlinePlus - Health Information from the National Library of Medicine. MedlinePlus. URL: https://medlineplus.gov/ [accessed 2022-03-20]
28. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. 2017 Presented at: The 15th Conference of the European Chapter of the Association for Computational Linguistics; Apr 3, 2017; Valencia, Spain. [doi: 10.18653/v1/e17-2068]
29. Vydiswaran V, Romero D, Zhao X, Yu D, Gomez-Lopez I, Lu J, et al. "Bacon bacon bacon": food-related tweets and sentiment in metro Detroit. In: Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018. 2018 Presented at: The 12th International AAAI Conference on Web and Social Media; Jun 25, 2018; Stanford, California.
30. Mayring P. Qualitative Content Analysis. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research. URL: https://www.qualitative-research.net/index.php/fqs/article/view/1089/2385 [accessed 2022-09-12]
31. Vydiswaran V, Yu D, Zhao X, Carr E, Martindale J, Xiao J, et al. Identifying medication abuse and adverse effects from tweets: University of Michigan at# SMM4H 2020. In: Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task. 2020. pp. 90?94. 2020 Presented at: The 28th International Conference on Computational Linguistics; Dec 8, 2020; Barcelona, Spain (Online). [doi: 10.18653/v1/w19-3209]
32. Wojcik S, Hughes A. Sizing Up Twitter Users. Pew Research Center. URL: https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/ [accessed 2022-09-12]

Abbreviations

ADE: adverse drug event
baCLAPA: bidirectional encoder representations from transformers–assisted collocated long short-term memory with attentive pooling and aggregated representation
BERT: bidirectional encoder representations from transformers
CLAPA: collocated long short-term memory with attentive pooling and aggregated representation
LSTM: long short-term memory
NLP: natural language processing
SMM4H: Social Media Mining For Health
SVM: support vector machine
WEB-RADR: web-recognizing adverse drug reactions
WHO: World Health Organization

Edited by C Lovis; submitted 21.03.22; peer-reviewed by R Sun; comments to author 12.06.22; revised version received 13.08.22; accepted 07.09.22; published 28.09.22.

Please cite as:
Yu D, Vydiswaran VGV
An Assessment of Mentions of Adverse Drug Events on Social Media With Natural Language Processing: Model Development and Analysis
JMIR Med Inform 2022;10(9):e38140
URL: https://medinform.jmir.org/2022/9/e38140
doi:10.2196/38140
PMID:36170004

Original Paper

Evaluating the Impact of a Point-of-Care Cardiometabolic Clinical Decision Support Tool on Clinical Efficiency Using Electronic Health Record Audit Log Data: Algorithm Development and Validation

Xiaowei Yan[1], MS, PhD; Hannah Husby[1], MPH; Satish Mudiganti[1], MS; Madina Gbotoe[1], MS; Jake Delatorre-Reimer[2], BS; Kevin Knobel[3], MD; Andrew Hudnut[4], MD; J B Jones[1], MBA, PhD

[1]Center for Health Systems Research, Sutter Health, Walnut Creek, CA, United States

[2]Department of Clinical Informatics, NorthBay Healthcare, Fairfield, CA, United States

[3]Sutter Gould Medical Foundation, Sutter Health, Modesto, CA, United States

[4]Sutter Medical Group, Sutter Health, Sacramento, CA, United States

Corresponding Author:
Xiaowei Yan, MS, PhD
Center for Health Systems Research
Sutter Health
2121 N California Blvd
Suite 310
Walnut Creek, CA, 94596
United States
Phone: 1 925 287 4025
Email: YanSX@sutterhealth.org

Abstract

Background: Electronic health record (EHR) systems are becoming increasingly complicated, leading to concerns about rising physician burnout, particularly for primary care physicians (PCPs). Managing the most common cardiometabolic chronic conditions by PCPs during a limited clinical time with a patient is challenging.

Objective: This study aimed to evaluate a Cardiometabolic Sutter Health Advanced Reengineered Encounter (CM-SHARE), a web-based application to visualize key EHR data, on the EHR use efficiency.

Methods: We developed algorithms to identify key clinic workflow measures (eg, total encounter time, total physician time in the examination room, and physician EHR time in the examination room) using audit data, and we validated and calibrated the measures with time-motion data. We used a pre-post parallel design to identify propensity score–matched CM-SHARE users (cases), nonusers (controls), and nested-matched patients. Cardiometabolic encounters from matched case and control patients were used for the workflow evaluation. Outcome measures were compared between the cases and controls. We applied this approach separately to both the CM-SHARE pilot and spread phases.

Results: Time-motion observation was conducted on 101 primary care encounters for 9 PCPs in 3 clinics. There was little difference (<0.8 minutes) between the audit data–derived workflow measures and the time-motion observation. Two key unobservable times from audit data, physician entry into and exiting the examination room, were imputed based on time-motion studies. CM-SHARE was launched with 6 pilot PCPs in April 2016. During the prestudy period (April 1, 2015, to April 1, 2016), 870 control patients with 2845 encounters were matched with 870 case patients and encounters, and 727 case patients with 852 encounters were matched with 727 control patients and 3754 encounters in the poststudy period (June 1, 2016, to June 30, 2017). Total encounter time was slightly shorter (mean −2.7, SD 1.4 minutes, 95% CI −4.7 to −0.9; mean –1.6, SD 1.1 minutes, 95% CI −3.2 to −0.1) for cases than controls for both periods. CM-SHARE saves physicians approximately 2 minutes EHR time in the examination room (mean −2.0, SD 1.3, 95% CI −3.4 to −0.9) compared with prestudy period and poststudy period controls (mean −1.9, SD 0.9, 95% CI −3.8 to −0.5). In the spread phase, 48 CM-SHARE spread PCPs were matched with 84 control PCPs and 1272 cases with 3412 control patients, having 1119 and 4240 encounters, respectively. A significant reduction in total encounter time for the CM-SHARE group was observed for short appointments (≤20 minutes; 5.3-minute reduction on average) only. Total physician EHR time was significantly reduced for both longer and shorter appointments (17%-33% reductions).

Conclusions: Combining EHR audit log files and clinical information, our approach offers an innovative and scalable method and new measures that can be used to evaluate clinical EHR efficiency of digital tools used in clinical settings.

(JMIR Med Inform 2022;10(9):e38385) doi:10.2196/38385

KEYWORDS

digital health; electronic health record; EHR audit logs; workflow efficiency; cardiometabolic conditions

Introduction

Background

Approximately 34% of the population in the United States aged ≥18 years has a cardiometabolic condition (ie, diabetes mellitus [DM], hypertension, and high cholesterol), which are among the most common and costly health problems [1]. Managing these chronic conditions is an important area of focus for primary care physicians (PCPs) in the United States. However, the effective management of these chronic conditions can be challenging for patients and PCPs.

Patients spend <1% of their time with their PCPs and the rest on their own, attempting to adopt the care plan prescribed by their PCP into their daily lives [2]. During their limited time at the point of care with patients, PCPs rarely have enough time to review all the critical but scattered data in an electronic health record (EHR); given that, on average, PCPs only have 15 minutes with patients for a face-to-face visit [3]. Health care providers have expressed dissatisfaction with EHR systems [4-6] used to manage patients, which have generally been poorly designed for facilitating care delivery. Increasing evidence indicates that an EHR imposes an additional burden on physicians [7-10]. In particular, PCPs reported having the highest burnout associated with EHR use [11]. The causes of physician burnout are multifactorial, including the increasing complexity and cognitive burden of using the EHR and decreased face-to-face time with patients [12,13]. Moreover, longer EHR use time is negatively associated with patient satisfaction [14], especially with increased daytime EHR use, potentially occurring in the examination room, implying that physicians have less time to communicate with patients, which may adversely affect patient-physician relationships [15-17]. Therefore, technology or interventions that aim to reduce EHR time in the examination room can potentially improve health delivery quality from both physician and patient perspectives. Digital health solutions, including clinical decision support tools, hold the promise of helping physicians improve patient clinical outcomes or quality of care [18-21]; however, it is less clear whether these solutions have an impact on clinical EHR efficiency [3].

A well-designed and integrated digital tool can be sufficiently seamless so that the user feels unencumbered by the effort to open the additional platform and perceive the EHR and digital tool as one system [21]. Therefore, using principles of user-centered design, we developed Cardiometabolic Sutter Health Advanced Reengineered Encounter (CM-SHARE), a web-based application designed to simplify care delivery for patients with cardiometabolic conditions (DM, hypertension, and dyslipidemia) [22]. CM-SHARE extracts essential health data elements from the EHR in real time at the point of care and displays them in novel ways for both physicians and patients. The main features of CM-SHARE include a snapshot view, graphs, medication dispensing history, and risk calculators, where the snapshot view provides an intuitive overview of patient-specific data gathered from different areas that are critical to review for patients with cardiometabolic conditions, whereas graphs and medication dispensing history use graphic views of longitudinal laboratories, vitals, medication dispense and adherence history, and risk calculators that allow physicians to change the values of different risk factors to help educate patients on how changes that modify different risk factors can affect cardiovascular risk. The primary design intent of CM-SHARE is to reduce the time physicians spend “hunting and clicking” for information in the EHR. A previous study showed that CM-SHARE, a voluntary-use digital health solution, was successfully integrated into a real-world primary care setting with high adoption and consistent use in caring for patients with cardiometabolic conditions [12]. In this study, following principles of the digital health technology development and deployment [23,24], we first tested CM-SHARE among a small group of pilot users (ie, PCPs) before spreading to a much broader group of PCP end users. Although there is an increasing availability of digital tools created by health care or high-technology companies, little is known about whether the impact of these digital tools on pilot users is sustained when the technology is spread to a much broader group of users [25,26]. Therefore, we assess CM-SHARE’s impact on physician workflow and whether CM-SHARE achieves its intended goal of improving provider EHR and encounter efficiency in the pilot users and then assess whether similar impacts are sustained after disseminating to a broader group of PCPs.

Clinical workflow and EHR have typically been measured using time-motion studies, which are costly and not scalable [27-30]. In recent years, EHR audit log data have been increasingly used as an alternative approach to estimate the workflow and time used in the EHR, and this approach is scalable and reproducible [31-34]. The audit logs record who (user), when (time), where (location), and what EHR function has been used and are routinely collected in health care systems [33-36]. EHR audit data have been used previously in emergency departments [36] and specialty settings [37,38] to assess the efficiency and have been validated as a resource for analyzing workflows [39,40].

Objectives

The objective of this study was to thoroughly evaluate the impact of CM-SHARE on physicians’ clinical workflow in primary care encounters. We conducted data collection and analyses in the following order: time-motion observations to develop and validate audit data–derived clinical workflow algorithms, extrapolated for non-EHR clinical work (eg, patient examination and conversation with patients), followed by the application of these algorithms to clinical workflow analysis among CM-SHARE pilot users, and, finally, clinical workflow analysis among a broader group of "spread users" to whom CM-SHARE was made available. Our analysis focused on assessing the impact of CM-SHARE on physician workflows and EHR use. We hypothesized that CM-SHARE's visual display of health data would (1) reduce the overall encounter time, EHR time, and physician EHR time in the examination room; (2) result in differential reductions in EHR time based on the scheduled visit length and the encounter primary diagnosis; and (3) lead to observable reductions in pilot users that are sustained when CM-SHARE is spread to a broader group of PCPs.

Methods

Overview

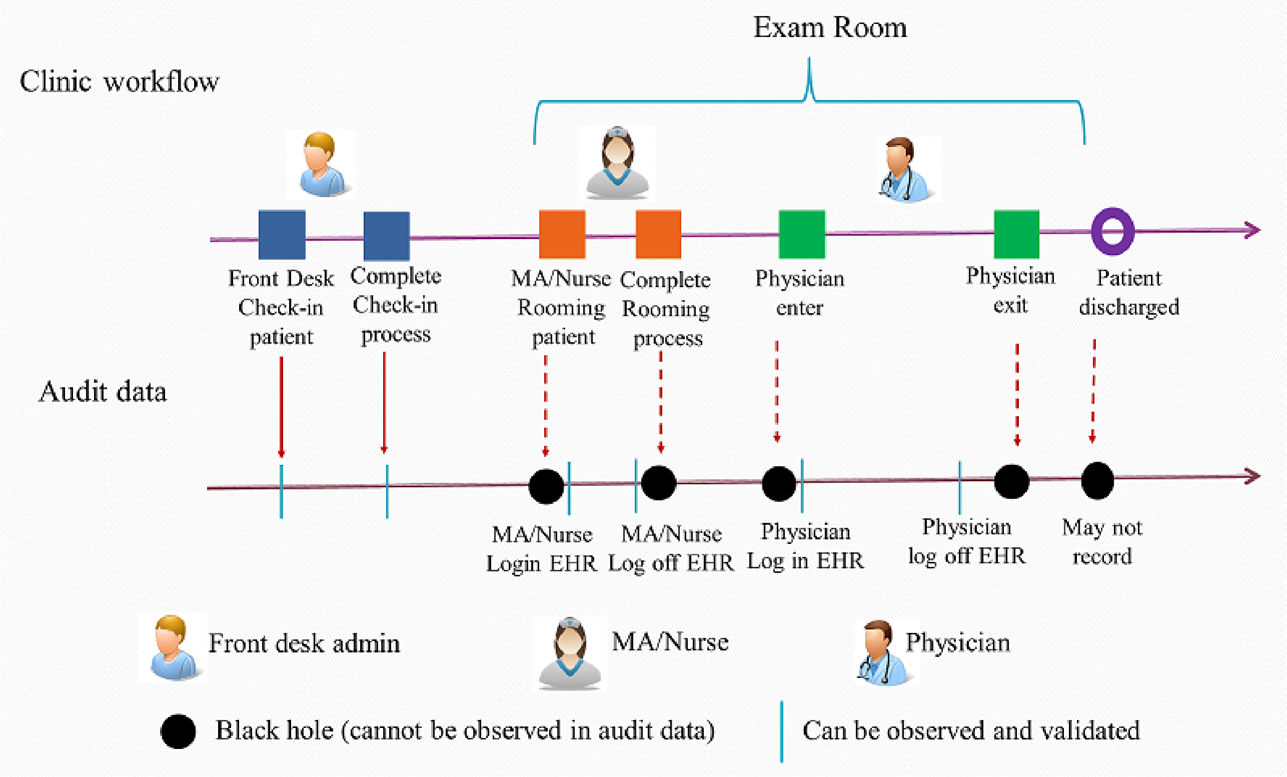
There were 3 main components to this study. First, we conducted a time-motion study to collect the main workflow time and duration of primary care encounters observed on 3 randomly selected workdays for 9 PCPs. We used these time-motion data to validate the workflow steps and to refine algorithms that capture the workflow based on the EHR audit log data. Second, we evaluated CM-SHARE using a pre-post parallel design, where cases were defined as encounters in which CM-SHARE was launched. Finally, we estimated the impact of CM-SHARE on the audit data–derived workflow outcomes in the matched cohort. The second and third study components were first conducted on the original set of pilot CM-SHARE users involved in the application development. We then repeated these analyses in the spread phase of the study, in which CM-SHARE was implemented and used by a new, broader group of PCPs who were not involved in any aspect of the original design and pilot phase.

Validation of Audit Data Workflow

We developed algorithms to identify key steps (eg, check-in), tasks performed (eg, EHR use), locations of tasks (eg, examination room), and roles (eg, nurses and physicians) involved in clinical ambulatory workflows using time-stamped audit log files from the EHR and performed a time-motion observation of 101 encounters from 9 PCPs at 3 different primary care clinics. We compared the time-motion–observed times of key workflow points in a patient encounter with workflow measures from the EHR audit log files. The time-motion data-tracking form can be found in Multimedia Appendix 1.

The critical clinical workflow points included check-in, rooming time (when a medical assistant [MA] or nurse takes a patient to the examination room), nurse exiting the examination room, physician entry into the examination room, physician's total EHR time in the examination room, physician exiting the examination room, and patient check-out time. We used the term "black holes" to identify essential steps in a clinical workflow, which cannot be observed in the audit data as they do not involve interaction with the EHR but can be observed and recorded in time and motion observations. These black holes include steps such as the physician entering the examination room and the physician exiting the examination room (Figure 1). These are important events to account for as they affect the overall time of the encounter. For example, a physician may enter the examination room and spend several seconds to minutes conversing with the patient before logging into the computer.

Figure 1. Illustration of outpatient visit workflow and corresponding audit data. EHR: electronic health record; MA: medical assistant.



Measures that can be reliably observed in audit data include the physician's start time, accessing the EHR in the examination room, and the time of the last EHR access in the examination room. We compared these key workflow time points recorded in the audit log data with time-motion observations and estimated the distribution of each black hole. We further conducted an ANOVA for each black hole in relation to the EHR user (ie, PCP) and patient. This assessment allowed us to understand the variation among PCPs and among patients. We then imputed the time duration of each black hole based on the distribution, and a random number was generated based on the empirical distribution and imputed for the black hole.

The time duration of each key workflow point was calculated based on the imputed audit data and compared again with the time points from the time-motion observation data. Discrepancies were further analyzed, and audit data for encounters with large discrepancies (discrepancy ≥2 SD) were manually reviewed. EHR activity and time points that were closest to the time recorded in time-motion observations were selected for specific users, and the algorithms used to capture each key workflow time point from the audit data were updated. After this initial validation process, we applied the algorithms to 7474 office encounters on a random working day across all Sutter Health primary care clinics to assess the generalizability. We identified key clinical workflow steps using this method, including check-in time, rooming time, nurse leave time, and physician time in the examination room on the computer. However, in this study, we focused only on the total encounter time (defined as the duration between patient check-in time to the time the patient exits the examination room), the physician's total time in the examination room, the physician's time spent in the EHR in the examination room, and the physician's total EHR click per encounter (Figure 1).

Study Population

The CM-SHARE application was developed and tested at Sutter Health, a large not-for-profit health system in Northern California that serves a racially and economically diverse patient population. CM-SHARE was implemented in 2 phases. The pilot study was initiated in April 2016 with 6 PCPs from 2 different primary care clinics. These PCPs were involved in the development of the application and had frequent communication (once a month in the first year after the initial launch) with the study team during the pilot-testing period. The spread of CM-SHARE started in October 2019 to a new group of PCPs at a large Sutter medical group that previously did not have access to CM-SHARE. In contrast, these new PCPs were lightly touched by the study team; for example, they were offered group (as opposed to one-on-one training as in the pilot phase) training using a CM-SHARE user manual or training provided by designated EHR trainers who were responsible for training clinicians on all EHR features and capabilities, not just CM-SHARE. Spread users were informed that CM-SHARE was specifically developed for patients with cardiometabolic conditions. When and for whom CM-SHARE was used was completely voluntary and up to a physician's discretion. Details of CM-SHARE features and adoption have been previously published [22].

At the patient level, primary care patients who had at least one cardiometabolic condition and had at least one visit with pilot PCPs or spread PCPs during the pilot test period or spread period were eligible for the study.

Outcomes

To assess efficiency, we measured (1) physician's total time in the examination room, (2) physician's EHR time in the examination room, (3) total encounter time (ie, from check-in to check-out), and (4) total physician clicks in the EHR for an encounter [41].

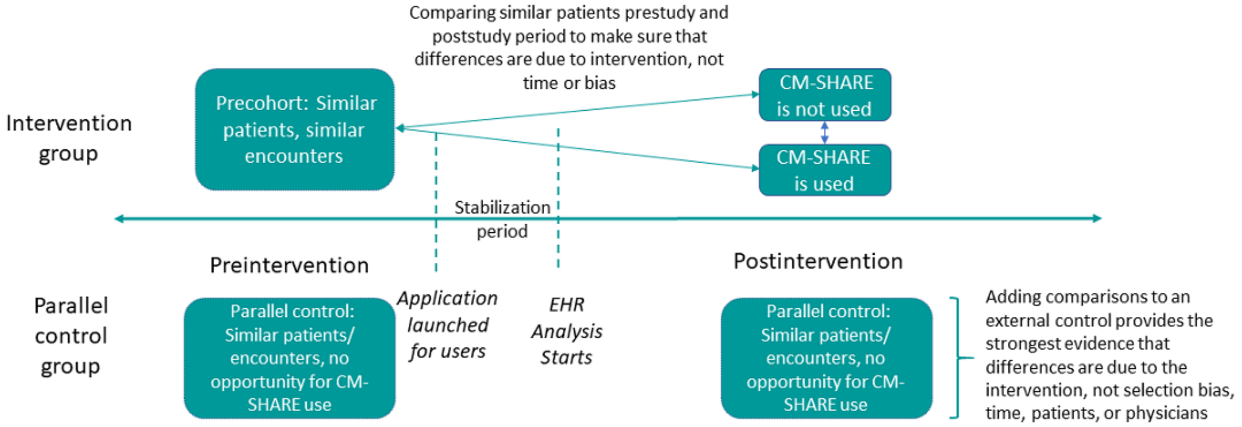
As illustrated in Figure 1, the physician's total time in the examination room was defined as the time between when the physician entered and exited the examination room, estimated based on audit data and imputed data for black holes described previously in the time-motion explanation. The time at which the physician entered the examination room was estimated by identifying the first time the physician logged into the EHR in the examination room, subtracting the imputed value for a black hole (ie, the time between the physician entering the examination room and logging into the EHR), and the physician exit time was estimated based on the EHR log-off time by the physician in the examination room and accounting for the imputed black hole (ie, the time between physician EHR log-off time and the exit examination room time). The physician's EHR time, estimated based on audit data, was defined as the cumulative time the physician spent in the EHR in the examination room, which is a subset of the physician's total time spent physically present in the examination room. The total number of physician EHR mouse clicks for each encounter, captured using audit data, was defined based on the cumulative number of EHR log entries (which record the EHR features accessed by a user) for a physician for a given encounter, including all previsit EHR activities (preparation for the visit), during the visit, and postvisit EHR activities (eg, clicking on a diagnosis, on medication tags in EHR, and clicking on a patient message). The total number of clicks reflects EHR information searching and access, and usually, more clicks implied more complex encounters as more patient medical information was reviewed.

Study Design

Overview

A pre-post matched design was applied, in which we defined the prestudy period as 12 months before the initial CM-SHARE launch in April, 2016, and the CM-SHARE stabilization time as at least 2 months after the initial launch time where the data were not used in the evaluation. The poststudy period was defined as 12 months after the stabilization period (Figure 2). Owing to the disruption of the COVID-19 pandemic and to allow for the stabilization of CM-SHARE use, we identified spread users who consistently used CM-SHARE starting in May 2020 (12 months after the initial spread) as stable CM-SHARE users. Data from October 15, 2018, to October 14, 2019, were used as the prestudy period to determine matched physicians. We compared efficiency measures for encounters occurring between May 2020 and December 2020.

Figure 2. Pre-post matched study design for CM-SHARE evaluation. CM-SHARE: Cardiometabolic Sutter Health Advanced Reengineered Encounter; EHR: electronic health record.


Comparing similar patients prestudy and poststudy period to make sure that differences are due to intervention, not time or bias
Intervention group
Precohort: Similar patients, similar encounters
CM-SHARE is not used
CM-SHARE is used
Stabilization period
Preintervention
Postintervention
Parallel control group
Parallel control: Similar patients/ encounters, no opportunity for CM-SHARE use
Application launched for users
EHR Analysis Starts
Parallel control: Similar patients/ encounters, no opportunity for CM-SHARE use
Adding comparisons to an external control provides the strongest evidence that differences are due to the intervention, not selection bias, time, patients, or physicians

Propensity Matching

There are two levels of endogeneity, which we accounted for in part via propensity matching. At the first level, physicians who had used CM-SHARE regularly (ie, used it at least once a week), denoted as case PCPs, may be systematically and unobservably different from those who did not regularly use CM-SHARE. To that end, case PCPs were matched to other Sutter-wide PCPs based on physician panel information, assuming that a physician's clinical practice pattern, including EHR use and information access, was affected by the composition of their patient panel. A propensity score–matching method was used to identify matched physicians, and a separate method was used to identify matched patients. At the physician level, a logistic regression model was developed, in which covariates included annual patient volume, mean number of appointments per day, average age of the patient (at time of encounter, categorized into <30, 30-49, 50-64, and ≥65 years), proportion of female patients, proportion of different ethnicities (Hispanic and non-Hispanic) in the patient panel, proportion of patients with each cardiometabolic condition (DM, dyslipidemia, and hypertension), practice type (family medicine and internal medicine), and proportion of patients with different levels of morbidity based on the Charlson Comorbidity Index (0, 1-2, and ≥3). The model outcome was the CM-SHARE pilot status (yes or no). Greedy matching was used in this study.

At the second level, patients for whom CM-SHARE was used may also be unobservably different from those for whom CM-SHARE was not used. Therefore, within the pilot physicians and matched control physicians, we performed one-to-one matching at the patient level to determine cases (for which CM-SHARE was used) and control patients (within the matched physicians). The case patients, for whom CM-SHARE was used, were matched to the control patients at the prestudy period and separately at the poststudy period.

Patients were grouped into pre- or poststudy period based on the first visit date in the prestudy period and poststudy period. In the prestudy period, eligible patients included those who had encounters with pilot physicians and matched control physicians. In the poststudy period, control patients were those who were eligible and had encounters with matched control physicians. Data from the prestudy period were used for prestudy period matching, and poststudy period data were used for poststudy period matching.

We developed a separate logistic regression model for patients where the covariates were individual patient-level features, including age, sex, race, ethnicity, percentage of 15-minute scheduled appointments, percentage of 30-minute scheduled appointments, percentage of the level of service (LOS) level 3 visits, percent of LOS level 4 visits, percentage of DM as primary encounter diagnosis, percentage of hypertension as primary encounter diagnosis, and percentage of dyslipidemia as primary encounter diagnosis. The outcome of this model was the CM-SHARE launch status in any patient encounter (yes or no). Greedy matching was also used.

We further defined eligible encounters to require that the encounter diagnoses contain at least one cardiometabolic diagnosis code, and encounters were made by matched case and control patients in the pre- and postperiod. EHR audit data for the eligible encounters were extracted and used in the analysis, and outcome measures were derived from those encounters.

Stratification

Furthermore, based on the feedback from pilot PCPs on the different use cases for CM-SHARE and guidelines on the content and services provided at ambulatory encounters, the complexity of a patient visit and documentation requirements were usually reflected by LOS or length of the appointment [42]. We expected EHR and CM-SHARE utility to differ according to LOS or length of the appointment. Therefore, we stratified the encounters according to the scheduled length of appointments. To compare the workflow measures, we first estimated the mean difference (and 95% CI) between cases and matched controls in the prestudy period and separately in the poststudy period. We used 2-tailed *t* tests to test the mean difference between prestudy period control versus case patients, poststudy period control versus case patients, and prestudy period controls and poststudy period controls. The 95% CI was estimated by fitting a mixed linear regression model that matched the case and control patients treated as the same cluster. The cluster was taken as the random effect in the model. In addition, the primary diagnosis usually reflects the main reason for the visit; thus, we conducted the abovementioned analysis stratified by the primary diagnosis and assessed whether the impact of CM-SHARE varies according to the reason for the visit.

Sutter Health uses an instance of Epic Systems software (Epic Systems Corporation) as its EHR [24]. All analyses were performed using SAS Enterprise Guide 7.1 (SAS Institute).

Ethics Approval

This study was reviewed and approved by the Sutter Health Institutional Review Board (IRB #: 833549).

Results

Validation Results

Among the 101 encounters, time-motion observations recorded check-in start and end times for all encounters, rooming start time, time the physician enters the examination room, and patient check-out time for >91 (90%) encounters. MA or nurse EHR log-in and log-off times in the examination room and physician EHR log-in and log-off times were collected for 72 to 76 (71%-76%) encounters, respectively. Physician exiting examination room time was collected for 88% (89/101) of the encounters (Multimedia Appendix 2). We were able to capture all EHR workflow time points from the audit log data. Table 1 shows the summary statistics of the differences (time-motion–observed time and time derived from audit log data) for each workflow time point.

Table 1. Summary statistics for the difference between time-motion–observed time and time derived from audit log data (N=101).

Workflow event	Encounter, n (%)	Observed and audit data	
		Values, mean (SD)	Values, median (IQR)
Check-in start	101 (100)	0.6 (1.5)	0.3 (0.1 to 0.7)
Check-in end	101 (100)	1.0 (2.9)	0.2 (−0.1 to 1.0)
Room start	—[a]	—	—
MA[b] logs in	76 (75)	0.2 (1.7)	0.4 (−0.1 to 0.9)
MA logs off	79 (78)	0.1 (0.5)	0.3 (−0.1 to 0.7)
Physician enters	—	—	—
Physician logs in	76 (75)	−0.2 (1.5)	0.2 (−0.6 to 1.0)
Physician logs off	72 (71)	0.3 (1.9)	−0.1 (−0.5 to 0.9)
Physician exits	87 (86)	—	—
Observed time between physician log-off and physician exit time	89 (88)	0.8 (1.3)	0.3 (0.1 to 0.7)

[a] Not available in audit log data; represents a workflow "black hole," as illustrated in Figure 1.

[b] MA: medical assistant.

Audit data can capture most clinical workflow time points with high accuracy. The difference between time-motion–observed time and time derived from audit data is <1 minute, and for most time points, the difference is <30 seconds. As shown in Table 1, the largest differences between time-motion time and audit data–derived time were for check-in start (mean 0.6, SD 1.5) and end times (mean 1.0, SD 2.9) as time-motion time recorded the time the patient was called to the front desk to check in and the time the patient left the front desk, whereas audit data recorded the time front desk MAs started the EHR check-in process. We also observed a large variation in the patient exiting the examination room time, with a mean of 2.9 (SD 4.1) minutes, as the time patients exit the examination room was not available in the audit data, and we used the last EHR log-off time by the physician in the examination room as the surrogate to compare with time-motion observations.

As illustrated in Figure 1, we calculated the duration for each black hole based on time-motion data. The duration of these "black holes" varies from approximately half a minute to approximately 2.6 minutes, where the black holes associated with MAs are short (<1 minute in most encounters), and the longest unobserved black hole was the interval between the physician entering the examination room and the EHR log-in (mean duration 4.6, SD 5.6 minutes). ANOVA showed that physicians explained 62% of the duration variance, and patient demographic factors explained approximately 15%. The mean duration between physician exit and patient completion of the encounter was 2.7 (SD 4.1) minutes, and the physician explained 39% of the variance, and 17% was explained by patient factors (Table 2).

Table 2. Observed duration for each black hole using time-motion data and output from ANOVA.

"Black hole"	Time interval (minute), mean (SD)	Variation explained by clinical staff (%)	Variation explained by patient characteristics[a] (%)
MA[b] room to MA log-in	0.73 (1.01)	11.6	35.1
MA log-off to exit	0.51 (1.39)	6.9	46.7
Physician entering examination room to log in	2.62 (1.61)	62.2	14.8
Physician log-off to exit	0.82 (1.29)	11.1	23
Physician exit to patient exit	2.72 (4.11)	39.1	17.5

[a] Patient age, sex, race, and ethnicity.

[b] MA: medical assistant.

Results Among Pilot Users

We found 6 matched control physicians for only 50% (3/6) of pilot physicians. Among the matched physicians, in the pre–CM-SHARE period, 870 control patients associated with 2845 encounters were matched with the same number of patients (870 encounters where CM-SHARE was launched for their encounters). In the poststudy period, 727 patients associated with 852 encounters during which CM-SHARE was launched (cases) were matched with 727 control patients associated with 3754 encounters.

Among all the eligible encounters with patients with cardiometabolic conditions (N=6599; ie, 2845+3754), <10% (595/6599, 9.02%) of the encounters also had a cardiometabolic condition (DM, hypertension, and dyslipidemia) listed as a primary diagnosis of the encounter.

As shown in Table 3 (additional table in Multimedia Appendix 3), the total encounter time was slightly shorter (mean −2.7, SD 1.1 minutes, 95% CI −4.7 to −0.9; mean −1.6, SD 1.1 minutes, 95% CI −3.2 to −0.1) for cases compared with prestudy period controls, as well as for poststudy period controls for 15-minute appointments only, but not for 30-minute appointments. The time saved may be explained by the reduction in the total EHR time for physicians in the examination room, in which CM-SHARE saves approximately 2 minutes (mean −2.0, SD 1.3 minutes; 95% CI −3.4 to −0.9) compared with controls in the prestudy period and a similar amount of time in the poststudy period (mean −1.9, SD 0.9 minutes; 95% CI −3.8 to −0.5). CM-SHARE had no impact on physicians' total time in the examination room or on physicians' total EHR clicks.

Table 3. Summary of difference in time and 95% CI in comparing controls versus cases workflow measures during pilot period.

Workflow measure and scheduled appointment time (minutes)	Difference between matched control in preperiod and matched cases (case and control; n=788)			Difference between matched control in postperiod and matched cases (case and control; n=669)		
	Values, n (%)	Mean (95% CI)	*P* value	Values, n (%)	Mean (95% CI)	*P* value
Total encounter time (minutes)						
≤20	325 (41.2)	−2.7 (−4.7 to −0.9)	.002	310 (46.3)	−1.6 (−3.2 to −0.1)	.02
≥30	463 (58.8)	−0.6 (−3.1 to 2.2)	.11	359 (53.7)	−0.3 (−2.4 to 1.1)	.18
Total physician time in the examination room (minutes)						
≤20	325 (41.2)	−0.6 (−1.9 to 2.0)	.35	310 (46.3)	0.5 (−0.7 to 3.2)	.46
≥30	463 (58.8)	−1.0 (−2.9 to 2.4)	.41	359 (53.7)	−0.7 (−2.1 to 2.0)	.29
Physician EHR[a] time in the examination room (minutes)						
≤20	325 (41.2)	−2.0 (−3.9 to −0.9)	.006	310 (46.3)	−1.9 (−3.8 to −0.5)	.009
≥30	463 (58.8)	−1.3 (−3.4 to 0.5)	.12	359 (53.7)	−1.1 (−3.1 to 0.7)	.15
Physician total clicks in the EHR						
≤20	N/A[b]	N/A	N/A	N/A	6 (−24 to 7)	.29
≥30	N/A	N/A	N/A	N/A	12 (−27 to 10)	.33

[a] EHR: electronic health record.

[b] N/A: not applicable.

However, there was a significant reduction in the total encounter time, total physician time in the EHR, and total physician EHR clicks for two subsets of encounters: encounters with DM as the primary diagnosis and encounters with hypertension as the primary diagnosis. For diabetes encounters, the average total encounter time was 51.3 (SD 5.7) minutes and 49.5 (SD 5.4) minutes for controls in the prestudy period and the poststudy period, respectively, and was reduced to 47.6 (SD 5.1) minutes in CM-SHARE encounters (Multimedia Appendix 4). The mean reduction was 2.1 to 3.5 minutes within matched pairs (Table 4). A substantial reduction was observed in total physician EHR time for diabetes and hypertension encounters in the CM-SHARE group, showing an approximately 30% reduction (the reduction was approximately 2.9-3.5 minutes for hypertension and 4.1-4.3 minutes for diabetes encounters; Table 4). Physician clicks within the EHR were also reduced significantly when using CM-SHARE by 25% in hypertension encounters (from 173 to 129) and 14% for diabetes encounters (from 126 to 108; Multimedia Appendix 4; Table 4).

Table 4. Summary of difference in time and 95% CI in comparing controls versus cases workflow measures during the pilot period.

Workflow measure and primary diagnosis	Difference between matched control in prestudy period and matched cases (case and control) (n=283)			Difference between matched control in poststudy period and matched cases (case and control) (n=312)		
	Value, n (%)	Mean (95% CI)	*P* value	Value, n (%)	Mean (95% CI)	*P* value
Total encounter time (minutes)						
Diabetes	129 (45.6)	−2.8 (−3.9 to −0.2)	.03	145 (46.5)	−2.1 (−3.7 to −0.1)	.03
Hypertension	114 (40.3)	−3.5 (−5.2 to −0.1)	.04	124 (39.7)	−3.4 (−4.9 to −0.9)	.008
Hyperlipidemia	40 (14.1)	−3.7 (−6.2 to 1.8)	.19	43 (13.8)	−4.2 (−5.9 to 0.6)	.09
Total physician time in the examination room (minutes)						
Diabetes	129 (45.6)	−2.1 (−3.9 to 0.8)	.12	145 (46.5)	−2.2 (−3.7 to 0.7)	.15
Hypertension	114 (40.3)	−0.5 (−1.2 to 1.3)	.36	124 (39.7)	−0.6 (−1.3 to 1.2)	.57
Hyperlipidemia	40 (14.1)	−0.1 (−1.0 to 1.7)	.61	43 (13.8)	1.0 (−0.7 to 2.3)	.15
Physician EHR[a] time in the examination room (minutes)						
Diabetes	129 (45.6)	−4.3 (−5.2 to −2.4)	.007	145 (46.5)	−4.1 (−5.0 to −2.6)	.005
Hypertension	114 (40.3)	−2.9 (−3.8 to −0.5)	.02	124 (39.7)	−3.5 (−4.6 to −1.4)	.009
Hyperlipidemia	40 (14.1)	−2.2 (−3.7 to 1.0)	.14	43 (13.8)	−1.4 (−3.1 to 0.8)	.10
Physician total clicks in EHR						
Diabetes	N/A[b]	N/A	N/A	145 (46.5)	−19 (−34 to −2)	.04
Hypertension	N/A	N/A	N/A	124 (39.7)	−50 (−70 to −22)	.006
Hyperlipidemia	N/A	N/A	N/A	43 (13.8)	4 (−11 to 15)	.35

[a] EHR: electronic health record.

[b] N/A: not applicable.

Results Among Spread Users

We matched 48 CM-SHARE spread physicians with 84 control physicians and further matched 1272 patients in the CM-SHARE group with 3412 control patients, associated with 1119 and 4240 eligible encounters, respectively. As shown in Multimedia Appendix 5 and Table 5, a significant reduction in total encounter time for the CM-SHARE group was only observed for encounters with appointments ≤20 minutes (5.3-minute reduction on average) but not for encounters with longer appointment times. However, the total physician's EHR time was significantly reduced for both longer and shorter appointments (reduced by 17%-31%, respectively), and a 16% reduction was observed in physicians' total clicks for both longer and shorter appointments.

Furthermore, <10% of eligible encounters had cardiometabolic as the primary diagnosis. Owing to the limited sample size for matched cases, we only observed a reduction in total encounter time for diabetes as the primary diagnosis (mean −3.2, 95% CI −4.9 to −0.9) and, to a lesser degree, for hypertension encounters (mean −2.9, 95% CI −4.0 to −0.1). For hypertension encounters, we also observed an approximately 33% reduction of physician EHR time in the examination room (mean −2.1, 95% CI −4.7 to −0.2) and a 19% reduction in physician total EHR clicks (mean −24, 95% CI −38 to −12; Multimedia Appendix 6; Table 6).

Table 5. Summary of differences in time and 95% CI in comparing controls versus cases for the encounter-related workflow measures in the Cardiometabolic Sutter Health Advanced Reengineered Encounter spread period, stratified by scheduled appointment time.

Workflow measure and scheduled appointment time (minutes)	Difference between matched control in postperiod and matched cases (case and control)	
	Values, mean (95% CI)	*P* value
Total encounter time (minutes)		
≤20	−5.3 (−7.5 to −0.7)	.002
≥30	3.7 (−6.1 to 3.9)	.37
Total physician time in the examination room (minutes)		
≤2	−2.0 (−4.2 to 1.7)	.15
≥30	3.4 (−4.9 to 2.1)	.23
Physician EHR[a] time in the examination room (minutes)		
≤20	−4.0 (−5.7 to −1.8)	<.001
≥30	−2.1 (−4.5 to −0.3)	.003
Physician total clicks in EHR		
≤20	−11 (−17 to −2)	.02
≥30	−13 (−19 to −4)	.008

[a]EHR: electronic health record.

Table 6. Summary of difference in time and 95% CI in comparing controls versus cases workflow measures during spread period, stratified by primary diagnosis.

Workflow measure and primary diagnosis	Difference between matched control in postperiod and matched cases (case and control)			
	Values, n (%)		Mean (95% CI)	*P* value
	Case (n=132)	Control (n=353)		
Total encounter time (minutes)				
Diabetes	54 (40.9)	157 (44.5)	−3.2 (−4.9 to −0.9)	.01
Hypertension	41 (31.1)	144 (40.8)	−2.9 (−4.0 to −0.1)	.04
Hyperlipidemia	37 (28.0)	52 (14.7)	−3.9 (−6.7 to 1.1)	.14
Total physician time in the examination room (minutes)				
Diabetes	54 (40.9)	157 (44.5)	−1.5 (−4.2 to 1.9)	.51
Hypertension	41 (31.1)	144 (40.8)	−1.9 (−5.2 to 3.7)	.49
Hyperlipidemia	37 (28.0)	52 (14.7)	−1.4 (−5.0 to 3.4)	.78
Physician EHR[a] time in the examination room (minutes)				
Diabetes	54 (40.9)	157 (44.5)	−3.3 (−7.0 to 0.3)	.12
Hypertension	41 (31.1)	144 (40.8)	−2.1 (−4.7 to −0.2)	.03
Hyperlipidemia	37 (28.0)	52 (14.7)	−0.3 (−4.2 to 3.5)	.89
Physician total clicks in EHR				
Diabetes	54 (40.9)	157 (44.5)	−13 (−24 to −3)	.02
Hypertension	41 (31.1)	144 (40.8)	−24 (−38 to −12)	.009
Hyperlipidemia	37 (28.0)	52 (14.7)	−7 (−24 to 10)	.51

[a]EHR: electronic health record.

Discussion

Principal Findings

We successfully used EHR audit data to evaluate efficiency (time and clicks) in the EHR for physicians using a new web-based application and showed that physician EHR time in the examination room was reduced by 17% to 31%, and clicks in the EHR were reduced by 14% to 25%, varying by characteristics of the encounter (ie, scheduled length of the appointment and primary diagnosis of the encounter) when using the web-based application in a pilot phase. More importantly, we also replicated our method for evaluating efficiency with a group of spread users and found similar reductions in time and clicks on the computer in the examination room for both long and short appointments. Few studies report testing on validity [32] similar to ours, and when compared with them, the workflow times derived from audit data in our study are consistent with theirs [40,43,44]. Our study further offers a methodology for using audit data to evaluate the impact of a clinical decision support tool on physician workflow and efficiency.

Comparison With Prior Work

CM-SHARE development was motivated by the desire to design a dashboard with better data integration and an intuitive visual display to facilitate physician decisions and communication with patients [22]. We previously assessed the adoption in the pilot phase and identified the target patient population for which the CM-SHARE is most likely to be used [22]. In this study, the reduction in total EHR time using CM-SHARE was seen not only during pilot testing with physicians but also in the spread physicians who were neither individually trained as in the pilot phase nor frequently interacted with by the CM-SHARE study team. This demonstrates that the initial design intention of CM-SHARE was fulfilled. Furthermore, we observed a significant reduction in physician EHR time in the examination room and no change in the total physician time in the examination room, implying that physicians are likely to have more time to communicate with patients, improving patient satisfaction. From qualitative data taken during the initial pilot [22], we know that 2 pilot providers describe using CM-SHARE for patient education and that users describe CM-SHARE as leading to better discussions with patients, which seems to hold true in spread physicians and may explain the reduction in time spent on the computer with no changes in overall time spent with the patient.

These findings in the initial pilot users and spread sites further show that CM-SHARE is more valuable for patients with diabetes or hypertension-related encounters and less valuable for patients with dyslipidemia. Hyperlipidemia is a common comorbidity for hypertension and diabetes, and the management of dyslipidemia usually occurs either when patients encounter hypertension or diabetes or as one component in encounters related to cardiovascular disease management or prevention [45,46]. Standalone hyperlipidemia encounters are less common than those of hypertension or diabetes; therefore, it is less likely to detect a time reduction in hyperlipidemia-related encounters. In contrast, compared with diabetes-related encounters, physician EHR time for hypertension-related encounters is much shorter (10 minutes for hypertension vs 14 minutes for diabetes in controls), implying that less EHR information is required for physicians to manage patients with hypertension compared with patients with diabetes. We observed a similar percentage (4/14, 4/11, or ~30%) of physician EHR time saved by using CM-SHARE for those 2 conditions, indicating that the information that affects clinical decisions, EHR processes, and clinical tasks for an encounter may not vary significantly for those 2 cardiometabolic conditions. It also suggests considering target population needs when designing and implementing clinical support tools to optimize the benefits to those populations.

Interestingly, among pilot users, we only observed CM-SHARE reducing physicians' EHR time during encounters with shorter scheduled appointment lengths (ie, ≤20 minutes) and not for longer appointments (ie, ≥30 minutes); however, a reduction in EHR time for both long and short appointments was observed in spread physicians. On the basis of the American Academy of Family Physicians and Centers for Medicare and Medicaid Services guidelines on the evaluation and management of office visits [42], for appointment lengths <20 minutes, the number and complexity of problems needing to be addressed during the encounter and the complexity of data to be reviewed were relatively fewer or simpler, whereas ≥ 30-minute appointments usually imply more problems to be addressed and, thus, more patient data to be reviewed. Longer appointments are usually used for more complex patients with more comorbidities [47]. Given the EHR content provided by CM-SHARE, which mostly includes the information directly related to cardiometabolic conditions, and, to a lesser degree, on comorbidities, we expect that CM-SHARE's features may not be sufficient to produce similar reductions in outcomes when addressing comorbidities other than cardiometabolic conditions. It is surprising that CM-SHARE reduces EHR time for longer appointments in spread physicians but not in pilot physicians, in addition to the larger sample size for the spreading evaluation. A possible reason is that the practice pattern and use of CM-SHARE may differ between pilot physicians and spread physicians. Pilot physicians participated in the tool design and were well aware of the target condition that CM-SHARE was designed for, the limitations of CM-SHARE, and the "design thinking" that may drive them more likely to go back to the EHR to manage other comorbidities [48], whereas spread physicians may be likely to "think out-of-box" to optimize the utility of CM-SHARE functions and to use the CM-SHARE as a tool to manage common risk factors for patients with comorbidities. More studies, especially qualitative interviews, are needed to understand this discrepancy between physicians and the difference in whether and how CM-SHARE is used in short and long appointments.

In summary, CM-SHARE has shown improvement in data integration and reduction of EHR time for certain encounters (eg, encounters with a shorter scheduled appointment time and encounters with chief complaints of diabetes and hypertension). A similar design and evaluation approach (eg, user-centered design, workflow integration, and pseudoexperimental design) has the potential to be generalized to other similar clinical decision support tools deployed in real-world settings. If similar improvements in the physician EHR efficiency are observed in other studies, it will provide great insight into redesigning the EHR user interface and reorganizing disease-related contents with a better, more user-friendly visual display.

Limitations

This study had several limitations. First, this study was conducted at a single center with a single EHR system, and the results may not be directly generalizable to other health care systems or systems with a different EHR vendor [41]. However, the methodology, including the use of audit data, validation, imputation, and the study design, is generalizable. Second, audit data may overestimate physicians' EHR time in the examination room. For example, patient examination time or time away from the EHR talking to patients between EHR activities might be counted as EHR time. Finally, this was not a randomized trial, and 2-level matching was applied; therefore, we may not be able to control all confounding variables and nested correlations between physicians and patients, which may be related to EHR efficiency.

Conclusions

Combining audit log files and clinical information from the EHR, we were able to evaluate the impact of a clinical decision support tool (CM-SHARE) on the clinical workflow times and physicians' EHR efficiency. The CM-SHARE web-based application significantly reduced physicians' EHR time in the examination room, particularly for hypertension- or diabetes-related encounters, and less for complex encounters. Our approach offers an innovative way of evaluating digital tools used in clinical settings.

Acknowledgments

The authors would like to acknowledge and thank Dr Ruenell Adams-Jacobs, Dr Scott Cannon, Dr Julieta Dominguez-Jones, Dr Kenneth Hua**,** and Dr Edward Yu for their clinical guidance and involvement in the piloting of the Cardiometabolic Sutter Health Advanced Reengineered Encounter application. The authors would also like to acknowledge Gil Flamino for his development support, user engagement, technology strategy, and management and *Zijun Shen* and *Ridhima Nerlekar* for the initial data curation.

Authors' Contributions

XY was involved in the conceptualization (lead), methodology (lead), formal analysis (lead), software (support), visualization (equal), investigation (lead), writing of the original draft (lead), and review and editing of the manuscript (equal). HH was involved in project administration (lead), investigation (support), visualization (support), preparing the original draft (equal), and review and editing of the manuscript (equal). SM was involved in the methodology (supporting), software (lead), investigation (supporting), data curation (lead), visualization (equal), and review and editing of the manuscript (supporting). MG took part in the investigation (supporting), data curation (supporting), visualization (supporting), and review and editing of the manuscript (supporting). JDR participated in conceptualization (supporting), validation (equal), review and editing of the manuscript (supporting), and formal analysis (supporting). AH was involved in the investigation (supporting), validation (equal), and review and editing of the manuscript (supporting). KK participated in the investigation (supporting), validation (equal), and review and editing of the manuscript (supporting). JBJ was involved in the conceptualization (equal), investigation (supporting), writing of the original draft (supporting), review and editing of the manuscript (equal), supervision (lead), and funding acquisition (lead).

Conflicts of Interest

This study was initiated and led by Sutter Investigators and was funded by AstraZeneca. The funders were involved in the decision to publish the results.

Multimedia Appendix 1
Time-motion data tracking form.
[DOCX File , 17 KB - medinform_v10i9e38385_app1.docx]

Multimedia Appendix 2
Summary of the number of encounters in time-motion observation and the encounters where audit log data were used in validation.
[DOCX File , 15 KB - medinform_v10i9e38385_app2.docx]

Multimedia Appendix 3
Summary of time duration for key encounter-related workflow measures and comparison between prestudy and poststudy period for matched cases and controls.
[DOCX File , 16 KB - medinform_v10i9e38385_app3.docx]

Multimedia Appendix 4
Summary of time duration for key encounter-related workflow measures by primary diagnosis and comparison between prestudy and poststudy period for matched cases and controls during the pilot period.
[DOCX File , 17 KB - medinform_v10i9e38385_app4.docx]

Multimedia Appendix 5
Summary of time duration for key encounter-related workflow measures and comparison between poststudy period for matched cases and controls in the Cardiometabolic Sutter Health Advanced Reengineered Encounter spread period.

[DOCX File , 16 KB - medinform_v10i9e38385_app5.docx]

Multimedia Appendix 6
Summary of time duration for key encounter-related workflow measures and comparison between poststudy period for matched cases and controls in the Cardiometabolic Sutter Health Advanced Reengineered Encounter spread period for encounters with Diabetes, Hypertension or Hyperlipidemia.
[DOCX File , 15 KB - medinform_v10i9e38385_app6.docx]

References

1. National Center for Health Statistics, Center For Disease Control And Prevention. Health, United States, 2016, with Chartbook on Long-Term Trends in Health. Washington, D.C., United States: U.S. Government Printing Office; 2017.
2. Bodenheimer T, Lorig K, Holman H, Grumbach K. Patient self-management of chronic disease in primary care. JAMA 2002 Nov 20;288(19):2469-2475. [doi: 10.1001/jama.288.19.2469] [Medline: 12435261]
3. Tai-Seale M, Olson CW, Li J, Chan AS, Morikawa C, Durbin M, et al. Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. Health Aff (Millwood) 2017 Apr 01;36(4):655-662 [FREE Full text] [doi: 10.1377/hlthaff.2016.0811] [Medline: 28373331]
4. Jankovic I, Chen JH. Clinical decision support and implications for the clinician burnout crisis. Yearb Med Inform 2020 Aug;29(1):145-154 [FREE Full text] [doi: 10.1055/s-0040-1701986] [Medline: 32823308]
5. Downing NL, Bates DW, Longhurst CA. Physician burnout in the electronic health record era: are we ignoring the real cause? Ann Intern Med 2018 Jul 03;169(1):50-51. [doi: 10.7326/M18-0139] [Medline: 29801050]
6. Auerbach AD, Khanna R, Adler-Milstein J. Letting a good crisis go to waste. J Gen Intern Med 2020 Apr;35(4):1289-1291 [FREE Full text] [doi: 10.1007/s11606-019-05552-z] [Medline: 31745851]
7. Collier R. Electronic health records contributing to physician burnout. Can Med Assoc J 2017 Nov 13;189(45):E1405-E1406 [FREE Full text] [doi: 10.1503/cmaj.109-5522] [Medline: 29133547]
8. Adler-Milstein J, Zhao W, Willard-Grace R, Knox M, Grumbach K. Electronic health records and burnout: time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. J Am Med Inform Assoc 2020 Apr 01;27(4):531-538 [FREE Full text] [doi: 10.1093/jamia/ocz220] [Medline: 32016375]
9. Johnson KB, Neuss MJ, Detmer DE. Electronic health records and clinician burnout: a story of three eras. J Am Med Inform Assoc 2021 Apr 23;28(5):967-973 [FREE Full text] [doi: 10.1093/jamia/ocaa274] [Medline: 33367815]
10. Gardner RL, Cooper E, Haskell J, Harris DA, Poplau S, Kroth PJ, et al. Physician stress and burnout: the impact of health information technology. J Am Med Inform Assoc 2019 Feb 01;26(2):106-114 [FREE Full text] [doi: 10.1093/jamia/ocy145] [Medline: 30517663]
11. Goldberg DG, Soylu TG, Grady VM, Kitsantas P, Grady JD, Nichols LM. Indicators of workplace burnout among physicians, advanced practice clinicians, and staff in small to medium-sized primary care practices. J Am Board Fam Med 2020;33(3):378-385 [FREE Full text] [doi: 10.3122/jabfm.2020.03.190260] [Medline: 32430369]
12. Collier R. Rethinking EHR interfaces to reduce click fatigue and physician burnout. Can Med Assoc J 2018 Aug 20;190(33):E994-E995 [FREE Full text] [doi: 10.1503/cmaj.109-5644] [Medline: 30127043]
13. Melnick ER, Harry E, Sinsky CA, Dyrbye LN, Wang H, Trockel MT, et al. Perceived electronic health record usability as a predictor of task load and burnout among US physicians: mediation analysis. J Med Internet Res 2020 Dec 22;22(12):e23382 [FREE Full text] [doi: 10.2196/23382] [Medline: 33289493]
14. Marmor RA, Clay B, Millen M, Savides TJ, Longhurst CA. The impact of physician EHR usage on patient satisfaction. Appl Clin Inform 2018 Jan;9(1):11-14 [FREE Full text] [doi: 10.1055/s-0037-1620263] [Medline: 29298451]
15. Gadd CS, Penrod LE. Dichotomy between physicians' and patients' attitudes regarding EMR use during outpatient encounters. Proc AMIA Symp 2000:275-279 [FREE Full text] [Medline: 11079888]
16. Frankel R, Altschuler A, George S, Kinsman J, Jimison H, Robertson NR, et al. Effects of exam-room computing on clinician-patient communication: a longitudinal qualitative study. J Gen Intern Med 2005 Aug;20(8):677-682 [FREE Full text] [doi: 10.1111/j.1525-1497.2005.0163.x] [Medline: 16050873]
17. Shachak A, Reis S. The impact of electronic medical records on patient-doctor communication during consultation: a narrative literature review. J Eval Clin Pract 2009 Aug;15(4):641-649. [doi: 10.1111/j.1365-2753.2008.01065.x] [Medline: 19522722]
18. Kharbanda EO, Asche SE, Sinaiko AR, Ekstrom HL, Nordin JD, Sherwood NE, et al. Clinical decision support for recognition and management of hypertension: a randomized trial. Pediatrics 2018 Feb;141(2):e20172954 [FREE Full text] [doi: 10.1542/peds.2017-2954] [Medline: 29371241]
19. Mesko B, Győrffy Z. The rise of the empowered physician in the digital health era: viewpoint. J Med Internet Res 2019 Mar 26;21(3):e12490 [FREE Full text] [doi: 10.2196/12490] [Medline: 30912758]

20. Thomas Craig KJ, Willis VC, Gruen D, Rhee K, Jackson GP. The burden of the digital environment: a systematic review on organization-directed workplace interventions to mitigate physician burnout. J Am Med Inform Assoc 2021 Apr 23;28(5):985-997 [FREE Full text] [doi: 10.1093/jamia/ocaa301] [Medline: 33463680]
21. Taheri Moghadam S, Sadoughi F, Velayati F, Ehsanzadeh SJ, Poursharif S. The effects of clinical decision support system for prescribing medication on patient outcomes and physician practice performance: a systematic review and meta-analysis. BMC Med Inform Decis Mak 2021 Mar 10;21(1):98 [FREE Full text] [doi: 10.1186/s12911-020-01376-8] [Medline: 33691690]
22. Jones JB, Liang S, Husby HM, Delatorre-Reimer JK, Mosser CA, Hudnut AG, et al. Development, integration, and adoption of an electronic health record-linked digital health solution to support care for diabetes in primary care. Clin Diabetes 2019 Oct;37(4):338-346 [FREE Full text] [doi: 10.2337/cd18-0057] [Medline: 31660006]
23. Sheon AR, Van Winkle B, Solad Y, Atreja A. An algorithm for digital medicine testing: a NODE.Health perspective intended to help emerging technology companies and healthcare systems navigate the trial and testing period prior to full-scale adoption. Digit Biomark 2018;2(3):139-154 [FREE Full text] [doi: 10.1159/000494365] [Medline: 31032473]
24. The Lancet. Is digital medicine different? Lancet 2018 Jul 14;392(10142):95. [doi: 10.1016/S0140-6736(18)31562-9] [Medline: 30017135]
25. van Dyk L. A review of telehealth service implementation frameworks. Int J Environ Res Public Health 2014 Jan 23;11(2):1279-1298 [FREE Full text] [doi: 10.3390/ijerph110201279] [Medline: 24464237]
26. Granja C, Janssen W, Johansen MA. Factors determining the success and failure of eHealth interventions: systematic review of the literature. J Med Internet Res 2018 May 01;20(5):e10235 [FREE Full text] [doi: 10.2196/10235] [Medline: 29716883]
27. Unertl KM, Novak LL, Johnson KB, Lorenzi NM. Traversing the many paths of workflow research: developing a conceptual framework of workflow terminology through a systematic literature review. J Am Med Inform Assoc 2010;17(3):265-273 [FREE Full text] [doi: 10.1136/jamia.2010.004333] [Medline: 20442143]
28. Zheng K, Guo MH, Hanauer DA. Using the time and motion method to study clinical work processes and workflow: methodological inconsistencies and a call for standardized research. J Am Med Inform Assoc 2011;18(5):704-710 [FREE Full text] [doi: 10.1136/amiajnl-2011-000083] [Medline: 21527407]
29. Lopetegui M, Yen P, Lai A, Jeffries J, Embi P, Payne P. Time motion studies in healthcare: what are we talking about? J Biomed Inform 2014 Jun;49:292-299 [FREE Full text] [doi: 10.1016/j.jbi.2014.02.017] [Medline: 24607863]
30. Kannampallil T, Abraham J. Evaluation of health information technology: methods, frameworks and challenges. In: Cognitive Informatics for Biomedicine. Cham: Springer; 2015.
31. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan W, Sinsky CA, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann Fam Med 2017 Sep;15(5):419-426 [FREE Full text] [doi: 10.1370/afm.2121] [Medline: 28893811]
32. Rule A, Chiang MF, Hribar MR. Using electronic health record audit logs to study clinical activity: a systematic review of aims, measures, and methods. J Am Med Inform Assoc 2020 Mar 01;27(3):480-490 [FREE Full text] [doi: 10.1093/jamia/ocz196] [Medline: 31750912]
33. Bowes WA. Measuring use of electronic health record functionality using system audit information. Stud Health Technol Inform 2010;160(Pt 1):86-90. [Medline: 20841655]
34. Sinsky CA, Rule A, Cohen G, Arndt BG, Shanafelt TD, Sharp CD, et al. Metrics for assessing physician activity using electronic health record log data. J Am Med Inform Assoc 2020 Apr 01;27(4):639-643 [FREE Full text] [doi: 10.1093/jamia/ocz223] [Medline: 32027360]
35. Gray JE, Feldman H, Reti S, Markson L, Lu X, Davis RB, et al. Using digital crumbs from an electronic health record to identify, study and improve health care teams. AMIA Annu Symp Proc 2011;2011:491-500 [FREE Full text] [Medline: 22195103]
36. Ben-Assuli O, Shabtai I, Leshno M. The impact of EHR and HIE on reducing avoidable admissions: controlling main differential diagnoses. BMC Med Inform Decis Mak 2013 Apr 17;13:49 [FREE Full text] [doi: 10.1186/1472-6947-13-49] [Medline: 23594488]
37. Ben-Assuli O, Shabtai I, Leshno M. Using electronic health record systems to optimize admission decisions: the Creatinine case study. Health Informatics J 2015 Mar;21(1):73-88 [FREE Full text] [doi: 10.1177/1460458213503646] [Medline: 24692078]
38. Read-Brown S, Hribar MR, Reznick LG, Lombardi LH, Parikh M, Chamberlain WD, et al. Time requirements for electronic health record use in an academic ophthalmology center. JAMA Ophthalmol 2017 Nov 01;135(11):1250-1257 [FREE Full text] [doi: 10.1001/jamaophthalmol.2017.4187] [Medline: 29049512]
39. Wu DT, Smart N, Ciemins EL, Lanham HJ, Lindberg C, Zheng K. Using EHR audit trail logs to analyze clinical workflow: a case study from community-based ambulatory clinics. AMIA Annu Symp Proc 2017;2017:1820-1827 [FREE Full text] [Medline: 29854253]
40. Hribar MR, Read-Brown S, Goldstein IH, Reznick LG, Lombardi L, Parikh M, et al. Secondary use of electronic health record data for clinical workflow analysis. J Am Med Inform Assoc 2018 Jan 01;25(1):40-46 [FREE Full text] [doi: 10.1093/jamia/ocx098] [Medline: 29036581]

41. Melnick ER, Ong SY, Fong A, Socrates V, Ratwani RM, Nath B, et al. Characterizing physician EHR use with vendor derived data: a feasibility study and cross-sectional analysis. J Am Med Inform Assoc 2021 Jul 14;28(7):1383-1392 [FREE Full text] [doi: 10.1093/jamia/ocab011] [Medline: 33822970]
42. Millette KW. A step-by-step time-saving approach to coding office visits. Fam Pract Manag 2021;28(4):21-26. [Medline: 34254761]
43. Overhage JM, McCallie D. Physician time spent using the electronic health record during outpatient encounters: a descriptive study. Ann Intern Med 2020 Feb 04;172(3):169-174. [doi: 10.7326/M18-3684] [Medline: 31931523]
44. Tai-Seale M, McGuire TG, Zhang W. Time allocation in primary care office visits. Health Serv Res 2007 Oct;42(5):1871-1894 [FREE Full text] [doi: 10.1111/j.1475-6773.2006.00689.x] [Medline: 17850524]
45. THETA. Frequency of testing for dyslipidemia: a systematic review and budget impact analysis. Ont Health Technol Assess Ser 2014;14(7):1-27 [FREE Full text] [Medline: 26316921]
46. Morrison F, Shubina M, Turchin A. Encounter frequency and serum glucose level, blood pressure, and cholesterol level control in patients with diabetes mellitus. Arch Intern Med 2011 Sep 26;171(17):1542-1550 [FREE Full text] [doi: 10.1001/archinternmed.2011.400] [Medline: 21949161]
47. Swanson KM, Matulis JC, McCoy RG. Association between primary care appointment lengths and subsequent ambulatory reassessment, emergency department care, and hospitalization: a cohort study. BMC Prim Care 2022 Mar 06;23(1):39 [FREE Full text] [doi: 10.1186/s12875-022-01644-8] [Medline: 35249539]
48. Mummah SA, Robinson TN, King AC, Gardner CD, Sutton S. IDEAS (integrate, design, assess, and share): a framework and toolkit of strategies for the development of more effective digital interventions to change health behavior. J Med Internet Res 2016 Dec 16;18(12):e317 [FREE Full text] [doi: 10.2196/jmir.5927] [Medline: 27986647]

Abbreviations

CM-SHARE: Cardiometabolic Sutter Health Advanced Reengineered Encounter
DM: diabetes mellitus
EHR: electronic health record
LOS: level of service
MA: medical assistant
PCP: primary care physician

Edited by C Lovis, J Hefner; submitted 01.04.22; peer-reviewed by C Gaudet-Blavignac, N Apathy, F Lau; comments to author 16.05.22; revised version received 10.07.22; accepted 26.07.22; published 06.09.22.

Please cite as:
Yan X, Husby H, Mudiganti S, Gbotoe M, Delatorre-Reimer J, Knobel K, Hudnut A, Jones JB
Evaluating the Impact of a Point-of-Care Cardiometabolic Clinical Decision Support Tool on Clinical Efficiency Using Electronic Health Record Audit Log Data: Algorithm Development and Validation
JMIR Med Inform 2022;10(9):e38385
URL: https://medinform.jmir.org/2022/9/e38385
doi:10.2196/38385
PMID:36066940

Original Paper

Issues With Variability in Electronic Health Record Data About Race and Ethnicity: Descriptive Analysis of the National COVID Cohort Collaborative Data Enclave

Lily Cook[1], MA, PhD; Juan Espinoza[2], MD; Nicole G Weiskopf[1], PhD; Nisha Mathews[3], PhD; David A Dorr[1], MD; Kelly L Gonzales[4,5,6,7], MPH, PhD; Adam Wilcox[8], PhD; Charisse Madlock-Brown[9], PhD; N3C Consortium[10]

[1]Department of Medical Informatics and Clinical Epidemiology, School of Medicine, Oregon Health & Science University, Portland, OR, United States

[2]Department of Pediatrics, Children’s Hospital Los Angeles, Los Angeles, CA, United States

[3]College of Human Sciences and Humanities, University of Houston, Clear Lake-Pearland, TX, United States

[4]Citizen of the Cherokee Nation, Portland, OR, United States

[5]Joint School of Public Health, Oregon Health & Science University-Portland State University, Portland, OR, United States

[6]Founding Indigenous Member, BIPOC Decolonizing Data Council, Portland, OR, United States

[7]Indigenous Equity Institute, Portland, OR, United States

[8]Department of Medicine, Institute for Informatics, Washington University in St. Louis, St. Louis, MO, United States

[9]Tennessee Clinical and Translational Science Institute, University of Tennessee Health Science Center, Memphis, TN, United States

[10]See Acknowledgments

Corresponding Author:
Lily Cook, MA, PhD
Department of Medical Informatics and Clinical Epidemiology
School of Medicine
Oregon Health & Science University
Biomedical Information Communication Center
3280 S.W. Sam Jackson Park Rd.
Portland, OR, 97239
United States
Phone: 1 503 494 4502
Email: lilyjune25@gmail.com

Abstract

Background: The adverse impact of COVID-19 on marginalized and under-resourced communities of color has highlighted the need for accurate, comprehensive race and ethnicity data. However, a significant technical challenge related to integrating race and ethnicity data in large, consolidated databases is the lack of consistency in how data about race and ethnicity are collected and structured by health care organizations.

Objective: This study aims to evaluate and describe variations in how health care systems collect and report information about the race and ethnicity of their patients and to assess how well these data are integrated when aggregated into a large clinical database.

Methods: At the time of our analysis, the National COVID Cohort Collaborative (N3C) Data Enclave contained records from 6.5 million patients contributed by 56 health care institutions. We quantified the variability in the harmonized race and ethnicity data in the N3C Data Enclave by analyzing the conformance to health care standards for such data. We conducted a descriptive analysis by comparing the harmonized data available for research purposes in the database to the original source data contributed by health care institutions. To make the comparison, we tabulated the original source codes, enumerating how many patients had been reported with each encoded value and how many distinct ways each category was reported. The nonconforming data were also cross tabulated by 3 factors: patient ethnicity, the number of data partners using each code, and which data models utilized those particular encodings. For the nonconforming data, we used an inductive approach to sort the source encodings into categories. For example, values such as “Declined” were grouped with “Refused,” and “Multiple Race” was grouped with “Two or more races” and “Multiracial.”

Results: “No matching concept” was the second largest harmonized concept used by the N3C to describe the race of patients in their database. In addition, 20.7% of the race data did not conform to the standard; the largest category was data that were missing. Hispanic or Latino patients were overrepresented in the nonconforming racial data, and data from American Indian or Alaska Native patients were obscured. Although only a small proportion of the source data had not been mapped to the correct concepts (0.6%), Black or African American and Hispanic/Latino patients were overrepresented in this category.

Conclusions: Differences in how race and ethnicity data are conceptualized and encoded by health care institutions can affect the quality of the data in aggregated clinical databases. The impact of data quality issues in the N3C Data Enclave was not equal across all races and ethnicities, which has the potential to introduce bias in analyses and conclusions drawn from these data. Transparency about how data have been transformed can help users make accurate analyses and inferences and eventually better guide clinical care and public policy.

(JMIR Med Inform 2022;10(9):e39235) doi:10.2196/39235

KEYWORDS

social determinants of health; health equity; bias; data quality; data harmonization; data standards; terminology; data aggregation

Introduction

The United States has had more COVID-19 cases and deaths than any other country [1]. Black or African American, Hispanic or Latino, and American Indian or Alaska Native (AI/AN) communities have experienced disproportionate morbidity and mortality from COVID-19 [2-5]. Compared with the non-Hispanic White population, the Black or African American population has a higher prevalence of COVID-19, as well as higher mortality and hospitalization rates from the virus [2]. The Centers for Disease Control and Prevention (CDC) reported that, between February 2020 and May 2020, Hispanic or Latino and non-White individuals under 65 years of age were 2 to 3 times more likely to die from COVID-19 than their non-Hispanic White counterparts [4]. COVID-19 incidence for AI/AN persons is estimated to be 3.5 times higher than for non-Hispanic White persons [5]. The full consideration of the social, economic, and health impacts of COVID-19 on these communities relies on data sets structured to answer such questions.

Resources have been created with the intention of tracking, quantifying, and analyzing the impact of COVID-19 within and across populations [6-8]. The largest such resource in the United States is the National COVID Cohort Collaborative (N3C), a National Institutes of Health (NIH)–funded collaboration between the National Center for Advancing Translational Sciences (NCATS) and the Center for Data to Health [8]. The N3C Data Enclave is also one of the largest collections of COVID-19 patient-level data globally [7], providing harmonized electronic health record (EHR) data from 56 health care institutions and networks across the country. Currently, 1615 researchers representing 186 research institutions have been granted access to the Enclave to work on 215 research projects [9].

Large data sets like N3C, whether centralized or distributed, face a substantial challenge in the form of data heterogeneity, stemming from varying data collection, documentation, and coding practices [10]. These upstream processes may result in data quality problems and other artifacts that can lead to data loss and possibly misleading signals in the data [11]. The encodings used to represent race and ethnicity vary across institutions and data models and require specialized harmonization [12]. Indeed, 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 are collected and structured by health care organizations. The Institute of Medicine’s landmark report on racial and ethnic disparities in health care, *Unequal Treatment: Confronting Racial and Ethnic Disparities in Healthcare*, highlighted the need for standardized collection and reporting of race and ethnicity data [13].

Data standardization and harmonization is one of the best tools for combating heterogeneity and ensuring that observed signals are genuine. The N3C provides a unique opportunity to assess how different health care systems in various locations collect and conceptualize information about their patients’ race and ethnicity and to examine efforts to integrate these categories across different data models. In this paper, we discuss race and ethnicity from the perspective of data standards and database harmonization.

The standard most commonly used by health care systems to collect and organize data about race and ethnicity was created for the 2000 US Census. The Office of Management and Budget (OMB) released this standard in 1997 [14], and shortly afterward, the CDC added encodings to the OMB Standard; both are shown in Table 1 [15]. To maintain clarity and consistency, we used these terms throughout this paper.

The 1997 OMB classification system was then adopted with minor changes by Health Level Seven International (HL7), the creator of the standard most widely used by health care systems to transmit and receive health records [16]; any references to “the health care standard” in our paper refer to how this information is currently structured in HL7 Fast Healthcare Interoperability Resources (FHIR).

The current health care standard uses terminology in a manner different from how it is used colloquially. For the purposes of collecting and organizing self-reported patient demographic data, race and ethnicity are considered distinct concepts, and ethnicity refers only to Hispanic or Latino origin. Thus, ethnicity has 3 minimum codes: Patients can either be Hispanic or Latino or non-Hispanic or Latino. However, this category is intended to be hierarchical, and “granular” ethnicity refers to the 41 subcategories (e.g., Panamanian, Venezuelan) that are required to roll up into Hispanic or Latino.

Table 1. Office of Management and Budget (OMB) revisions to the Standards for the Classification of Federal Data on Race and Ethnicity, 1997.

OMB category	HL7[a] code	Category definition
Race: American Indian or Alaska Native	1002-5	A person having origins in any of the original peoples of North and South America (including Central America) and who maintains tribal affiliation or community attachment
Race: Asian	2028-9	A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam
Race: Black or African American	2054-5	A person having origins in any of the black racial groups of Africa. Terms such as "Haitian"; or "Negro"; can be used in addition to "Black or African American"
Race: Native Hawaiian or Other Pacific Islander	2076-8	A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands
Race: White	2106-3	A person having origins in any of the original peoples of Europe, the Middle East, or North Africa
Ethnicity: Hispanic or Latino	2135-2	A person of Mexican, Puerto Rican, Cuban, South or Central American, or other Spanish culture or origin, regardless of race. Ethnicity is considered a distinct category from race

[a] HL7: Health Level Seven International.

The 2009 Institute of Medicine Subcommittee on Standardized Collection of Race/Ethnicity Data for Healthcare Quality Improvement report provided direction for health care systems on how to implement the federal standard [17]. Because the health care standard treats race and ethnicity as separate concepts, it is recommended that the question about Hispanic of Latino origin be presented first when gathering demographic information from patients. The standard has 5 minimum categories for race: (1) AI/AN, (2) Asian, (3) Black or African American, (4) Native Hawaiian or Other Pacific Islander, and (5) White. The health care standard for race data is hierarchical, with almost 900 different subcategories that could be used to describe more granular race categories, all of which are required to collapse (or "roll up") into o1 of the 5 major categories. "Other" race is deprecated within HL7, although "unknown" and "asked but not answered" are permissible [16]. For patients who identify as multiracial, the 1997 OMB Standard and the Institute of Medicine Subcommittee both recommend allowing for the selection of more than one race rather than offering a single "multiracial" category [14,17]. However, the OMB acknowledged that allowing for multiple selections creates complications during tabulation and analysis, and the Institute of Medicine noted that "some health information technology systems are unable to support the collection and reporting of data in a 'Select one or more' manner" [17].

The health care standard only recommends a structure for how information about patient race and ethnicity should be stored; in practice, there are wide variations in how health care systems collect this information. Studies have documented that it is frequently missing from the patient record, and when it is collected, it is often of poor quality [18-23]. Our objective was to explore variations in how health care 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 data in N3C by focusing on conformance to standard definitions, missingness, and misclassification.

Methods

Data Source

Although the size of the N3C Data Enclave has continued to grow, at the time of our analysis (July and August of 2021), the N3C Data Enclave contained health records from 6.5 million patients tested for COVID-19, including 2.1 million who had tested positive. The data in the Enclave are updated weekly with new information. To keep our numbers consistent, we used Release-v40-2021-07-30 to conduct analyses whenever possible; small numerical inconsistencies may appear as the result of occasions when different release versions were used.

Significant technical and regulatory hurdles were addressed to make the N3C Data Enclave available to researchers seeking insight into COVID-19. The clinical data are stored in the Observational Medical Outcomes Partnership (OMOP) Common Data Model; institutions using Accrual to Clinical Trials (ACT), National Patient-Centered Clinical Research Network (PCORnet), and TriNetX common data models have their data mapped to OMOP, while those already using OMOP have their data ingested directly. The Common Data Model Harmonization project provided syntactic mapping with conversion logic and semantic mapping to the OMOP vocabulary. N3C met with subject matter experts from source common data models and the Observational Health Data Sciences and Informatics community to finalize these mappings, which are available to the public on GitHub [12].

Ethical Review

The protocol for this study was approved by the Institutional Review Board at Oregon Health and Science University (IRB ID STUDY00022764). This study was granted a waiver because the study design—a retrospective review of existing records—involved minimal risk. Waiver of the formal written consent process did not adversely affect the rights or welfare of the participants. This study was performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

Analyses

To quantify the variability in how health care institutions are reporting data about patient race and ethnicity to N3C, we used a multistep process that included data processing, terminology harmonization, and descriptive analyses. First, we sorted the harmonized data into “conforming” and “nonconforming” categories. We defined race and ethnicity data as “conforming” if they had been mapped to 1 of the 5 minimum categories for race congruent with the health care standard: White, Black or African American, Native Hawaiian or Other Pacific Islander, Asian, or AI/AN [14-16]. Ethnicity data were conforming if they were harmonized into 1 of the 2 standard categories for ethnicity: “Hispanic or Latino” or “Not Hispanic or Latino.” All other data—including missing data—were deemed “nonconforming.”

Next, we delved into these categories by comparing these harmonized data to their original source encodings. To make the comparison, we tabulated the original source codes, enumerating how many patients had been reported with each encoded value and how many distinct ways each category was reported. This allowed us to get a better idea of how the health care institutions were reporting the data to the N3C and to approximate how well the source institutions were adhering to the health care standard. The nonconforming data were also cross tabulated by 3 factors: patient ethnicity, the number of data partners using each code, and which data models utilized those particular encodings. For the nonconforming data, we used an inductive approach to sort the source encodings into categories. For example, values such as “Declined” were grouped with “Refused,” and “Multiple Race” was grouped with “Two or more races” and “Multiracial.”

These analyses were conducted within the N3C Data Enclave using the software tools available within the platform during July 2021 and August 2021. Additional descriptive statistics were done with Excel. Figures were developed using Lucidchart, Excel, and Keynote.

Comparison With Other Data Sources and Repositories

To assess the external validity of the N3C race and ethnicity data, we compared race and ethnicity distributions across multiple data sources, including the 2019 American Communities Survey (ACS) demographic and housing estimates and Cerner HealthFacts (CHF) [24]. ACS data are compiled by the US Census Bureau and provide yearly updates and estimates to key demographic, economic, housing, and social data. CHF is a data warehouse that includes almost 70 million patients treated at hospitals and clinics throughout the United States between 2001 and 2017 using the Cerner EHR platform. Total unadjusted COVID-19 cases and deaths from the CDC are also included for comparison [25,26].

Results

Harmonized, Mapped Data

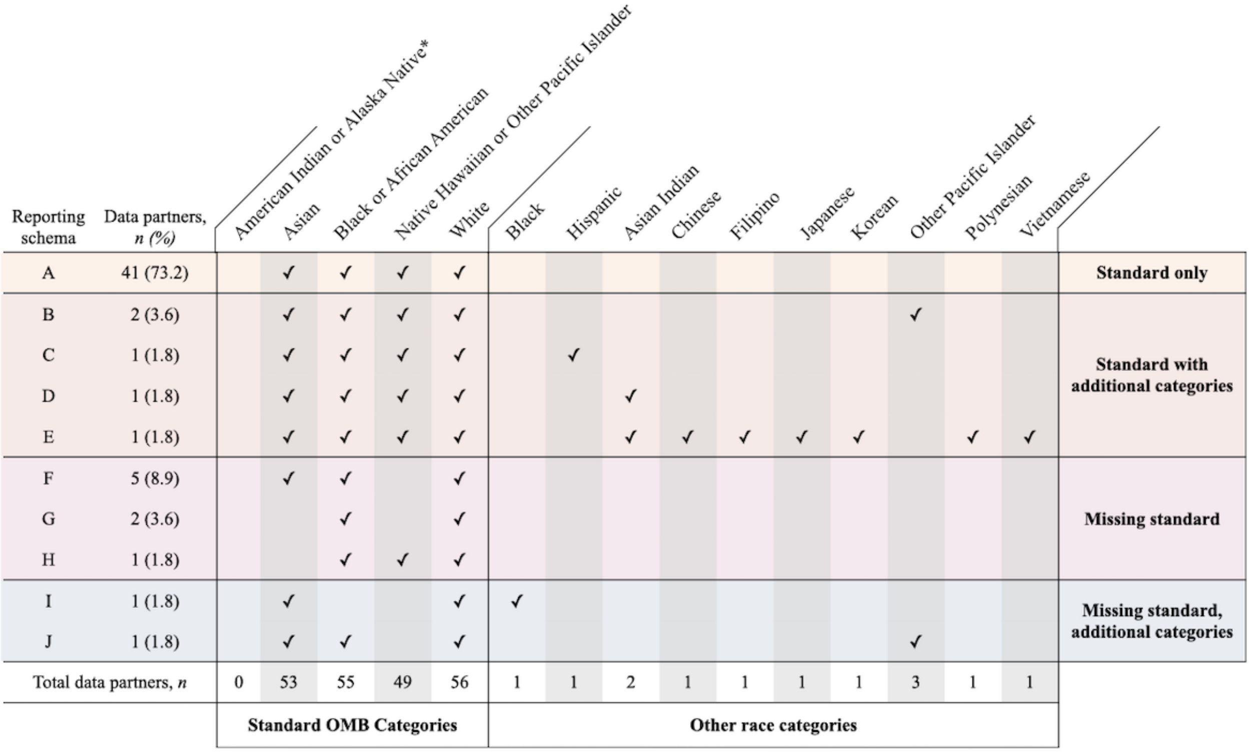
There are a total of 25 harmonized categories for race available in the N3C Data Enclave, representing mapped data contributed by 56 health care institutions. The top 10 concepts used by the N3C to describe the harmonized categories used to describe the race of the patients in the database are shown in Textbox 1. The top 3 harmonized categories—White, “No matching concept,” and Black or African American—account for 94.3% (6,140,139/6,513,464) of the data. No patients with race of AI/AN were found; at the request of the NIH, the health records of AI/AN patients were intentionally obscured during ingestion (see Discussion) [27].

Textbox 1. Top 10 harmonized concepts used by the National COVID Cohort Collaborative (N3C) to describe the race of patients in the database.

1. White
2. No matching concept
3. Black or African American
4. Asian
5. Null
6. Unknown
7. Other
8. Other race
9. Black
10. No information

Following ingestion and mapping to standard OMOP race categories, we identified 10 different reporting schema (combinations of reported race categories) among contributing institutions as shown in Figure 1. Of the 56 data partners, 41 (73%) had harmonized patient race data that adhered to the standard OMB race categories other than AI/AN, as noted in the previous paragraph. Of the data partners, 5 had harmonized data that included all the standard race categories, plus some additional categories such as Filipino or Korean that had not been correctly rolled up into main categories for tabulation purposes (both are subcategories of Asian). Data from 10 contributing institutions (10/56, 18%) omitted at least one of the standard race categories other than AI/AN. The only OMB race category present for all data partners was White. Ethnicity was as a separate field present for 51 (51/56, 91%) of the data partners.

Figure 1. Race data reporting schema by contributing sites. Although data partners did contribute data on American Indian or Alaska Native patients, as noted elsewhere, these data were intentionally obscured. OMB: Office of Management and Budget.

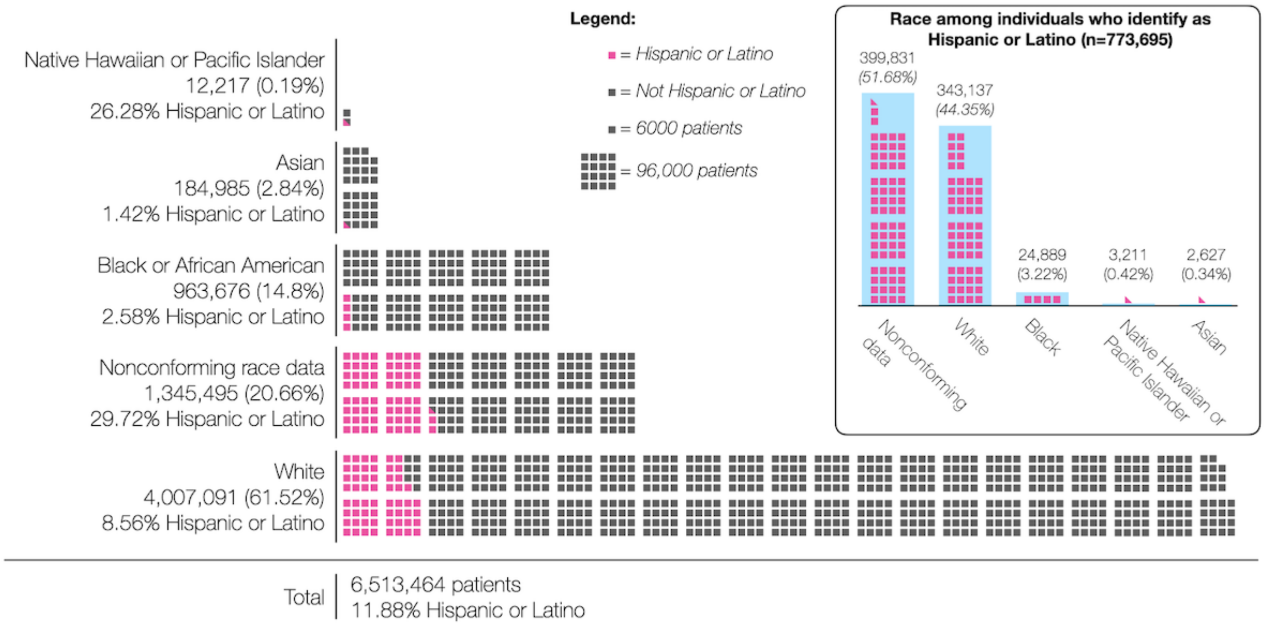


Reporting schema	Data partners, *n (%)*	American Indian or Alaska Native*	Asian	Black or African American	Native Hawaiian or Other Pacific Islander	White	Black	Hispanic	Asian Indian	Chinese	Filipino	Japanese	Korean	Other Pacific Islander	Polynesian	Vietnamese	
A	41 (73.2)		✓	✓	✓	✓											**Standard only**
B	2 (3.6)		✓	✓	✓	✓								✓			**Standard with additional categories**
C	1 (1.8)		✓	✓	✓	✓		✓									
D	1 (1.8)		✓	✓	✓	✓			✓								
E	1 (1.8)		✓	✓	✓	✓			✓	✓	✓	✓	✓		✓	✓	
F	5 (8.9)		✓	✓		✓											**Missing standard**
G	2 (3.6)			✓		✓											
H	1 (1.8)			✓	✓	✓											
I	1 (1.8)		✓			✓	✓										**Missing standard, additional categories**
J	1 (1.8)		✓	✓		✓								✓			
Total data partners, *n*		0	53	55	49	56	1	1	2	1	1	1	1	3	1	1	
		Standard OMB Categories					**Other race categories**										

Conforming Data

Of the data about race, 79.3% (5,167,969/6,513,464) had been harmonized to 1 the 5 main categories recommended in the health care standard. Examining the harmonized data that does conform to the standard, Figure 2 illustrates the racial and ethnic makeup of patients in N3C. The source data showed that White race was originally reported to the N3C by health care institutions a total of 21 different ways, most commonly using the PCORnet code 05 (1,442,961/4,007,091, 36.0% of all White patients). “Jewish” was the only granular category available in the source data for patients whose race had been mapped to White, and 141 of the patients whose race had been mapped to White had Jewish ancestry recorded in the source data.

Figure 2. Race and ethnicity data in the National COVID Cohort Collaborative (N3C) after harmonization.



The most common code found in the source data to report Black or African American patients was PCORnet’s encoding 03 (411,537/963,676, 42.7%). Although the source data contained 24 different encodings for this group, there were no granular subcategories of Black or African American available in either the source or the mapped data.

The source data for patients whose race had been harmonized to Asian showed 22 distinct encodings, most commonly using the PCORnet code 02 (77,426/184,985, 41.9% of all Asian patients). Source data revealed 6 more granular race categories had been rolled up into Asian during the harmonization process; these more granular data represented 546 patients. The most common of these granular subcategories was Asian Indian (n=388). However, in the harmonized data, there were 1534 additional Asian Indian patients who were not rolled up into the Asian category.

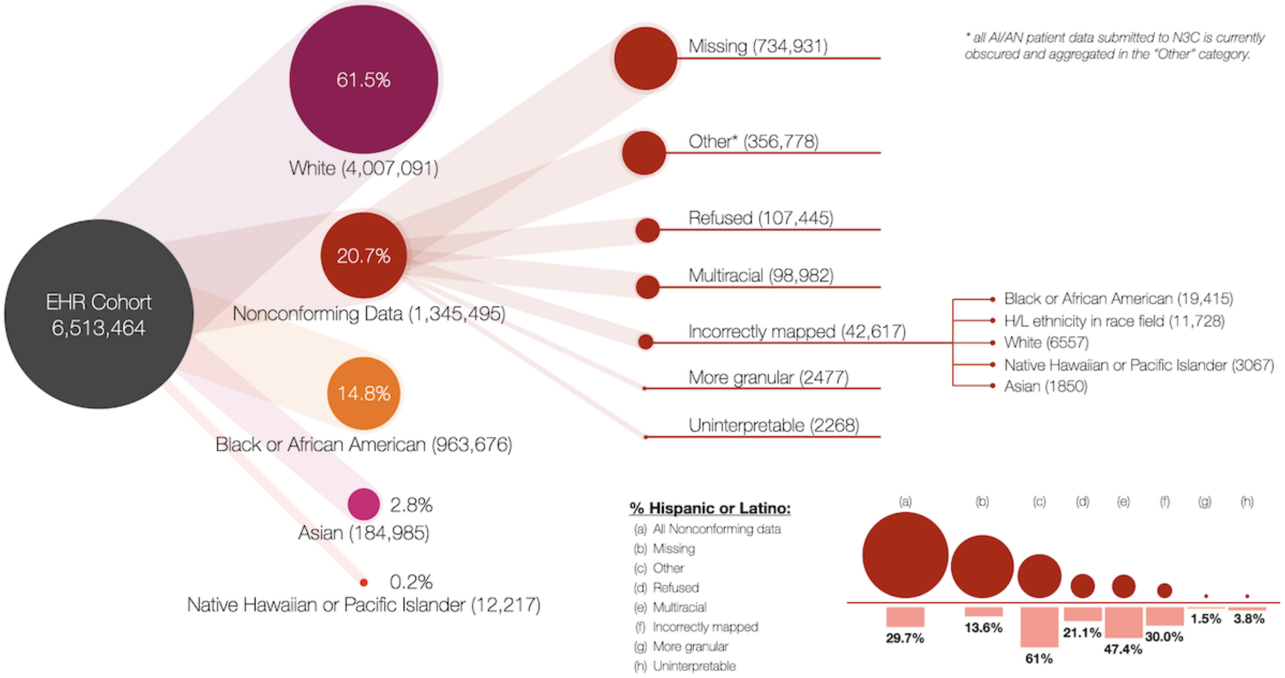
Native Hawaiian or Other Pacific Islander is the smallest of the standardized racial categories found in the Enclave, and the data were initially reported using 24 different encodings prior to harmonization. In the source data, we found 2 granular subcategories that had been rolled up into Native Hawaiian or Other Pacific Islander: Guamanian/Chamorro and Polynesian. This group contained the largest proportion of people who also identified as Hispanic or Latino.

Overall, 83.8% (5,456,162/6,513,464) of the data about ethnicity conformed to the standard. There was a total of 26 different encodings to represent Hispanic or Latino ethnicity in the source data, including 7 granular subcategories such as Puerto Rican, Mexican, and South American.

Nonconforming Data

About 20.7% (1,345,495/6,513,464) of the data in the N3C had not been harmonized to one of the 5 primary race categories described in the health care standard. As shown in Figure 3, nonconforming data could be divided into 7 categories.

Figure 3. Weighted tree diagram of nonconforming race data. AI/AN: American Indian or Alaska Native; EHR: electronic health record; H/L: Hispanic/Latino; N3C: National COVID Cohort Collaborative.



Missing

Incompleteness was the most common reason for race data to be nonconforming, and source data showed that 11.3% (734,931/6,513,464) of all patients in the N3C Data Enclave were marked as missing race data. Of the contributing health care institutions, 31 reported missing data in 29 distinct ways; most often, a zero was recorded to indicate that the data were incomplete (n=348,057). Of the patients missing data about race, 13.6% (99,853/734,931) were noted as being Hispanic or Latino in the ethnicity column.

Other

The second largest category of nonconforming race data was patients labeled by health care systems as "Other" race. The majority of these patients (217,476/356,778, 61.0%) was recorded as being of Hispanic or Latino ethnicity. Currently, all AI/AN patient data submitted to N3C are obscured and aggregated in the "Other" category [28]. This transformation has thus far rendered data from this cohort unavailable to researchers.

Refused

Patients who declined to answer questions about race represented 8.0% (107,445/1,345,495) of the nonconforming data. However, examining the data about ethnicity showed that 21.1% (22,683/107,445) of these patients were Hispanic or Latino.

Multiracial

Multiracial patients represented 7.4% (98,979/1,345,495) of the nonconforming data and 1.5% (98,979/6,513,464) of all the patients in the N3C Data Enclave. Of the 257 different codes used by systems to represent race in the nonconforming data, 119 of them were distinct codes used to represent multiracial patients. Much of the variety was due to some systems allowing patients to select multiple races, which was then reported as several selections in a single column. Although only 3.8%

(3764/98,979) of all multiracial patients actually had more than one race recorded, this 3.8% represented 101 different combinations of codes. The most common of these were combinations of White and Black or African American (n=1563).

Misclassified

Examining the source data revealed that 3.1% (42,617/1,345,495) of the nonconforming race data, 0.6% (42,617/6,513,464) of all data in the N3C, were not mapped to the appropriate standard race concepts. Although only 14.8% (963,676/6,513,464) of the patients in the Enclave are Black or African American, source data showed that 45.6% (19,415/42,617) of these misclassified patients should have had their race mapped to Black or African American. The next largest group of misclassified patients was those whose source institutions had recorded Hispanic or Latino ethnicity in the race field (n=11,728)—the N3C Data Enclave treats Hispanic or Latino separately from race. Confusingly, 19.3% (2258/11,728) of the patients whose race was reported as Hispanic or Latino were labeled as Not Hispanic or Latino in the ethnicity field. Patients identified as White represented 15.9% (6557/42,617) of the misclassified nonconforming data; 7.4% (3067/42,617) of these misclassified patients were Native Hawaiian or Other Pacific Islander, and 4.5% (1850/42,617) were Asian.

Uninterpretable

For 2268 patients, the source institution had provided a code such as "@" that did not conform to those recognized by any of the known data models. There were 2 encodings we were unable to decipher, both of which came from institutions using the TriNetX Common Data Model.

More Granular

Finally, 2477 patients did not map to 1 of the 5 categories because they had been labeled with a granular racial subcategory that had not been rolled up into 1 of the 5 main race categories. Nine racial subgroups are available in the nonconforming data in the N3C; 6 of these (Asian Indian, Filipino, Chinese, Korean, Vietnamese, and Japanese) should have been rolled up into the larger category of "Asian." The most widely reported subcategory we found in the nonconforming data was "Asian Indian" (n=1534). These granular data all came from health care systems using the OMOP Common Data Model.

For patients with Hispanic or Latino ethnicity, "nonconforming" was the single largest racial category (399,831/773,695, 51.7%). One data partner mapped 2533 patients whose race had originally been recorded as Hispanic or Latino to a racial category that was subsequently labelled only as "non-White."

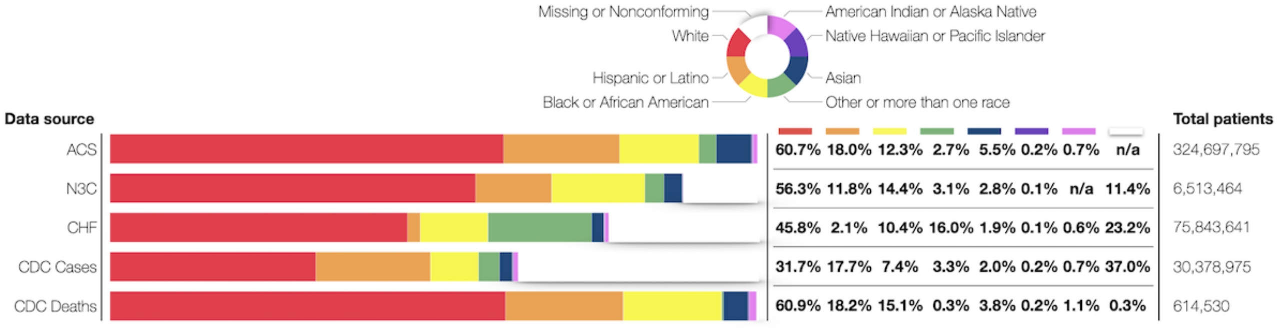
Variations by Common Data Model

Four common data models are used by the health care institutions contributing data to the N3C Data Enclave: OMOP, PCORnet, TriNetX, and ACT. Some OMOP sites also included data in the PEDSnet common data model, which is an extension of OMOP that includes pediatric-specific data fields and standards such as age-normalized anthropometrics [29]. The Enclave itself uses the OMOP model, and non-OMOP contributing institutions preprocess their data so they can be harmonized to the OMOP model. When stratifying the patient data by the data model used by their health care institution, we found that data about patient race from TriNetX had the best conformance; 86.2% (711,075/825,001) of the TriNetX data conformed to 1 of the 5 main categories. Only 66.2% (64,242/97,097) of the race data from OMOP PEDSnet, on the other hand, achieved conformance. We found that, depending on the data model, the conformance of data about ethnicity varied more widely than the race data; although 93.1% (2,146,229/2,305,731) of data from health care institutions using PCORnet's data model conform to the standard for reporting patient ethnicity, only 50.8% (271,304/534,179) of ethnicity data from institutions using the ACT model were adherent to the standard.

Comparison With Other Data Sources

Figure 4 shows how the distribution of race and ethnicity data in N3C compares with the United States overall (ie, the ACS) and 1 other EHR-based data repository, CHF. N3C, similar to CHF, has fewer Hispanic or Latino and Asian patients than the ACS but comparable rates for other groups. This is likely related to both the types of institutions that contribute data (and the patients they serve) as well as the large amount of missing or nonconforming data in both data sets.

Figure 4. Comparisons of race and ethnicity data across data sets. Data from American Indian or Alaska Native patients in the National COVID Cohort Collaborative (N3C) are labeled "not applicable" because these data were obscured until the completion of the Tribal Consultation. ACS: American Communities Survey; CDC: Centers for Disease Control and Prevention; CHF: Cerner HealthFacts.



Discussion

Principal Findings

Our analyses of the N3C Data Enclave revealed a number of facts that are important for researchers to consider when drawing conclusions based on these data. First, "no matching concept" was the second largest harmonized racial group in the N3C. A substantial portion of the records (20.7%) were in some way nonconforming, including 11.7% of all records that were missing race or ethnicity data (missing data were considered a subcategory of nonconformance in this study). Second, while data collection at the point of care needs improvement, there are also opportunities to improve the quality of these data at various points in the data pipeline. Finally, the impact of these data quality issues was not equal across all races and ethnicities. The magnitude and type of nonconformance varied across race and ethnicity, with patients of color and vulnerable communities overrepresented in the misclassified data and nonconforming data.

Implications for COVID-19 Research

The fact that the data were not randomly nonconforming means there is potential to introduce bias in analyses and conclusions drawn from these data. Because data in categories such as "other" or "missing" are often discarded by data users, any patients in those categories are at risk of being inadvertently excluded from research. Data we refer to as nonconforming included several categories that should have been either mapped or rolled up into a main category. For example, we found that the harmonized data included 18,885 patients who had been categorized as "Black" instead of being mapped into the standardized "Black or African American" category. This indicates a significant amount of heterogeneity in the data about race, an issue that may fracture research cohorts and create noise in the data. This could cause problems if data users conducting queries on the N3C database pull the information from one group and inadvertently omit the other. At best, this can be a rate-limiting factor for researchers who must then spend extra time harmonizing the data and doing the mapping themselves rather than studying COVID-19.

It is, however, necessary to put these findings in context. The problems with the data about race and ethnicity are not exclusive to the N3C; indeed, a report from the CDC entitled "Addressing Gaps in Public Health Reporting of Race and Ethnicity for COVID-19" documented the same issue with public health data [30]. Compared with the CDC data on COVID-19 cases, the N3C system has significantly less missingness at 11%, compared with 24% in the CDC COVID-19 case data set. Of the data, 79% conform to CDC standard racial categories (higher than the 64% in the CHF data set), making this repository useful for COVID-19 health disparities research. Moreover, because the N3C Data Enclave gives access to race and ethnicity source values, we were able to assess misclassification and can update the racial and ethnic categorization of patients for research purposes. Though granular data represent a small portion of the overall data, they can be used for small-scale projects analyzing differences within racial categories. Given that this data set has representation from all regions in the United States [8], it can be used to validate against the CDC COVID-19 positivity rates by race.

Mismappings

Although an exhaustive assessment of the causes of mismapping is not feasible given the various mapping and transformation steps that occur upstream of the N3C Data Enclave, many occur during site-level data entry and processing. For example, contributing sites employ standard scripts to map EHR data to common data models. When data preparation is automated using such scripts, patients who have been assigned deprecated codes at the point of care may ultimately be harmonized to "No matching concept." Misclassification of patients might also occur if multiple values have been entered into a single field, as when more than one race has been selected or when sites use the "single question" format when gathering demographic information. We hope to utilize the results of this analysis to add additional coding to these scripts to prevent misclassification and to identify ways to correct the race and ethnicity data postingestion.

"Hispanic or Latino," Race, and Ethnicity

Our finding that Hispanic or Latino patients are overrepresented in the nonconforming race data may reflect that the 2-concept system (ie, recording "race" and "ethnicity" separately) continues to be a source of variability. Although the health care standard recommends that self-reported race and ethnicity be collected as separate concepts, some health care systems combine them and offer "Hispanic or Latino" as a possible selection under Race. The current PCORnet common data model specification recommends mapping data from patients whose race has been recorded as Hispanic or Latino to "Other," which explains some of our finding that 61% of the data harmonized to the "Other" category come from Hispanic or Latino patients [31]. During the 2010 Census, the US Census Bureau tested a combined race-ethnicity question and found that including Hispanic origin as a racial category dramatically reduced both the item nonresponse rate and the selection of "some other race." The results of the Census testing suggest that the issues with Hispanic/Latino data are, at least in part, attributable to the 2-question structure [32]. However, it should be noted that Hispanic or Latino patients were also overrepresented in other categories of nonconforming data, such as "Refused" (21.1% Hispanic or Latino) and "Multiracial" (47.4% Hispanic or Latino). This suggests that the heterogeneity in the data from Hispanic/Latino patients may also be a result of the difficulty people have selecting from standardized categories that they feel do not adequately represent them.

Obscured Data From American Indian or Alaska Native Patients

The lack of accessible data on AI/AN populations is a limitation of the data set. The Urban Indian Health Institute has stated that "current standard data collection practices by many federal, state, and local entities effectively omit or misclassify AI/AN populations, both urban and rural. This is particularly concerning in the midst of the COVID-19 pandemic as these current standards of practice are resulting in a gross undercount of the impact COVID-19 has on Native people" [33]. A number of federal laws, treaties, and executive orders has established the sovereignty of Tribes and Tribal Nations over their data and the power to regulate research, although the gap between *recognition* of those rights and the *assertion* of those rights remains wide [34-36]. A legacy of harm, medical maltreatment, and research misconduct has engendered mistrust between the Tribal and clinical research communities [36]. To begin to address these issues, Tribal leaders, scholars, and advocates have established protocols and institutions to ensure human protections for research involving the AI/AN community [37]. In 2010, the US Department of Health and Human Services established a formal Tribal Consultation Policy to create a mechanism for collaboration at the federal level [38].

In December of 2021, the NCATS formally initiated a Tribal Consultation about the N3C Data Enclave. The NCATS Framing Letter, "NIH Tribal Consultation on the National COVID Cohort Collaborative (N3C)," states, "Ideally, NIH would have sought Tribal Consultation before the start of this program. However, given other COVID-related Consultations and urgency of the pandemic, NCATS decided to obscure AI/AN data until consultation could occur. During the consultation, the NIH will seek input on whether and how to make AI/AN data available through N3C" [28]. The N3C Tribal Consultation took place on February 11, 2022, and as of this writing, the testimony of the Tribal Leadership is being collected. NCATS expects to implement their recommendations by summer 2022.

It is important to note that our analysis is not an endorsement of the standard developed by the OMB and implemented by federal agencies but rather a description of how health care institutions around the United States have been implementing it. Indeed, our position is that deviations from the standard are a signal of the manner in which such categories are both arbitrary and reductionist. Our perspective is that race and ethnicity are not biological categorizations; instead, they should be viewed as social constructs that are highly context-dependent and tied to existing power dynamics. The US Census Bureau stresses this point, stating that these categories "generally reflect a social definition of race recognized in this country and not an attempt to define race biologically, anthropologically, or genetically" [39]. As variables in clinical research, the utility of race and ethnicity is that they can be used as highly imperfect proxies for the complex systemic factors (eg, racism, colonialism, socioeconomic barriers to health care delivery systems) that drive and perpetuate inequities [40,41].

Limitations

Finally, it should be noted that data partners, as the contributing health care institutions are referred to in the N3C, were provided with anonymity as a consideration for contributing data. This means that the provenance of these data is limited, so we do not know how they were initially collected. Finally, because most of the contributing health care institutions are recipients of a Center for Translational Science Award with an established relationship with the NCATS, it is likely that academic medical centers are overrepresented.

Conclusion

Twenty-eight years after Congress mandated the inclusion of racial and ethnic minority groups in federally funded clinical research with the NIH Revitalization Act [42], the ongoing lack of racially and ethnically diverse cohorts remains a challenge to improving equity in research and health care [43]. Because the COVID-19 pandemic disproportionately impacts communities of color by exacerbating existing health inequities, the accurate identification of these cohorts within N3C is crucial to identifying, understanding, and ultimately addressing these disparities. Data problems arise for many reasons, but primary among them is the discrepancy between how institutions conceptualize race and ethnicity and the far more varied ways people identify themselves [44]. The complex history of racial identification in the United States has resulted in shifting concepts of race and ethnicity [45]. Self-identified race and ethnicity are often dependent on physical attributes that, although heritable, correlate poorly with genetic similarity or ancestry. Nevertheless, race and ethnicity are well-established predictors of health outcomes and access to care. However, a multitude of factors that are both correlated with and are independent of race and ethnicity may affect group differences in health and health care. Race and ethnicity are only one of many elements considered to be social determinants of health—nonmedical factors that influence health outcomes and are known to have a significant relationship with these disparities [46,47]. Teasing out which factors influence health outcomes is challenging [48], and issues with data quality and inappropriate or poorly applied standards around race and ethnicity can greatly lessen our understanding of health disparities [17].

Though there are some limitations to the racial representation in this data set, it nevertheless remains a unique resource for COVID-19 research on racial disparities. COVID-19 has served to emphasize the deadliness of these disparities and has made social conditions far worse for many Black, Hispanic, and American Indian persons living in the United States. However, these inequities are not immutable. The COVID-19 pandemic provides an opportunity for clinicians, health systems, scientists, and policy makers to address social disparities and thereby improve the health and well-being of all persons in the United States for both known and future illnesses.

Databases such as N3C spur discovery by collecting and centralizing clinical data, making national, centralized data sets available to researchers. Although intended to increase the accessibility of data, governance can paradoxically create further restrictions. Centralization efforts require that data be transformed numerous times, and differences in how race and ethnicity are conceptualized, documented, and encoded by health care institutions affect the quality of the harmonized data. Across the full data life cycle, more transparency about these numerous decisions is critical if researchers are to make accurate inferences from analyses. Careful and systematic analyses are important to better guide clinical care and public policy but also to inform iterative improvement of collection and harmonization across the EHR data life cycle.

Acknowledgments

The members of the N3C Consortium are as follows: Melissa A Haendel, Christopher G Chute, Anita Walden, Jeremy Richard Harper, and Amit Saha.

The analyses described here were conducted with data and tools accessed through the National Center for Advancing Translational Science (NCATS) National Covid Cohort Collaborative (N3C) Data Enclave [9] and supported by NCATS U24TR002306. This research was possible because of the patients whose information is included within the data and the organizations and scientists who have contributed to the ongoing development of this community resource. LC has received support from the National Library of Medicine under Award Number T15LM007088. NGW has received funding under National Library of Medicine R21LM013645 and U24TR002306 from the NCATS of the US National Institutes of Health (NIH). JE has received funding from grants UL1TR001855 and UL1TR000130 from the NCATS of the US NIH. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The authors would like to acknowledge the assistance of Siegfried Gold, the N3C Data Liaison to the Social Determinants of Health domain team, and Adam Wilcox, who provided encouragement and guidance on the initial structure and goals of this work.

Conflicts of Interest

None declared.

References

1. COVID-19 Dashboard. Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. URL: https://coronavirus.jhu.edu/map.html [accessed 2021-08-07]
2. Mude W, Oguoma VM, Nyanhanda T, Mwanri L, Njue C. Racial disparities in COVID-19 pandemic cases, hospitalisations, and deaths: A systematic review and meta-analysis. J Glob Health 2021 Jul 26;11:05015 [FREE Full text] [doi: 10.7189/jogh.11.05015] [Medline: 34221360]
3. Gold JA, Rossen LM, Ahmad FB, Sutton P, Li Z, Salvatore PP, et al. Race, ethnicity, and age trends in persons who died from COVID-19 - United States, May-August 2020. MMWR Morb Mortal Wkly Rep 2020 Oct 23;69(42):1517-1521 [FREE Full text] [doi: 10.15585/mmwr.mm6942e1] [Medline: 33090984]
4. Wortham J, Lee J, Althomsons S, Latash J, Davidson A, Guerra K. Characteristics of Persons Who Died with COVID-19 — United States, February 12–May 18, 2020. In: Cockerham WC, Cockerham GB, editors. The COVID-19 Reader: The Science and What It Says About the Social. Milton Park, Abingdon-on-Thames, Oxfordshire, England, UK: Routledge; 2021:152-164.
5. Hatcher SM, Agnew-Brune C, Anderson M, Zambrano LD, Rose CE, Jim MA, et al. COVID-19 among American Indian and Alaska Native persons - 23 states, January 31-July 3, 2020. MMWR Morb Mortal Wkly Rep 2020 Aug 28;69(34):1166-1169 [FREE Full text] [doi: 10.15585/mmwr.mm6934e1] [Medline: 32853193]
6. Dagliati A, Malovini A, Tibollo V, Bellazzi R. Health informatics and EHR to support clinical research in the COVID-19 pandemic: an overview. Brief Bioinform 2021 Mar 22;22(2):812-822 [FREE Full text] [doi: 10.1093/bib/bbaa418] [Medline: 33454728]
7. Ferguson C. It took a pandemic, but the US finally has (some) centralized medical data.: MIT Technology Review; 2021 Jun 21. URL: https://www.technologyreview.com/2021/06/21/1026590/us-covid-database-n3c-nih-privacy/ [accessed 2021-08-30]
8. Haendel M, Chute C, Bennett T, Eichmann D, Guinney J, Kibbe W, N3C Consortium. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. J Am Med Inform Assoc 2021 Mar 01;28(3):427-443 [FREE Full text] [doi: 10.1093/jamia/ocaa196] [Medline: 32805036]
9. National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH). National COVID Cohort Collaborative (N3C) Data Enclave. URL: https://covid.cd2h.org/ [accessed 2021-07-30]
10. Jirkovsky V, Obitko M. Semantic Heterogeneity Reduction for Big Data in Industrial Automation. 2014 Presented at: 14th Conference on Information Technologies – Applications and Theory (ITAT 2014); September 25-29, 2014; Jasna, Slovakia URL: http://ceur-ws.org/Vol-1214/
11. Haendel M, Chute C. Interoperability Standards Priorities Task Force 2021 Presentation on N3C. 2021 Apr 16. URL: https://www.healthit.gov/hitac/events/interoperability-standards-priorities-task-force-2021-4 [accessed 2022-02-12]
12. Hong S. N3C Data Ingestion and Harmonization Workstream Race Mapping Crosswalk.: Github; 2020 Jul 27. URL: https://github.com/National-COVID-Cohort-Collaborative/Data-Ingestion-and-Harmonization/blob/1f48cb945b186292aec07ce7e2751796967a3a83/ETLProcess/scripts/n3c_xwalk_mapping.sql [accessed 2021-08-22]
13. Institute of Medicine (US) Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care. In: Smedley BD, Stith AY, Nelson AR, editors. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. Washington, DC: National Academies Press (US); 2003.
14. Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity. Office of Management and Budget. 1997. URL: https://obamawhitehouse.archives.gov/omb/fedreg_1997standards [accessed 2022-07-31]

15. Race and Ethnicity Code Set Version 1.0. Centers for Disease Control and Prevention. 2000 Mar 01. URL: https://www.cdc.gov/phin/resources/vocabulary/documents/CDC-Race-Ethnicity-Background-and-Purpose.pdf [accessed 2020-01-20]
16. US Core Implementation Guide: 5.0.1 - STU5 Release US. HL7 International. 2019 May 21. URL: https://www.hl7.org/fhir/us/core/StructureDefinition-us-core-race.html [accessed 2021-08-23]
17. Subcommittee on Standardized Collection of Race/Ethnicity Data for Healthcare Quality Improvement Board on Health Care Services. Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement. Agency for Healthcare Research and Quality. 2009. URL: https://www.ahrq.gov/research/findings/final-reports/iomracereport/index.html [accessed 2022-07-31]
18. Polubriaginof F, Ryan P, Salmasian H, Shapiro A, Perotte A, Safford M, et al. Challenges with quality of race and ethnicity data in observational databases. J Am Med Inform Assoc 2019 Aug 01;26(8-9):730-736 [FREE Full text] [doi: 10.1093/jamia/ocz113] [Medline: 31365089]
19. Grafova I, Jarrín OF. Beyond Black and White: mapping misclassification of Medicare beneficiaries race and ethnicity. Med Care Res Rev 2020 Jul 07;78(5):616-626 [FREE Full text] [doi: 10.1177/1077558720935733] [Medline: 32633665]
20. Jarrín OF, Nyandege A, Grafova I, Dong X, Lin H. Validity of race and ethnicity codes in Medicare administrative data compared with gold-standard self-reported race collected during routine home health care visits. Med Care 2020 Jan;58(1):e1-e8 [FREE Full text] [doi: 10.1097/MLR.0000000000001216] [Medline: 31688554]
21. Klinger EV, Carlini SV, Gonzalez I, Hubert SS, Linder JA, Rigotti NA, et al. Accuracy of race, ethnicity, and language preference in an electronic health record. J Gen Intern Med 2015 Jul 20;30(6):719-723 [FREE Full text] [doi: 10.1007/s11606-014-3102-8] [Medline: 25527336]
22. Cook LA, Sachs J, Weiskopf NG. The quality of social determinants data in the electronic health record: a systematic review. J Am Med Inform Assoc 2021 Dec 28;29(1):187-196 [FREE Full text] [doi: 10.1093/jamia/ocab199] [Medline: 34664641]
23. Pellegrin KL, Miyamura JB, Ma C, Taniguchi R. Improving accuracy and relevance of race/ethnicity data: results of a statewide collaboration in Hawaii. J Healthc Qual 2016;38(5):314-321. [doi: 10.1097/01.jhq.0000462679.40135.76]
24. Madlock-Brown C, Reynolds RB. Identifying obesity-related multimorbidity combinations in the United States. Clin Obes 2019 Dec 15;9(6):e12336. [doi: 10.1111/cob.12336] [Medline: 31418172]
25. Centers for Disease Control and Prevention. Demographic trends of COVID-19 cases and deaths in the US reported to the CDC. COVID Data Tracker. URL: https://covid.cdc.gov/covid-data-tracker [accessed 2021-08-26]
26. Health Disparities: Provisional Death Counts for Coronavirus Disease 2019 (COVID-19). Centers for Disease Control and Prevention National Center for Health Statistics. URL: https://www.cdc.gov/nchs/nvss/vsrr/covid19/health_disparities.htm [accessed 2021-08-26]
27. National Institutes of Health (NIH) Tribal Consultation on the National COVID Cohort Collaborative (N3C): Framing Letter. National Center for Advancing Translational Sciences. 2022 Feb 11. URL: https://ncats.nih.gov/files/Framing-Letter-N3C-508.pdf [accessed 2022-07-30]
28. Rutter JL. Dear Tribal Leader and Urban Indian Organization Leader Letter. National Center for Advancing Translational Sciences. 2021 Dec 20. URL: https://ncats.nih.gov/files/Dear-Tribal-Leader-Letter-N3C-508.pdf [accessed 2022-07-31]
29. Common Data Model. PEDSnet. URL: https://pedsnet.org/data/common-data-model/ [accessed 2021-08-23]
30. Beaulieu B. Council of State and Territorial Epidemiologists. Centers for Disease Control and Prevention. 2022 Apr 01. URL: https://preparedness.cste.org/wp-content/uploads/2022/04/RaceEthnicityData_FINAL.pdf [accessed 2022-07-31]
31. Common Data Model (CDM) Specification, Version 6.0. PCORnet. URL: https://pcornet.org/wp-content/uploads/2022/01/PCORnet-Common-Data-Model-v60-2020_10_221.pdf [accessed 2022-07-31]
32. Compton E, Bentley M, Ennis S, Rastogi S. 2010 Census Race and Hispanic Origin Alternative Questionnaire Experiment. United States Census Bureau. 2013 Feb 28. URL: https://www.census.gov/content/dam/Census/library/publications/2013/dec/2010_cpex_211.pdf [accessed 2022-07-31]
33. Osman I, Urban Indian Health Institute. Best Practices for American Indian and Alaska Native Data Collection. Seattle Indian Health Board. 2020. URL: https://aipi.asu.edu/sites/default/files/best-practices-for-american-indian-and-alaska-native-data-collection.pdf [accessed 2022-07-31]
34. Tsosie RA. Tribal Data Governance and Informational Privacy: Constructing 'Indigenous Data Sovereignty'. Montana Law Review 2019 Sep 16;80(2):229-268 [FREE Full text]
35. Hull SC, Wilson (Diné) DR. Beyond Belmont: ensuring respect for AI/AN communities through Tribal IRBs, laws, and policies. Am J Bioeth 2017 Jul 29;17(7):60-62 [FREE Full text] [doi: 10.1080/15265161.2017.1328531] [Medline: 28661757]
36. Garrison NA, Barton KS, Porter KM, Mai T, Burke W, Carroll SR. Access and management: Indigenous perspectives on genomic data sharing. Ethn Dis 2019 Dec 12;29(Supp):659-668. [doi: 10.18865/ed.29.s3.659]
37. NIH Tribal Consultation Report: NIH Draft Policy for Data Management and Sharing. National Institutes of Health. 2020 Sep 24. URL: https://osp.od.nih.gov/wp-content/uploads/Tribal_Report_Final_508.pdf [accessed 2022-07-31]
38. Department of Health and Human Services. Tribal Consultation Policy. Washington, DC: Health Resources and Services Administration; 2010.
39. About the Topic of Race. United States Census Bureau. URL: https://www.census.gov/topics/population/race/about.html [accessed 2022-07-31]

40. Mays VM, Ponce NA, Washington DL, Cochran SD. Classification of race and ethnicity: implications for public health. Annu Rev Public Health 2003 Jan;24(1):83-110 [FREE Full text] [doi: 10.1146/annurev.publhealth.24.100901.140927] [Medline: 12668755]
41. Williams DR, Lavizzo-Mourey R, Warren RC. The concept of race and health status in America. Public Health Rep 1994;109(1):26-41 [FREE Full text] [Medline: 8303011]
42. Section 289a-2, PL 103-43. National Institutes of Health (NIH) Revitalization Act. 1993 Jun 10. URL: https://www.govinfo.gov/content/pkg/USCODE-2011-title42/pdf/USCODE-2011-title42-chap6A-subchapIII-partH-sec289a-2.pdf [accessed 2022-07-31]
43. Flores LE, Frontera WR, Andrasik MP, Del Rio C, Mondríguez-González A, Price SA, et al. Assessment of the inclusion of racial/ethnic minority, female, and older individuals in vaccine clinical trials. JAMA Netw Open 2021 Feb 01;4(2):e2037640 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.37640] [Medline: 33606033]
44. Landale NS, Oropesa RS. White, Black, or Puerto Rican? Racial self-identification among mainland and island Puerto Ricans. Social Forces 2002 Sep 01;81(1):231-254. [doi: 10.1353/sof.2002.0052]
45. Hahn RA. The state of federal health statistics on racial and ethnic groups. JAMA 1992 Jan 08;267(2):268-271. [doi: 10.1001/jama.1992.03480020078035]
46. Social determinants of health. World Health Organization (WHO). URL: https://www.who.int/health-topics/social-determinants-of-health [accessed 2021-08-23]
47. Social Determinants of Health: Healthy People 2030. Health.gov. URL: https://health.gov/healthypeople/objectives-and-data/social-determinants-health [accessed 2021-08-23]
48. Cottrell EK, Hendricks M, Dambrun K, Cowburn S, Pantell M, Gold R, et al. Comparison of community-level and patient-level social risk data in a network of community health centers. JAMA Netw Open 2020 Oct 01;3(10):e2016852-e2016852 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.16852] [Medline: 33119102]

Abbreviations

ACS: American Communities Survey
ACT: Accrual to Clinical Trials
AI/AN: American Indian or Alaska Native
CDC: Centers for Disease Control and Prevention
CHF: Cerner HealthFacts
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
HL7: Health Level Seven International
N3C: National COVID Cohort Collaborative
NCATS: National Center for Advancing Translational Sciences
NIH: National Institutes of Health
OMB: Office of Management and Budget
OMOP: Observational Medical Outcomes Partnership
PCORnet: National Patient-Centered Clinical Research Network

Edited by C Lovis, J Hefner; submitted 09.05.22; peer-reviewed by M Boland, I Iyamu; comments to author 31.05.22; revised version received 21.06.22; accepted 26.07.22; published 06.09.22.

Please cite as:
Cook L, Espinoza J, Weiskopf NG, Mathews N, Dorr DA, Gonzales KL, Wilcox A, Madlock-Brown C, N3C Consortium
Issues With Variability in Electronic Health Record Data About Race and Ethnicity: Descriptive Analysis of the National COVID Cohort Collaborative Data Enclave
JMIR Med Inform 2022;10(9):e39235
URL: https://medinform.jmir.org/2022/9/e39235
doi:10.2196/39235
PMID:35917481

Original Paper

Use of a Semiautomatic Text Message System to Improve Satisfaction With Wait Time in the Adult Emergency Department: Cross-sectional Survey Study

Frederic Ehrler[1,2], PhD; Jessica Rochat[1,2], MSc; Johan N Siebert[2,3], MD; Idris Guessous[4,5], MD, PhD; Christian Lovis[1,2], MPH, MD; Hervé Spechbach[2,6], MD

[1]Division of Medical Information Sciences, University Hospitals of Geneva, Geneva, Switzerland

[2]Faculty of Medicine, University of Geneva, Geneva, Switzerland

[3]Department of Pediatric Emergency Medicine, Geneva Children’s Hospital, Geneva University Hospitals, Geneva, Switzerland

[4]Division of Primary Care, Geneva University Hospitals, Geneva, Switzerland

[5]Department of Health and Community Medicine, Faculty of Medicine, University of Geneva, Geneva, Switzerland

[6]Ambulatory Emergency Care Unit, Department of Primary Care Medicine, Geneva University Hospitals, Geneva, Switzerland

Corresponding Author:
Frederic Ehrler, PhD
Division of Medical Information Sciences
University Hospitals of Geneva
Gabrielle Perret Gentil 4
Geneva, 1205
Switzerland
Phone: 41 79 553 16 03
Email: frederic.ehrler@hcuge.ch

Abstract

Background: Many factors influence patient satisfaction during an emergency department (ED) visit, but the perception of wait time plays a central role. A long wait time in the waiting room increases the risk of hospital-acquired infection, as well as the risk of a patient leaving before being seen by a physician, particularly those with a lower level of urgency who may have to wait for a longer time.

Objective: We aimed to improve the perception of wait time through the implementation of a semiautomatic SMS text message system that allows patients to wait outside the hospital and facilitates the recall of patients closer to the scheduled time of meeting with the physician.

Methods: We performed a cross-sectional survey to evaluate the system using a tailored questionnaire to assess the patient perspective and the Unified Theory of Acceptance and Use of Technology questionnaire for the caregiver perspective. We also monitored the frequency of system use with logs.

Results: A total of 110 usable responses were collected (100 patients and 10 caregivers). Findings revealed that 97 of 100 (97%) patients were satisfied, with most patients waiting outside the ED but inside the hospital. The caregiver evaluation showed that it was very easy to use, but the adoption of the system was more problematic because of the perceived additional workload associated with its use.

Conclusions: Although not suitable for all patients, our system allows those who have a low-severity condition to wait outside the waiting room and to be recalled according to the dedicated time defined in the Swiss Emergency Triage Scale. It not only has the potential to reduce the risk of hospital-acquired infection but also can enhance the patient experience; additionally, it was perceived as a real improvement. Further automation of the system needs to be explored to reduce caregiver workload and increase its use.

(JMIR Med Inform 2022;10(9):e34488) doi:10.2196/34488

KEYWORDS

emergency; patient satisfaction; service-oriented health care; quality of care; health service; emergency department

Introduction

Background

Patients triaged with low priority in the emergency department (ED) are likely to have a long wait before being seen by a physician, as those with life-threatening and serious conditions are prioritized over patients that are less acute [1]. A side effect of long wait times is the risk that patients leave the ED without being seen by a physician, with this risk increasing significantly after a 1-hour wait [2]. It has also been shown that long wait times can result in staff interruptions by frustrated patients and lead to violent behavior [3,4]. Additionally, it has been reported that a long wait time increases the risk of contracting hospital-acquired infections [5]. As an example, Beggs et al [6] showed that the number of new cases of airborne infections increased substantially with time spent in the waiting room and the number of people waiting. However, reducing overcrowding in the ED waiting room is not a simple task [7]. The space available is often limited and the nature of the ED does not allow for a control on its occupation, which varies significantly over the course of a single day [8,9].

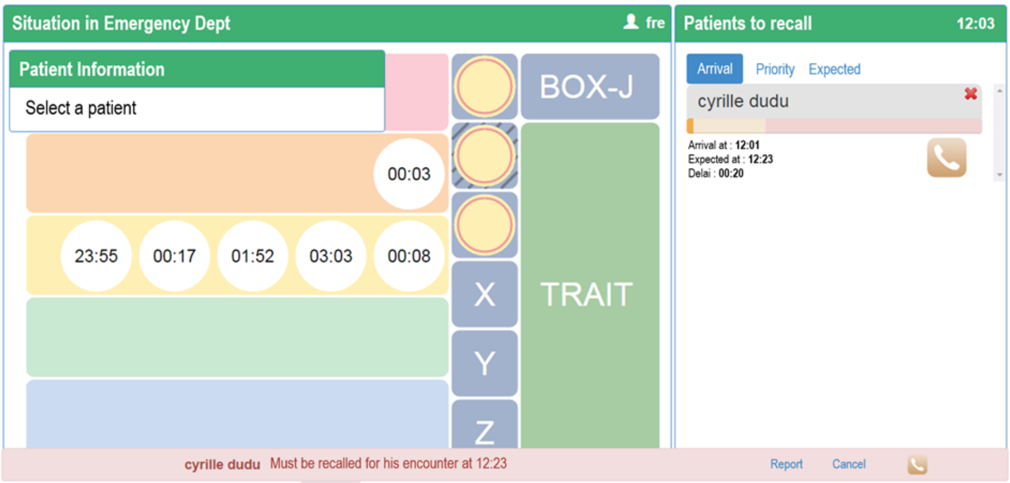
Several attempts have been performed to improve the wait time experience in the ED, either by minimizing the duration between triage and patient care or by acting on the actual perception of wait time [7,10]. Although organizational measures can improve ED efficiency, such as fast track [11], improved triage [12], and better team communication, they will never prevent overcrowding situations, as ED staff cannot be adjusted as quickly as the influx evolves. By contrast, improving the patient experience during the wait has been favored through interventions such as providing information to the patient about the expected wait duration [13], comfort improvement in the waiting room [14], or giving a pager to patients, which allowed them to wait in a place other than the waiting room [15]. These interventions were shown to have a positive impact and are promising strategies to be further explored.

To reduce ED congestion and improve the perception of wait time, we developed a semiautomated message (SMS text message) system that allows patients to wait outside the emergency waiting room and to be recalled closer to the actual time of the medical consultation. In this study, we explore the perceptions of this system by patients and caregivers.

A Semiautomatic SMS Text Message System

The system was initially developed at the pediatric department of Geneva University Hospitals (Geneva, Switzerland) [16], adapted later for the adult setting and deployed in September 2017, and finally introduced in the gynecology and obstetrics setting in 2019. It aims at improving patient flow in EDs by providing caregivers with a system to monitor the flow and ED occupancy. The system allows triaged patients with a low-severity grade to wait outside the ED and to be called back by a recall SMS text message system shortly before they are to be seen by a physician. A screen available to nurses provides real-time occupancy of the emergency rooms and wait times by triage level (Figure 1).

Figure 1. Main screen of the SMS text message recall system. The left-hand side represents the waiting queue in the emergency department waiting room, with each line representing the emergency level and each circle a patient currently waiting. The vertical middle row represents the emergency rooms and their occupancy, with each patient also represented by a circle. The right-hand side is the SMS text message recall system. The patients enrolled are presented with information on their arrival time and expected meeting time with the physician.



cyrille dudu

Arrival at : 12:01
Expected at : 12:23
Delai : 00:20

cyrille dudu Must be recalled for his encounter at 12:23

Once triaged, each patient can be registered in the SMS text message system by a nurse. A screen displays the patient's key administrative information, allowing the administrative clerk to verify the validity of the patient's telephone number. The nurse estimates the length of the wait and validates the patient's registration in the system. The patient then receives a confirmation message and can leave the ED while remaining virtually in the queue. Whether they are physically present or not, all patients are moved forward normally in the queue and recalled based on their arrival time and emergency level. All registered patients waiting outside the ED are visible on a screen with a time bar individually associated with them and showing the expected time to being seen by a physician. The time bar progressively changes color based on the elapsed time and actions that need to be taken by the caregiver responsible for calling the patient back. A green bar indicates that no recall is needed yet since the meeting with the physician is still distant. The bar turns orange 20 minutes before the patient's scheduled return, suggesting that the triage nurse call the patient back, without being mandatory. If the scheduled return time has passed, the bar turns red.

Dispatch of the first recall SMS text message is left to the discretion of the triage nurse to determine the most opportune time to return for the visit. If the patient does not arrive within 20 minutes of the first SMS text message, the system is automated to send reminder messages every 20 minutes (total of four SMS text messages). At any time, the nurse has the possibility to inform the patient about the evolution of the situation at the ED by sending a predefined message to the patient. Depending on the situation, the message sent will inform the patient that the visit is postponed due to a strong influx of patients or that an early return is possible due to an improved situation. If the patients do not arrive after three reminder SMS text messages, a final fourth SMS text message is sent to inform them that the position in the queue is no longer guaranteed, but the visit is still possible.

Methods

Study Design

This study is a cross-sectional descriptive investigation using a mixed methods methodology, including an assessment of the patients' experience during their wait in the ED through a tailor-made questionnaire, analysis of the system log to understand the use trend, and an assessment of nurses' acceptance of the SMS text message recall system. The survey was conducted between March 13 and April 28, 2017, at the 24-hour ED outpatient unit at Geneva University Hospitals, the largest public hospital in Switzerland with 70,000 patient admissions each year. Utilization logs were collected from October 2017 to August 2019.

ED Setting: Emergency Outpatient Unit

Medical and traumatic pathologies are treated in 12 consultation rooms. Patients wait in a semienclosed waiting room with seating, a television, water, and newspapers. The staff (clinicians and nurses) are the same for the entire unit. The median length of stay is 3.5 hours, with a median waiting time of 1.5 hours.

When patients arrive at the main ED entrance, they are first seen by a triage nurse who decides whether the patient is a candidate for the outpatient unit, based on the Swiss Emergency Triage Scale [17]. Level 1 is a life-/limb-threatening situation where the patient must be seen immediately by a physician. Level 2 must be seen within 20 minutes, level 3 within 120 minutes, and level 4 is considered nonurgent. A total of 80% of patients who come to the ED are classified as level 3 and 10% as level 4. After triage, the patient goes through an administrative registration process and is then directed to one of the subunits by following colored lines on the floor. These lead to a nurse's desk where a nurse escorts the patient to the waiting room. Whenever possible, the nurses inform patients of the estimated waiting time. As soon as a consultation room and physician are available, the patient is taken to the room by the nurse. After the medical visit, the patient can either go home or may have to undergo an additional examination and return to the waiting room. A small percentage of patients (8%) are hospitalized and 5% leave the unit without being seen by a physician [18]. The probability of leaving the ED prematurely is linked to flow concern as demonstrated in previous studies [19].

Study Participants

Patients presenting to the ED outpatient unit with a triage level of 3 and 4 (according to the 4-level Swiss Triage Scale) were invited to participate in the questionnaire part of the study if they were 16 years of age or older and spoke French. We used a convenience sampling method and arbitrarily defined the sample size as 100 participants. Exclusion criteria were patients not capable of discernment (eg, unconscious, intoxicated, extreme trauma, or cognitive impairment), unable to read/understand French, vision problems, severe pain or overly aggressive, and those who had already completed the questionnaire.

Measurement Instruments

Patient Satisfaction Questionnaire

A 12-item questionnaire was designed to assess the patient experience among those who had used the SMS text message system. This questionnaire was of our own design. It contained an item aiming to determine the minimum expected wait time before patients find the system useful. It also explored where the patient waited until being taken care of, whether the advertised waiting time matched the actual waiting time, and whether the content of the SMS text message was clear. Users were asked if they felt stressed during the wait, if they had enough time to come back to the emergency room, and if they were satisfied with the system overall. In addition, the actual wait time for each patient who completed the questionnaire was extracted from the hospital clinical information system.

Caregiver Acceptance Questionnaire

The 21-item Unified Theory of Acceptance and Use of Technology (UTAUT) questionnaire is a unified technology acceptance model formulated by Venkatesh et al [20] as a conceptual framework to understand users' intended use and acceptance of new information technologies, which can be determined by 5 constructs: (1) performance efficiency (4 questions), (2) effort expectancy (4 questions), (3) social influence (4 questions), (4) facilitating conditions (4 questions), and (5) behavioral intention to use the system in the future (3 questions). Each question is scored on a 7-point Likert scale. The questionnaire was distributed anonymously to all nurses working with the system.

System Use Logs

System use was assessed by analysis of the system use logs. A log, including a time stamp, was generated each time a caregiver entered a patient into the SMS text message system, as well as each time a SMS text message was sent.

Procedure and Ethical Considerations

The Geneva Institutional Ethics Committee approved the study protocol (Réq-2016-00555). Patient participation in the study was voluntary, and oral consent was obtained prior to the intervention. After verification of the inclusion criteria, the nurse asked the patients if they agreed to use the SMS text message recall system. Information about the study and confidentiality were given verbally. If accepted, the patients were allowed to wait wherever they wanted (ie, in or outside the ED). We did not verify where the patient waited as it would have been difficult to trace. We arbitrarily decided to set the number of questionnaires to be completed at 100.

Once back in the ED, the patient was immediately brought to a consultation room. The patient was given the study questionnaire by a nurse while waiting for the physician. The nurse remained available for any questions and to assist the patient in completing the questionnaire if necessary. Instructions were given to the medical staff to see the patients immediately after completion of the questionnaire. Once completed, questionnaires were collected by nurses and placed in a dedicated box in a secure room. Questionnaires were collected each morning by a scientific collaborator, and the responses were entered into an Excel (Microsoft Corporation) file. To link the questionnaire data to data extracted from the hospital clinical information system, we used a mapping file linking the questionnaire ID to the patient ID. Once all data were included in the Excel file, only the questionnaire ID was retained to ensure anonymous analysis.

Statistical Analysis

Descriptive statistics were generated to present the demographic and medical characteristics of participants. The difference of the average mean between each level of patient satisfaction was analyzed using an ANOVA performed on SPSS 26 (IBM Corp) software. The caregiver acceptance questionnaire was analyzed by computing the proportion of each response for a given item. UTAUT scores were reported as the average score given to all items of a given dimension for all participants. System logs were analyzed by looking at the number of SMS text messages sent each month during the observation period.

Results

Demographics

Patient questionnaires were distributed between March 13 and April 28, 2017, by a total of 20 nurses during two different shifts (7:30 AM to 4 PM and 3 PM to 11:30 PM). The total number of collected questionnaires was 100. One patient was excluded because he visited the unit twice during the study period. The questionnaire took an average of 10 minutes to complete. Baseline patient demographics and data related to the medical encounter are shown in Table 1. Of the 100 respondents, 87 (87%) were classified with an emergency level of 3, and 12 (12%) were classified in level 4. No patients were classified as levels 1 or 2 as these acuity triage levels require immediate care. The nurses' questionnaire was proposed to all nurses of the unit (n=25) but was only completed by 10 nurses.

Table 1. Demographics of participants and information on their medical encounter.

	Participants (N=100)
Age (years), mean (SD)	38 (14.75)
Sex, n (%)	
Male	60 (60)
Female	40 (40)
Triage level, n (%)	
3	87 (87)
4	12 (12)
Missing	1 (1)
Wait time (hours), n (%)	
<1	32 (32)
1-2	45 (45)
2-3	14 (14)
3-4	8 (8)
>4	1 (1)

Patient Satisfaction Questionnaire

As presented in Table 2, of the total 100 respondents, 97% (n=97) were satisfied with the SMS text message system. Among these, approximately 75% (n=75) were totally satisfied with their waiting time and 56% (n=56) were satisfied. A total of 79 (79%) respondents waited outside of the ED but inside the hospital, as the facility offers the possibility to wait in pleasant places such as the cafeteria, adjacent green spaces, and the meditation room. The fact that patients waited close to the ED was confirmed by the fact that 86% (n=86) of patients returned to the ED on foot. Therefore, 92 (92%) patients had sufficient time to return to the ED once recalled. A total of 95 (95%) patients considered the SMS text message to be clear and 72 (72%) did not feel particularly stressed waiting outside the ED.

Table 2. Questionnaire results.

	Participants (N=100), n (%)
Where did you spend your time while waiting?	
At home	2 (2)
Outside the hospital	13 (13)
Inside the hospital	80 (80)
Other	6 (6)
How do you rate your actual wait time compared to the wait time announced by the nurses?	
Longer	25 (25)
Shorter	49 (49)
Equal	25 (25)
Not informed	1 (1)
The SMS text message content was clearly understandable?	
Totally agree	72 (72)
Partly agree	23 (23)
Neither agree nor disagree	4 (4)
Partly disagree	1 (1)
Totally disagree	0 (0)
Did you experience a feeling of stress linked to your absence from the emergency waiting room?	
Totally agree	8 (8)
Partly agree	10 (10)
Neither agree nor disagree	11 (11)
Partly disagree	22 (22)
Totally disagree	50 (50)
Did you have enough time to return to the emergency room after receiving the recall message?	
Totally agree	59 (59)
Partly agree	33 (33)
Neither agree nor disagree	4 (4)
Partly disagree	2 (2)
Totally disagree	0 (0)
How did you return to the emergency room after receiving the recall message?	
On foot	86 (86)
Public transport	8 (8)
Private transport	2 (2)
Other	4 (4)
Are you satisfied with the SMS text message recall service?	
Totally agree	75 (75)
Partly agree	22 (22)
Neither agree nor disagree	3 (3)
Partly disagree	0 (0)
Totally disagree	0 (0)
Were you satisfied with your waiting time?	
Totally agree	28 (28)
Partly agree	28 (28)

	Participants (N=100), n (%)
Neither agree nor disagree	20 (20)
Partly disagree	12 (12)
Totally disagree	11 (11)

By asking patients what would be the minimum duration of expected wait that would trigger an interest to be enrolled in the system in a future encounter (Table 3), we found that 45 of the 100 (45%) patients were interested in the system regardless of the waiting time. After 30 minutes of expected waiting time, 75 (75%) patients were interested in the system, and 87 (87%) patients were interested after 1 hour.

Table 3. Patients interested in using the SMS text message system after n minutes.

Number of minutes	Patients, n (%)
0	45 (45)
10	51 (51)
20	63 (63)
30	75 (75)
40	75 (75)
50	75 (75)
60	87 (87)
70	87 (87)
80	88 (88)
90	93 (93)
100	93 (93)
110	93 (93)
120	100 (100)

Satisfaction and Waiting Time

To determine whether wait time duration influenced the level of patient satisfaction with wait time, we assessed if the differences in mean wait time across the five wait time satisfaction modalities (ie, totally disagree, partly disagree, neither agree nor disagree, partly agree, and totally agree) were statistically significant (descriptive statistics are presented in Table 4). As the homogeneity of variance using Levene was not statistically significant (*P*=.42), meaning that the variances were equal across groups, an ANOVA was performed. We found no significant differences between wait time means as a function of wait time satisfaction (*P*=.32; F_4=1.193).

Table 4. Average wait duration according to user satisfaction with wait time.

Satisfaction with wait time	Wait time (min), mean (SD)	Participants (N=100), n (%)
Totally disagree	86.9091 (64.28)	11 (11)
Partly disagree	105.0833 (58.78)	12 (12)
Neither agree nor disagree	101.5000 (58.06)	20 (20)
Partly agree	98.3929 (47.88)	28 (28)
Totally agree	75.0357 (46.22)	28 (28)
Total	91.9495 (53.10)	99 (99)

Caregiver Acceptance Questionnaire

The UTAUT questionnaire distributed to all nurses using the system was completed by 10 nurses (20% participation rate; Table 5). Nurses emphasized the good ergonomics of the system as they rated effort expectancy with an average score of 6.0. This was also confirmed by the facilitating condition dimension, including the resources and knowledge necessary to use the system, which were ranked above 5. Behavioral intention was high as most users intended to use the system frequently in the future on a daily basis. The expected gain on performance was less obvious for respondents. Although most users found the system useful (mean 4.5, SD 1.9), they did not find that the system increased their productivity (mean 3.2, SD 1.6) or speed at work (mean 3.0, SD 1.4). Hedonic motivation ranked below 4 as users did not find the system enjoyable or fun to use.

Finally, social influence scored the lowest (mean 2.3, SD 1.9) as all users did not observe a positive influence on their peers or hierarchy toward the use of the system.

Table 5. Score distribution for each UTAUT dimension.

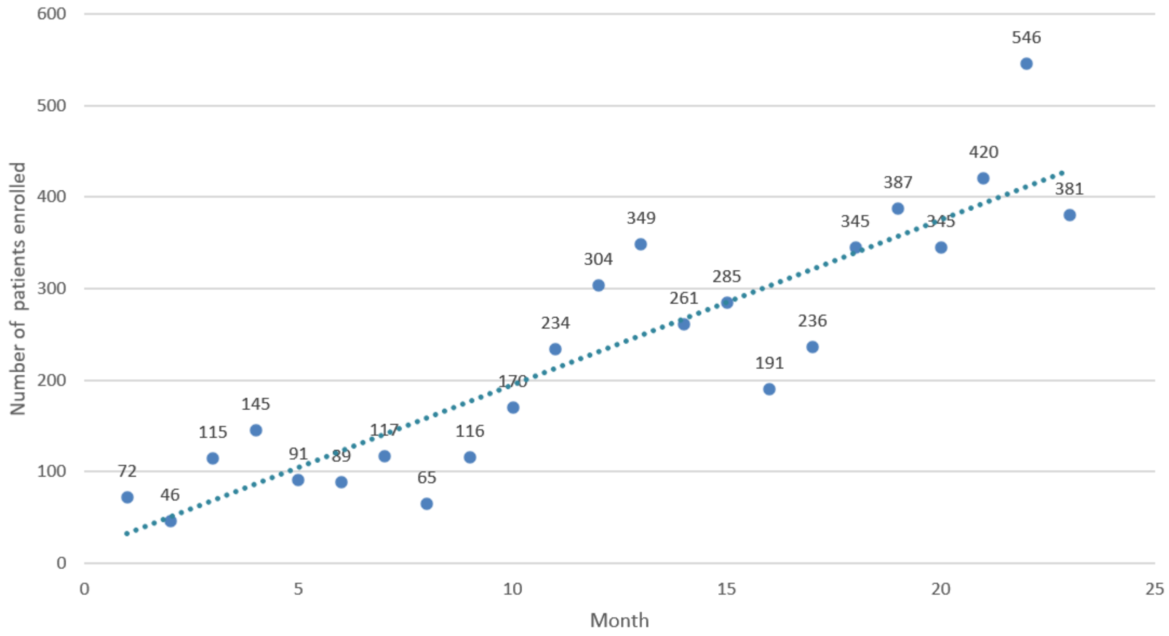
UTAUT[a] dimension	Nurses' scores, n (%)							Score, mean (SD)
	1	2	3	4	5	6	7	
Performance expectancy (n=33)	1 (3)	13 (39)	4 (12)	4 (12)	5 (15)	4 (12)	2 (6)	3.6 (1.8)
Effort expectancy (n=44)	0 (0)	3 (7)	0 (0)	3 (7)	6 (14)	9 (20)	23 (52)	6.0 (1.4)
Social influence (n=19)	12 (63)	0 (0)	0 (0)	4 (21)	2 (11)	0 (0)	1 (5)	2.3 (1.9)
Facilitating condition (n=41)	1 (2)	3 (7)	1 (2)	6 (15)	11 (27)	6 (15)	13 (32)	5.3 (1.6)
Hedonic motivation (n=24)	5 (21)	5 (21)	3 (13)	4 (17)	2 (8)	2 (8)	3 (13)	3.5 (2.1)
Behavioral intention (n=30)	0 (0)	2 (7)	3 (10)	4 (13)	6 (20)	6 (20)	9 (30)	5.3 (1.6)

[a] UTAUT: Unified Theory of Acceptance and Use of Technology.

Log Analysis

Figure 2 shows the number of unique patients entered into the SMS text message system from its introduction on October 1, 2017, to August 31, 2019. Although not always continuous, there was a clear trend of an increase in system use over time, ranging from 46 patients enrolled in November 2017 to 546 in July 2019. This corresponds to the trend of a linear function (18*x + 14) meaning that each month 18 additional patients are included in the system.

Figure 2. Number of unique patients enrolled in the SMS text message system each month (October 1, 2017, to August 31, 2019) and linear trend.



Discussion

Principal Findings

In this study, we found that 87 of the 100 (87%) patients with low-to-moderate urgency were interested in waiting outside the ED waiting room when the expected wait time was 1 hour or more. In a previous study, we observed that patients perceived the wait to be acceptable if it did not exceed 1 hour [21]. After 2 hours, they preferred to leave the ED before seeing a physician [22]. We observed that waiting outside the emergency room was perceived as a source of stress for <20% of participants, possibly related to the perceived reduced control over the situation when outside the room. Indeed, patients waiting outside the waiting room have no view on the current situation and can easily imagine being forgotten by ED staff [23]. Patients may also be concerned that their condition may worsen [24]. Thus, it may be worthwhile to send a recall SMS text message at regular intervals to indicate the patient's current place in the waiting queue to provide reassurance about their position and the progression of the ED process [25]. The messages could also inform the patient about the proper actions in case their condition worsens, such as approaching a specific person or to go to a specific desk to advise staff. This type of concern has already been highlighted in another report showing that some patients want to remain visible to the caregiver to avoid being forgotten [26].

By comparing the relationship between wait time and satisfaction with our previous study [21] performed in a similar setting, we observed a reduced negative influence of the average wait time on patient satisfaction. Whereas in previous studies longer wait times led to significantly less satisfied patients, this relationship was no longer observed in our study [27]. This may indicate that patients are less concerned about the length of wait if they can wait in another location than in a waiting room where they have little to do but remain seated until they are taken care of. This correlates well with our results indicating that most patients were willing to use the system if the wait was longer than 1 hour.

Use of the system by the nursing staff began at a low frequency but increased steadily over time. Nurses' initial reaction to the system was negative or neutral, and they initially perceived the tool as an additional burden to their workload. This phenomenon has already been observed in other studies [28,29]. The use of the system by many patients allowed it to predict potential benefits of the tool, such as reduced interruptions due to inpatient patients and reduced aggressive behavior in the waiting room due to long wait times [30]. However, informal feedback from nurses using the system highlighted the difficulty of using it when the ED was crowded. This is probably because busy nurses have less time to use the system in addition to regular duties that results in a contradictory effect that prevents the system from being used when it would be most useful. There are two solutions that can be considered to deal with this problem. Either the system can be used by administrative staff or the system can be automated. At our institution, the drive to develop this system has been a top-down process, and we plan to employ administrative workers to offload these tasks from nursing staff.

Limitations

A strong limitation of the paper is the absence of a strict control group. To compare the effect of our intervention on the relationship between wait time and patient satisfaction, we used the results of a previous study [21] conducted in the same setting where we explored the factors associated with wait perception as a preintervention finding. However, since the questionnaire was not the same, the comparison is limited. The selection of patients based on their interest in using the SMS text message system must be taken into account as it certainly has an impact on the high satisfaction rate, as well as on the low-stress rate related to a wait outside the ED. Indeed, a patient with a high-stress level could refuse to use the system.

Another limitation is the use of a questionnaire of our own design. Since the questionnaire has not been scientifically validated, we cannot guarantee that it measured accurately the investigated constructs. Unfortunately, we did not record the acceptance rate of the system, and it would have been interesting to see how many patients refused the system and preferred to stay in the waiting room. The low participation rate of nurses is also a limitation, and it will be useful to conduct a further survey following the training of administrative staff to take over tasks.

Conclusions

Waiting in the emergency waiting room is a source of frustration for the patient. In addition to the increase in aggressive attitudes in some patients when the ED waiting room is crowded, it also puts patients at risk of hospital-acquired infections. We observed a high level of satisfaction with our SMS text message recall system, allowing a wait outside the ED, but the adoption was more difficult among nurses. Relying on further automation of the system may be an interesting solution to reduce caregiver workload, but this must be done with caution given the high unpredictability of the ED waiting process.

Acknowledgments

We are grateful to Rosemary Sudan (freelance technical editor) for providing editorial assistance.

Authors' Contributions

All authors contributed to the conception, design, analysis, and interpretation of data for this work. All authors contributed to drafting, revising, and approving the final version of the manuscript.

Conflicts of Interest

JNS, FE, and CL have individual intellectual property rights on the SMS text message recall system and, as employees of Geneva University Hospitals, might receive indirect institutional reward in case of commercialization. The other authors have no conflicts of interest to declare.

References

1. Farrohknia N, Castrén M, Ehrenberg A, Lind L, Oredsson S, Jonsson H, et al. Emergency department triage scales and their components: a systematic review of the scientific evidence. Scand J Trauma Resusc Emerg Med 2011 Jun 30;19:42 [FREE Full text] [doi: 10.1186/1757-7241-19-42] [Medline: 21718476]
2. Li D, Brennan J, Kreshak A, Castillo E, Vilke G. Patients who leave the emergency department without being seen and their follow-up behavior: a retrospective descriptive analysis. J Emerg Med 2019 Jul;57(1):106-113. [doi: 10.1016/j.jemermed.2019.03.051] [Medline: 31078346]

3. Cohen EL, Wilkin HA, Tannebaum M, Plew MS, Haley LL. When patients are impatient: the communication strategies utilized by emergency department employees to manage patients frustrated by wait times. Health Commun 2013;28(3):275-285. [doi: 10.1080/10410236.2012.680948] [Medline: 22716025]
4. Morphet J, Griffiths D, Plummer V, Innes K, Fairhall R, Beattie J. At the crossroads of violence and aggression in the emergency department: perspectives of Australian emergency nurses. Aust Health Rev 2014 May;38(2):194-201. [doi: 10.1071/AH13189] [Medline: 24670224]
5. Quach C, McArthur M, McGeer A, Li L, Simor A, Dionne M, et al. Risk of infection following a visit to the emergency department: a cohort study. CMAJ 2012 Mar 06;184(4):E232-E239 [FREE Full text] [doi: 10.1503/cmaj.110372] [Medline: 22271915]
6. Beggs CB, Shepherd SJ, Kerr KG. Potential for airborne transmission of infection in the waiting areas of healthcare premises: stochastic analysis using a Monte Carlo model. BMC Infect Dis 2010 Aug 20;10:247 [FREE Full text] [doi: 10.1186/1471-2334-10-247] [Medline: 20727178]
7. Yen K, Gorelick MH. Strategies to improve flow in the pediatric emergency department. Pediatr Emerg Care 2007 Oct;23(10):745-9; quiz 750. [doi: 10.1097/PEC.0b013e3181568efe] [Medline: 18090112]
8. Otsuki H, Murakami Y, Fujino K, Matsumura K, Eguchi Y. Analysis of seasonal differences in emergency department attendance in Shiga Prefecture, Japan between 2007 and 2010. Acute Med Surg 2016 Apr;3(2):74-80. [doi: 10.1002/ams2.140] [Medline: 29123756]
9. Downing A, Wilson R. Temporal and demographic variations in attendance at accident and emergency departments. Emerg Med J 2002 Nov;19(6):531-535 [FREE Full text] [doi: 10.1136/emj.19.6.531] [Medline: 12421778]
10. Yuzeng S, Hui LL. Improving the wait time to triage at the emergency department. BMJ Open Qual 2020 Feb;9(1):e000708 [FREE Full text] [doi: 10.1136/bmjoq-2019-000708] [Medline: 32019749]
11. Hampers LC, Cha S, Gutglass DJ, Binns HJ, Krug SE. Fast track and the pediatric emergency department: resource utilization and patients outcomes. Acad Emerg Med 1999 Nov;6(11):1153-1159 [FREE Full text] [doi: 10.1111/j.1553-2712.1999.tb00119.x] [Medline: 10569389]
12. Sun BC, Adams J, Orav EJ, Rucker DW, Brennan TA, Burstin HR. Determinants of patient satisfaction and willingness to return with emergency care. Ann Emerg Med 2000 May;35(5):426-434. [Medline: 10783404]
13. Xie B, Youash S. The effects of publishing emergency department wait time on patient utilization patterns in a community with two emergency department sites: a retrospective, quasi-experiment design. Int J Emerg Med 2011 Jun 14;4(1):29. [doi: 10.1186/1865-1380-4-29] [Medline: 21672236]
14. Goodarzi H, Javadzadeh H, Hassanpour K. Assessing the physical environment of emergency departments. Trauma Mon 2015 Nov;20(4):e23734 [FREE Full text] [doi: 10.5812/traumamon.23734] [Medline: 26839860]
15. Scolnik D, Matthews P, Caulfeild J, Williams C, Feldman B. Pagers in a busy paediatric emergency waiting room: a randomized controlled trial. Paediatr Child Health 2003 Sep;8(7):422-426 [FREE Full text] [doi: 10.1093/pch/8.7.422] [Medline: 20019948]
16. Ehrler F, Lovis C, Rochat J, Schneider F, Gervaix A, Galetto-Lacour A, et al. [InfoKids: changing the patients' journey paradigm in an Emergency Department]. Rev Med Suisse 2018 Sep 05;14(617):1538-1542. [Medline: 30226668]
17. Rutschmann O, Hugli O, Marti C, Grosgurin O, Geissbuhler A, Kossovsky M, et al. Reliability of the revised Swiss Emergency Triage Scale: a computer simulation study. Eur J Emerg Med 2018 Aug;25(4):264-269 [FREE Full text] [doi: 10.1097/MEJ.0000000000000449] [Medline: 28099182]
18. Grosgurin O, Cramer B, Schaller M, Sarasin F, Rutschmann O. Patients leaving the emergency department without being seen by a physician: a retrospective database analysis. Swiss Med Wkly 2013;143:w13889. [doi: 10.4414/smw.2013.13889] [Medline: 24317804]
19. Weiss SJ, Ernst AA, Derlet R, King R, Bair A, Nick TG. Relationship between the National ED Overcrowding Scale and the number of patients who leave without being seen in an academic ED. Am J Emerg Med 2005 May;23(3):288-294. [doi: 10.1016/j.ajem.2005.02.034] [Medline: 15915399]
20. Venkatesh V, Morris M, Davis G, Davis F. User Acceptance of Information Technology: toward a unified view. MIS Q 2003;27(3):425 [FREE Full text] [doi: 10.2307/30036540]
21. Spechbach H, Rochat J, Gaspoz J, Lovis C, Ehrler F. Patients' time perception in the waiting room of an ambulatory emergency unit: a cross-sectional study. BMC Emerg Med 2019 Aug 01;19(1):41 [FREE Full text] [doi: 10.1186/s12873-019-0254-1] [Medline: 31370794]
22. Shaikh S, Jerrard D, Witting M, Winters M, Brodeur M. How long are patients willing to wait in the emergency department before leaving without being seen? West J Emerg Med 2012 Dec;13(6):463-467 [FREE Full text] [doi: 10.5811/westjem.2012.3.6895] [Medline: 23359833]
23. Rochat CP, Gaucher N, Bailey B. Measuring anxiety in the pediatric emergency department. Pediatr Emerg Care 2018 Aug;34(8):558-563. [doi: 10.1097/PEC.0000000000001568] [Medline: 30020249]
24. Dubé L, Teng L, Hawkins J, Kaplow M. Emotions, the neglected side of patient-centered health care management: the case of emergency department patients waiting to see a physician. In: Advances in Health Care Management: Volume 3. Bingley, West Yorkshire: Emerald Publishing Limited; 2002:161-193.

25. Andrade CC, Devlin AS. Stress reduction in the hospital room: applying Ulrich's theory of supportive design. J Environ Psychol 2015 Mar;41:125-134. [doi: 10.1016/j.jenvp.2014.12.001]
26. Yoon J, Sonneveld M. Anxiety of patients in the waiting room of the emergency department. In: Proceedings of the fourth international conference on Tangible, embedded, and embodied interaction. 2010 Jan Presented at: TEI '10; January 24-27, 2010; Cambridge, MA p. 279-286. [doi: 10.1145/1709886.1709946]
27. Davenport P, O'Connor SJ, Szychowski J, Landry A, Hernandez S. The relationship between emergency department wait times and inpatient satisfaction. Health Mark Q 2017;34(2):97-112. [doi: 10.1080/07359683.2017.1307066] [Medline: 28467280]
28. Mareš J. Resistance of health personnel to changes in healthcare. Kontakt 2018 Oct 12;20(3):e262-e272. [doi: 10.1016/j.kontakt.2018.04.002]
29. Safi S, Thiessen T, Schmailzl KJ. Acceptance and resistance of new digital technologies in medicine: qualitative study. JMIR Res Protoc 2018 Dec 04;7(12):e11072 [FREE Full text] [doi: 10.2196/11072] [Medline: 30514693]
30. Efrat-Treister D, Moriah H, Rafaeli A. The effect of waiting on aggressive tendencies toward emergency department staff: Providing information can help but may also backfire. PLoS One 2020;15(1):e0227729 [FREE Full text] [doi: 10.1371/journal.pone.0227729] [Medline: 31995583]

Abbreviations

ED: emergency department
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by G Eysenbach; submitted 26.10.21; peer-reviewed by A Benis, R Rivera; comments to author 24.11.21; revised version received 14.01.22; accepted 29.04.22; published 06.09.22.

Please cite as:
Ehrler F, Rochat J, Siebert JN, Guessous I, Lovis C, Spechbach H
Use of a Semiautomatic Text Message System to Improve Satisfaction With Wait Time in the Adult Emergency Department: Cross-sectional Survey Study
JMIR Med Inform 2022;10(9):e34488
URL: https://medinform.jmir.org/2022/9/e34488
doi:10.2196/34488
PMID:36066921

Review

The Value of Electronic Health Records Since the Health Information Technology for Economic and Clinical Health Act: Systematic Review

Shikha Modi[1], MBA, PhD; Sue S Feldman[2], RN, MEd, PhD

[1]Department of Political Science, Auburn University, Auburn, AL, United States

[2]Department of Health Services Administration, University of Alabama at Birmingham, Birmingham, AL, United States

Corresponding Author:
Shikha Modi, MBA, PhD
Department of Political Science
Auburn University
7074 Haley Center
Auburn, AL, 36849
United States
Phone: 1 2563355796
Email: szs0308@auburn.edu

Abstract

Background: Electronic health records (EHRs) are the electronic records of patient health information created during ≥1 encounter in any health care setting. The Health Information Technology Act of 2009 has been a major driver of the adoption and implementation of EHRs in the United States. Given that the adoption of EHRs is a complex and expensive investment, a return on this investment is expected.

Objective: This literature review aims to focus on how the value of EHRs as an intervention is defined in relation to the elaboration of value into 2 different value outcome categories, financial and clinical outcomes, and to understand how EHRs contribute to these 2 value outcome categories.

Methods: This literature review was conducted using PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). The initial search of key terms, EHRs, values, financial outcomes, and clinical outcomes in 3 different databases yielded 971 articles, of which, after removing 410 (42.2%) duplicates, 561 (57.8%) were incorporated in the title and abstract screening. During the title and abstract screening phase, articles were excluded from further review phases if they met any of the following criteria: not relevant to the outcomes of interest, not relevant to EHRs, nonempirical, and non–peer reviewed. After the application of the exclusion criteria, 80 studies remained for a full-text review. After evaluating the full text of the residual 80 studies, 26 (33%) studies were excluded as they did not address the impact of EHR adoption on the outcomes of interest. Furthermore, 4 additional studies were discovered through manual reference searches and were added to the total, resulting in 58 studies for analysis. A qualitative analysis tool, ATLAS.ti. (version 8.2), was used to categorize and code the final 58 studies.

Results: The findings from the literature review indicated a combination of positive and negative impacts of EHRs on financial and clinical outcomes. Of the 58 studies surveyed for this review of the literature, 5 (9%) reported on the intersection of financial and clinical outcomes. To investigate this intersection further, the category "Value–Intersection of Financial and Clinical Outcomes" was generated. Approximately 80% (4/5) of these studies specified a positive association between EHR adoption and financial and clinical outcomes.

Conclusions: This review of the literature reports on the individual and collective value of EHRs from a financial and clinical outcomes perspective. The collective perspective examined the intersection of financial and clinical outcomes, suggesting a reversal of the current understanding of how IT investments could generate improvements in productivity, and prompted a new question to be asked about whether an increase in productivity could potentially lead to more IT investments.

(JMIR Med Inform 2022;10(9):e37283) doi:10.2196/37283

KEYWORDS

electronic health records; EHRs; value; financial outcomes; clinical outcomes; health informatics; clinical informatics

Introduction

Electronic health records (EHRs) are described as electronic records of patient health information created by ≥1 encounter in any health care setting and include patient demographics, issues, medication information, laboratory data, radiology reports, and history [1]. EHRs enable health information exchange, clinical decision support, diagnostic support, patient health portals, and more [2]. EHR use has the potential to improve the quality of care and patient safety [3] and has become an important part of the modern health system because of government policies, technology developments, health care challenges, and market situations [4]. The Health Information Technology for Economic and Clinical Health (HITECH) Act has been a major driver of the increase in the adoption and implementation of EHRs [5].

The HITECH Act of 2009 was passed to decrease health care costs, improve quality, and increase patient safety through incentives for providers (physicians) and organizations that provided proof of their meaningful use (MU) of certified EHR systems [5]. Approximately US $27 billion in incentives was given to physicians and hospitals that adopted and used EHRs according to federally defined “meaningful use” criteria [6]. Out of US $27 billion, US $406 million was allotted to Medicare Advantage Organizations for eligible providers. The Center for Medicare and Medicaid Services (CMS) provided subsidy payments of US $63,750 over 6 years for Medicaid or US $44,000 over 5 years for Medicare to individual physicians if they used certified EHRs beginning in 2011 and exhibited MU criteria [7]. It is worth noting that in 2018, the CMS refocused MU on increasing health information exchange and patient access to data, renaming MU as Promoting Interoperability Programs.

Given that it has been over a decade since the HITECH Act was passed, sufficient data are available to understand how EHR adoption investment adds value to the hospitals that have EHR systems in place. It is important to first define “value” to understand the value of EHR adoption from a comprehensive perspective.

When reviewing the cost and resources associated with EHR adoption, it is generally considered to be an expensive investment [8,9], with an expectation of a return or value on the investment. Typically, return on investment (ROI) is measured by dividing the net profit by the net investment [10]. ROI-related concerns about EHR adoption were considered to be a major barrier to the adoption of EHRs, primarily as the value was unknown [11]. Jang et al [9] calculated the ROI for EHR adoption by looking at the breakeven point of EHR adoption investment. This study focused on 17 community primary care practices targeting the financial aspect of EHR adoption but did not consider the financial aspect of multilayered decisions such as system selection, employee training, updating or maintaining systems, and training employees for updated systems [11].

Moving beyond ROI, value can be defined as “considering (someone or something) to be important or beneficial” [12]. To simplify this definition, anything that benefits or is important to an individual is considered to be valuable to that individual, regardless of it being an action or intervention. Value is defined in multiple ways within the health care industry. Payne et al [13] describe value as dollars (financial), productivity (clinical), or effectiveness (clinical). Payne et al [13] also suggest that health IT (HIT) literature is primarily focused on productivity (process) and effectiveness (outcome), followed by dollars (outcome). Feldman et al [14] explain value as a combination of tangible (dollars, financial) and intangible (doing the right thing; trust relationships, social) components. In terms of examining the EHR value component, another study analyzed the value of EHRs in terms of efficiency (clinical) and cost savings (financial). This study further used efficiency to derive value by looking at the quality of care and cost savings from better claims management and reduced payments [11]. Riskin et al [15] highlighted the national focus on health reform and defined its value in terms of improved outcomes (clinical) and reduced costs (financial). Yeung [16] discussed EHR in terms of value as it is connected to improving services (clinical) delivered at local health departments. Hepp et al [17] evaluated the value of EHRs by looking at EHRs as a cost-effective strategy to improve medication safety (clinical). Adler-Milstein et al [18] analyzed different scopes of the value of EHRs by gauging process adherence (clinical), patient satisfaction (clinical), and efficiency outcomes (clinical).

The environment in which HIT is used may have an impact on the value that is derived from HIT [19]. For example, Peterson et al [11] suggested that current users of EHR systems focus on value in terms of improving workflows and, as a result, better clinical outcomes, whereas local health departments or community clinics may focus on value in terms of capturing patient information to improve the services that are provided [16] or for ambulatory settings on increasing medication safety [17]. Thinking about EHRs’ value more holistically, the value could equate with increased revenue and reduced cost (financial). For patients, it could mean improved health and prevention of illness (outcomes); for providers, it could signal reduced errors and an increase in the efficiency of care (process); and for the government, it could correspond with improvements in population health through timely public health reporting and population well-being (process and outcomes) [13]

The World Health Organization defines an outcome measure as “a change in the health of an individual, group of people, or population that is attributable to an intervention or series of interventions” [20]. Outcomes, in the conventional health services sense, are usually regarded as clinical outcomes [21]; however, to represent the scope of the Triple Aim of health care, the authors built upon the literature to broaden the definition of outcomes to include financial and social outcomes, in addition to traditional clinical outcomes.

This review of the literature aimed to describe how the value of EHRs, as an intervention, is defined in relation to the elaboration of value into 2 different value outcome categories, financial and clinical outcomes, and by understanding the contributions that EHRs make to these 2 value outcome categories.

Methods

This review was conducted using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [22]. This method has been used for other qualitative analyses of literature and is therefore regarded as a suitable method for this qualitative systematic review of the literature [23,24]. To capture the multidisciplinary evidence in this field, the following databases were used to conduct the initial search: PubMed, Scopus, and Embase. To capture the decade that followed the enactment of the HITECH Act, the literature published in English between January 2009 and December 2019 was used as a filter to refine the results. The initial keywords used were "electronic health records," "EHR," "value," "financial outcomes," and "clinical outcomes." To ensure the comprehensiveness of the literature search, all the outcome categories were searched separately and in conjunction with one another. The search strings and gathered results were extensive and lengthy and are recorded in Table 1. To optimize the chance of finding relevant studies on the value of EHR from the financial and clinical outcomes perspective after the enactment of the HITECH Act, the following filters were applied to the searches: (1) keywords in the title or abstract, (2) published in English, (3) published in the United States only, and (4) published between 2009 and 2019, when applicable.

A total of 971 articles was included in the initial literature screening, of which, after removing 410 (42.2%) duplicates, 561 (57.8) were incorporated in the title and abstract screening. During the title and abstract screening phase, articles were excluded from further review phase if they met any of the following criteria: (1) not relevant to the outcomes of interest, (2) not relevant to EHRs, (3) nonempirical, and (4) non–peer reviewed. After the application of the exclusion criteria, 80 studies remained for a full-text review. After evaluating the full text of the residual 80 studies, 26 (33%) studies were excluded as they did not address the impact of EHR adoption on the outcomes of interest. Following this, 4 additional studies were discovered through manual reference searches and were added to the total, resulting in 58 studies for analysis. Figure 1 displays this process in a flow diagram. Both authors were involved in the article search, selection, and review process.

The 58 studies selected for inclusion are exhibited in the *Results* section and are organized by outcome category. ATLAS.ti (version 8.2), a qualitative data analysis tool, was used to categorize and code the final 58 studies. All studies were uploaded into ATLAS.ti as full-text documents with names that included the first author, year of publication, and article title. Qualitative data analysis software was deemed fitting for this type of analysis as it allows for the possibility of applying a recurring and reiterative approach to data analysis that is efficient and would have been difficult to replicate using a spreadsheet application [25].

The coding process began by analyzing each article to understand the context in relation to how each outcome category is defined in the literature and learn about the evaluation process of the impact of EHRs on these outcome categories. For this study, overarching a priori categories (financial outcomes and clinical outcomes) were used, and the studies were further categorized under these 2 overarching categories. Additional categories that were developed included the following:

- Financial outcomes: cost, revenue, profit margins, reimbursement, and return on assets
- Clinical outcomes: productivity, workflow efficiency, medical errors, patient safety, patient satisfaction, clinical volume, readmission rates, length of stay (LOS), and quality indicators at individual patient levels

Additional categories were added as necessitated throughout the coding and category generation process, which was part of the larger data analysis process. For example, introduction and gap categories were generated as they assisted in the writing of the introduction and gap and supplied context for this review of literature; however, quotations included in these categories did not necessarily factor into the results presented.

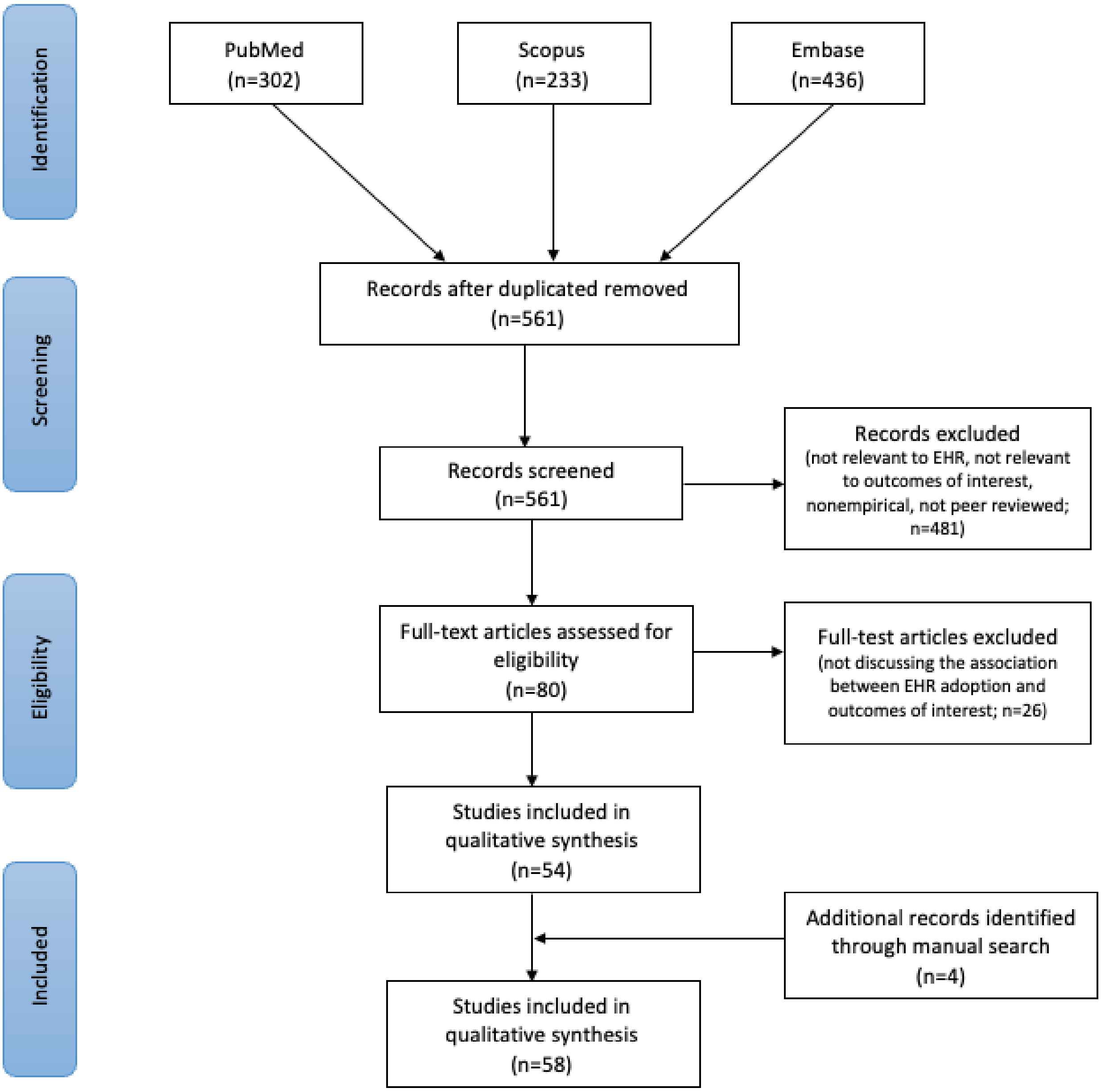
Table 1. Search strings from the literature search for the impact of electronic health records on financial and clinical outcomes (N=971).

Database and keywords	Results, n (%)	Filters	Results after applying filters, n (%)
PubMed			
([([([(((Finance*[Title] OR monetary[Title] OR economic*[Title] OR fiscal[Title] OR commercial[Title] OR cost[Title])) OR (Finance*[Other Term] OR monetary[Other Term] OR economic*[Other Term] OR fiscal[Other Term] OR cost[Other Term])) OR “Economics” [Mesh]]) OR ([(Clinical[Title] OR quality[Title] OR)] OR [Clinical[Other Term] OR quality[Other Term]]) AND ((((((Adopt*[Title] OR (Adopt*[Other Term]) OR implement*(Title)] OR implement*[Other Term])] AND [([(Follow-up-stud*[Title] OR prognos*[Title] OR predict*[Title] OR course[Title] OR followup-stud*[Title] OR efficacy[Title] OR complication[Title] OR chang*[Title] OR effective*[Title] OR evaluat*[Title] OR improve*[Title] OR indicat*[Title] OR impact*[Title] OR consequence*[Title] OR development*[Title] OR Result*[Title] OR outcome*[Title])] OR [Follow-up-stud*(Other Term) OR prognos*[Other Term] OR predict*(Other Term) OR course(Other Term) OR followup-stud*(Other Term) OR efficacy(Other Term) OR complication(Other Term) OR chang*(Other Term) OR effective*(Other Term) OR evaluat*(Other Term) OR improve*(Other Term) OR indicat*(Other Term) OR impact*(Other Term) OR consequence*(Other Term) OR development*(Other Term) OR Result*(Other Term) OR outcome*(Other Term)]) OR “follow-up studies” (mesh)]) AND ([([Electronic-health-record*(Title) OR electronic-medical-record*(Title) OR computerized-health-record*(Title) OR computerized-medical-record*(Title) OR EHR(Title) OR electronic-patient-record*(Title)]) OR (Electronic-health-record*[Other Term] OR electronic-medical-record*[Other Term] OR computerized-health-record*[Other Term] OR computerized-medical-record*[Other Term] OR EHR[Other Term] OR electronic-patient-record*[Other Term])] OR “electronic health records” [mesh])	193 (19.9)	Years: 2009-2019; language: English	179 (18.4)
([“electronic health records adoption”(Title/Abstract)] OR “EHR adoption”[Title/Abstract]) AND “financial outcomes”(Title/Abstract)	0 (0)	N/A[a]	0 (0)
([“electronic health records adoption”(Title/Abstract)] OR “EHR adoption”[Title/Abstract]) AND “financial”(Title/Abstract)	39 (4)	Years: 2009-2019; language: English	33 (3.4)
([“electronic health records adoption”(Title/Abstract)] OR “EHR adoption”[Title/Abstract]) AND “clinical outcomes”(Title/Abstract)	1 (0.1)	Years: 2009-2019; language: English	1 (0.1)
([“electronic health records adoption”(Title/Abstract)] OR “EHR adoption”[Title/Abstract]) AND “clinical”(Title/Abstract)	99 (10.2)	Years: 2009-2019; language: English	89 (9.2)
Scopus			
(TITLE-ABS-KEY [electronic-health-record* OR electronic-medical-record* OR computerized-health-record* OR computerized-medical-record* OR ehr OR electronic-patient-record* OR “electronic health record”] AND TITLE-ABS-KEY [finance* OR monetary OR economic* OR fiscal OR “economic”] AND TITLE-ABS-KEY [clinical OR quality] AND TITLE-ABS-KEY [“follow-cup studies” OR follow-up-stud* OR prognos* OR predict* chang* OR effective* OR evaluat* OR improve* OR indicat* OR impact* OR consequence* OR outcome*] AND TITLE-ABS-KEY [Adopt* OR implement*])	70 (7.2)	Years: 2009-2019; language: English; country: United States	35 (3.6)
TITLE-ABS-KEY (“EHR adoption” OR “electronic health records adoption” AND “financial outcomes”)	0 (0)	N/A	0 (0)
TITLE-ABS-KEY (“EHR adoption” OR “electronic health records adoption” AND “financial”)	61 (6.3)	Years: 2009-2019; language: English; country: United States	41 (4.2)
TITLE-ABS-KEY (“ehr adoption” OR “electronic health records adoption” AND “clinical outcomes”)	2 (0.2)	Years: 2009-2019; language: English; country: United States	2 (0.2)
TITLE-ABS-KEY (“EHR adoption” OR “electronic health records adoption” AND “clinical”)	173 (17.8)	Years: 2009-2019; language: English; country: United States	155 (16)
Embase			

Database and keywords	Results, n (%)	Filters	Results after applying filters, n (%)
(“electronic health record*”:ti,ab,kw OR “electronic medical record*”:ti,ab,kw OR “computerized health record*”:ti,ab,kw OR “computerized medical record*”:ti,ab,kw OR ehr:ti,ab,kw OR “electronic patient record*”:ti,ab,kw OR “electronic health record”:ti,ab,kw) AND (finance*:ti,ab,kw OR monetary:ti,ab,kw OR economic*:ti,ab,kw OR fiscal:ti,ab,kw OR “economic”:ti,ab,kw) AND (clinical:ti,ab,kw OR quality:ti,ab,kw) AND (“follow-up studies”:ti,ab,kw OR “follow up stud*”:ti,ab,kw OR prognos*:ti,ab,kw OR predict*:ti,ab,kw OR course:ti,ab,kw OR “followup stud*”:ti,ab,kw OR efficacy:ti,ab,kw OR complication:ti,ab,kw OR chang*:ti,ab,kw OR effective*:ti,ab,kw OR evaluat*:ti,ab,kw OR imptove*:ti,ab,kw OR indicat*:ti,ab,kw OR impact*:ti,ab,kw OR consequence*:ti,ab,kw OR development*:ti,ab,kw OR result*:ti,ab,kw OR outcome*:ti,ab,kw) AND (adopt*:ti,ab,kw OR implement*:ti,ab,kw)	350 (36)	Years: 2009-2019	303 (31.2)
(“electronic health records adoption”:ti,ab,kw OR “ehr adoption”:ti,ab,kw) AND “financial outcomes”:ti,ab,kw	0 (0)	N/A	0 (0)
(“electronic health records adoption”:ti,ab,kw OR “ehr adoption”:ti,ab,kw) AND “financial”:ti,ab,kw	42 (4.3)	Years: 2009-2019	35 (3.6)
(“electronic health records adoption”:ti,ab,kw OR “ehr adoption”:ti,ab,kw) AND “clinical outcomes”:ti,ab,kw	3 (0.3)	Years: 2009-2019	3 (0.3)
(“electronic health records adoption”:ti,ab,kw OR “ehr adoption”:ti,ab,kw) AND “clinical”:ti,ab,kw	104 (10.7)	Years: 2009-2019	95 (9.8)

[a]N/A: not applicable.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram [22]. EHR: electronic health record.


Identification
Screening
Eligibility
Included
PubMed (n=302)
Scopus (n=233)
Embase (n=436)
Records after duplicated removed (n=561)
Records screened (n=561)
Records excluded (not relevant to EHR, not relevant to outcomes of interest, nonempirical, not peer reviewed; n=481)
Full-text articles assessed for eligibility (n=80)
Full-test articles excluded (not discussing the association between EHR adoption and outcomes of interest; n=26)
Studies included in qualitative synthesis (n=54)
Additional records identified through manual search (n=4)
Studies included in qualitative synthesis (n=58)

Results

Information from the reviewed articles (n=58) was analyzed to ascertain how the value of EHRs is determined regarding financial and clinical outcomes relative to how they are defined earlier in this paper. In addition, findings from this review of the literature describe how EHR adoption affects each outcome category.

Financial Outcomes

Of the 58 studies reviewed, 21 (36%) studies incorporated segments that were coded under the "Value-Financial Outcomes" category. Different measures of financial outcomes were used in these studies, such as cost [26-29], revenue [28,29], profit margins [8,27], reimbursement [30], and return on assets [8]. These different financial outcome measures are described and detailed in Table 2. The included studies contained positive (17/58, 81%), negative (4/58, 19%), and no (3/58, 14%) association relationships between EHR adoption and financial outcomes. There were overlapping positive and negative impacts of EHR adoption on financial outcomes in some of the reviewed studies.

Table 2. Reviewed studies on the impact of EHR[a] adoption and financial and clinical outcomes.

Study	Journal or conference	Study period or data set	Objective	Outcome measures	Financial (n=21)	Clinical (n=54)
Adler-Milstein et al [18]	*Health Services Research*	AHA IT Supplement Survey (2008-2011), AHA Annual Survey (2009-2012), CMS[b] Hospital Compare data set (2009-2012), and CMS EHR Incentive Program Reports	To examine the relationship between EHR adoption and hospital outcomes	Efficiency (measured by the ratio of a hospital's total expenditures to adjusted patient days), process adherence, and patient satisfaction	✓[c]	✓
Appari et al [31]	*The American Journal of Managed Care*	Cross-sectional retrospective study, data on hospital patient safety performance (2008-2010) combined with IT systems data (2007; n=3002 nonfederal acute care hospitals)	To determine whether HIT[d] systems are associated with better patient safety in acute care settings	Adverse event indicators developed by AHRQ[e] (death among surgical patients with serious, treatable complications; collapsed lung that results from medical treatment [iatrogenic pneumothorax]; breathing failure after surgery [postoperative respiratory failure]; blood clots in the lung or a large vein after surgery [postoperative pulmonary embolism or deep venous thrombosis]; wounds that split open after surgery [postoperative wound dehiscence]; accidental cuts and tears [accidental puncture or laceration]; death after surgery to repair a weakness in the abdominal aorta [abdominal aortic aneurysm mortality rate]; and death among patients with hip fractures [hip fracture mortality rate])		✓
Bae et al [32]	*BioMed Central Health Services Research*	National Ambulatory Medical Care Survey (37,962 patient visits to 1470 primary care physicians from 2006 to 2009)	To analyze the impact of EHRs on primary care physicians' workloads	Duration measured in minutes of the face-to-face encounter between physicians and patients (patient face time) for direct patient care during the office visit and number of total patient office visits per physician per week (patient volume)		✓
Behkami et al [33]	*Studies in Health Technology and Informatics*	Simulation of clinic-type scenarios to capture the dynamic nature of policy interventions that affect the adoption of EHR	To describe a framework that allows decision-makers to efficiently evaluate factors that affect EHR adoption and test financial incentives	Revenue	✓	

Study	Journal or conference	Study period or data set	Objective	Outcome measures	Financial (n=21)	Clinical (n=54)
Bishop et al [34]	*Health Affairs*	Interviews of medical group leaders (n=21) who use electronic communication with patients extensively and staff from 6 of the groups	To understand how primary care practices can use electronic communication to manage clinical issues that are usually managed during clinic visits; determine perceived advantages and disadvantages of the electronic communication programs for patients, physicians, and practices; and determine the barriers to and facilitators of the implementation of the electronic communication programs	The convenience of access, patient satisfaction, efficiency, safety and quality of care, and workload		✓
Brown Jr et al [29]	*Journal of Addiction Medicine*	Data collected from paper patient charts (for preimplementation data) and electronic patient charts (for postimplementation data); patients, clinicians, and management stakeholders participated in surveys	To evaluate the impact of an EMR[f] system on the Opioid Agonist Treatment Program	Financial performance (revenue), quality (timeliness of medical assessments), productivity (clinic visits), patient satisfaction, and risk management (incident reports)	✓	✓
Bucher et al [35]	*Journal of the American College of Surgeons*	CMS SCIP[g] measuring compliance rates; HIMSS[h] hospital EHR adoption survey from 2006 to 2012	To analyze the impact of EHR adoption on hospital compliance with quality and process measures	Hospital compliance with SCIP core measures		✓
Burke et al [36]	*Journal of Innovation in Health Informatics*	Notes of outpatients with type 2 diabetes analyzed (n=537) for 5.5 years	To analyze the impact of EHR use on clinical quality measures and HbA_{1c}[i]	HbA_{1c} values		✓
Cheriff et al [37]	*International Journal of Medical Informatics*	The practice management system used to extract physician productivity data (n=203)	To describe the changes in physician productivity in an academic multispecialty group because of ambulatory EHR adoption	Average monthly charge, visit volume, and work-relative value units	✓	✓
Chiang et al [38]	*Journal of American Association for Pediatric Ophthalmology and Strabismus*	Academic pediatric ophthalmology practice data for the year 2006 (n=4 faculty providers)	To analyze the impact of EHR implementation on the volume and time for pediatric ophthalmology	Clinical volume		✓
Chiang et al [39]	*Transactions of the American Ophthalmological Society*	Outpatient clinical examinations (n=120,490) from faculty providers (n=23) at an academic ophthalmology department analyzed for 3 years	To evaluate clinical volume, time requirements, and nature of clinical documentation related to EHR implementation	Clinical volume, time requirements, and nature of clinical documentation		✓
Choi et al [40]	*Journal of Medical Systems*	Retrospective chart review study—a convenience sample of 60 to 80 charts reviewed every month from (January 1, 2006, to October 4, 2009, n=3997; October 5, 2009, to December 31, 2010, n=984)	To analyze the organizational performance and regulatory compliance before and after implementation of the Anesthesia Information Management System	Documentation of medication and patient status		✓

Study	Journal or conference	Study period or data set	Objective	Outcome measures	Financial (n=21)	Clinical (n=54)
Collum et al [8]	*Healthcare Management Review*	AHA[j] Annual survey (2007-2010), AHA IT Supplement (2007-2010), and CMS Medicare Cost Reports (2007-2011)	To examine how EHR adoption affects hospital financial performance	Profit margins and return on assets	✓	
Dandu et al [41]	*Clinical Orthopedics and Related Research*	Data were collected from a combination of the Physician Compare data set (2016), Meaningful Use Eligible Professional public use files (2011-2016), and Medicare Utilization and Payment data sets (2012-2016)	To evaluate the impact of EHRs on provider productivity, billing, and orthopedic surgery	Billing, outpatient volume, and surgical volume	✓	✓
Daniel et al [42]	*Academic Emergency Medicine*	Health plan and electronic hospital data from a large urban ED[k] (November 1, 2004, to March 31, 2005, n=1509 ED encounters compared with September 1, 2005, to February 17, 2006, n=779 ED encounters)	To evaluate the use of paper-based EHR in an ED on LOS[l] and plan payments	Plan payment for ED encounters and ED LOS	✓	✓
Deily et al [43]	*Health Research and Educational Trust*	Administrative claims data in Pennsylvania from 1998 to 2004 (n=491,832)	To examine whether HIT at nonhospital facilities improves health outcomes and decreases resource use at hospitals within the same network and whether the effect of HIT differs as providers obtain more experience with it	Incidence of obstetric trauma and preventable complications; LOS		✓
Edwardson et al [44]	*Medical Care Research and Review*	Financial panel data from the pediatric primary care network comprising 260 providers across 42 practices (2008-2013)	To examine the effect of EHR adoption on charge capture	Average per-patient charge, average per-patient collections, and charge-to-collection ratios	✓	
Ehrlich et al [45]	*Applied Clinical Informatics*	Survey responses from 32 ophthalmologists after implementation, 28 at 3 months, 35 at 7 months, 40 at 13 months, and 39 at 24 months after implementation (implementation in 2012)	To comprehend and describe the perceptions of ophthalmologists during EHR implementation in an academic department of ophthalmology	Documentation quality, workflow, and efficiency		✓
Flatow et al [46]	*Applied Clinical Informatics*	Retrospective chart review for all patients admitted to the surgical intensive care unit (n=3742; January 1, 2009, to December 31, 2010)	To evaluate key quality measures of a surgical intensive care unit following EHR implementation in a tertiary hospital	LOS, mortality, central line–associated bloodstream infection rates, clostridium difficile colitis rates, readmission rates, and number of coded diagnoses		✓
Furukawa et al [47]	*Journal of the American Medical Informatics Association*	Data collected from Medicare Patient Safety Monitoring System (2010-2013) and HIMSS Analytics database (2008-2013)	To evaluate the impact of meaningful use capabilities on in-hospital adverse drug events	Rate of adverse drug events		✓

Study	Journal or conference	Study period or data set	Objective	Outcome measures	Financial (n=21)	Clinical (n=54)
Han et al [48]	*American Journal of the Medical Sciences*	A prospective observational study (n=797 patients) at an urban teaching hospital from July 2010 to June 2011 in the MICU[m]	To determine the effect of EHR on MICU mortality, hospital LOS, and medication errors	MICU mortality, hospital LOS, and medication errors		✓
Hepp et al [17]	*Value in Health*	The decision-analytic model was used to estimate the cost-effectiveness of CPOE[n] in a multidisciplinary medical group for the years 2010 to 2014 (n=400 providers)	To assess the cost-effectiveness of CPOE in the reduction of medication errors and adverse drug events in an ambulatory setting	Costs (CPOE system costs, personnel costs, administrative costs, and prescribing costs), financial incentives (Health Information Technology for Economic and Clinical Health meaningful use incentives and pay-for-performance incentives), medication error probability, and adverse drug event probability	✓	✓
Herasevich et al [49]	*Critical Care Medicine*	A prospective study at Mayo Clinic, Rochester, Minnesota (n=1159 patients) from February 16, 2008, to February 16, 2009	To design and test an electronic algorithm that includes patient characteristics and ventilator settings, allowing notification to bedside providers about potentially injurious ventilator settings to improve the safety of ventilator care and decrease the risk of ventilator-related lung injury	Prevalence of acute lung injury		✓
Hessels et al [50]	*Online Journal of Nursing Informatics*	Data on 854,258 adult patients discharged from 70 New Jersey hospitals and 7679 nurses working in those same hospitals for the year 2006	To examine the relationship between the EHR adoption stage, missed nursing care, nursing practice environment, and adverse outcomes and satisfaction of patients who are hospitalized	Prolonged LOS and patient satisfaction		✓
Howley et al [51]	*Journal of the American Medical Informatics Association*	Compared practice productivity and reimbursement of ambulatory practices (n=30) for 2 years after EHR implementation to their per-EHR implementation baseline	To evaluate how EHR implementation affects the financial performance of ambulatory practices	Reimbursement and practice productivity (number of patient visits)	✓	✓
Jones et al [52]	*American Journal of Managed Care*	Database with 2021 hospitals collected by linking the AHA Annual Survey database, Hospital Compare database, and HIMSS database for the years 2004 and 2007	To analyze longitudinal data on EHR adoption to evaluate the impact of new EHR adoption on quality improvement	Composite measures of hospital process quality for acute myocardial infarction, health failure, and pneumonia		✓
Katzer et al [53]	*Applied Clinical Informatics*	Prehospital patient care reports (n=154) at Georgetown University's student-run Emergency Medical Services organization	To describe whether implementing an electronic patient care report system influenced improvement in physical exam documentation	Mean physical exam documentation		✓

Study	Journal or conference	Study period or data set	Objective	Outcome measures	Financial (n=21)	Clinical (n=54)
Kritz et al [54]	*Journal of Evaluation in Clinical Practice*	Opioid treatment program clinics (7 clinics) in New York State—paper patient charts and electronic patient charts (to analyze pre- and postimplementation data), assessment meetings and surveys with patients, direct care providers, and supervisors or managers	Prospective, comparative study using a pre- and postimplementation design to establish whether EHR implementation yielded any improvements	Revenue, quality, productivity, risk management, and satisfaction	✓	✓
Lam et al [55]	*BioMed Central Health Services Research*	Data from physicians with practices at the University of Washington Department of Ophthalmology for the years 2008 to 2012 (n=8 physicians)	To analyze the impact of EHR adoption on patient visit volume at an academic ophthalmology department	Patient volume per provider		✓
Lim et al [28]	*Journal of American Medical Association Ophthalmology*	Population-based, cross-sectional study (n=348)	To evaluate the adoption rate and perceptions of financial and clinical outcomes of EHRs among ophthalmologists in the United States	Net revenues and productivity	✓	✓
Love et al [3]	*Journal of American Medical Informatics Association*	2007 state-wide survey of Massachusetts physicians (n=541)	To characterize and describe physicians' attitudes toward EHR's potential to cause new errors, improve health care quality, and change physician satisfaction	Medical errors, quality of care, and physician satisfaction		✓
Lowe et al [56]	*Journal of Wound Ostomy Continence Nurses Society*	Data were collected from a regional Veterans Affairs database and computerized patient medical records for a year after implementation of the EMR wound care template (October 1, 2006, to September 30, 2007) and 2 years before the intervention	To evaluate the impact of a 1-year intervention of an EMR wound care template on the completeness of wound care documentation and medical coding and compare results with the preintervention period	Documentation of wound care and documentation of coding for diagnoses and procedures		✓

Study	Journal or conference	Study period or data set	Objective	Outcome measures	Financial (n=21)	Clinical (n=54)
McCullough et al [57]	*Health Affairs*	AHA Annual Survey, HIMSS Analytics, and CMS Hospital Compare database for the years 2004 to 2007 (n=3401 nonfederal acute care US hospitals)	To analyze the impact of HIT on the quality of care in US hospitals	Quality indicators: percentage of patients with heart failure given angiotensin-converting enzyme inhibitor or angiotensin II receptor blocker for left ventricular systolic dysfunction; the percentage of smokers with heart failure and pneumonia who were given smoking cessation advice; the percentage of patients with pneumonia assessed and given pneumococcal vaccination if indicated; the percentage of patients with pneumonia whose initial blood culture in the ED preceded their first dose of the hospital-administered antibiotics; and the percentage of patients with pneumonia given the most appropriate initial antibiotic		✓
McCullough et al [58]	*Generating Evidence and Methods to improve patient outcomes)*	Manual review of the paper and electronic charts for 6007 patients across 35 small primary care practices	To analyze the quality measure performance in small practices before and after EHR adoption	Clinical quality measures: antithrombotic therapy, BMI recorded, smoking status recorded, smoking cessation intervention offered, HbA_{1c} testing and control, cholesterol testing and control, and BP^{o} control		✓
Mirani and Harpalani [27]	*ACM Transactions on Management Information Systems*	“Data and Reports” and “Hospital Cost Report” from the CMS website for 2008 to 2010	To analyze the impact of the Medicare EHR incentive program on acute care hospitals	The average cost of ancillary services per patient, profit margins, inpatient bed debts, outpatient bed debts, and patient stay durations	✓	✓
Mitchell et al [59]	*The Journal of Rural Health*	AHA EHR adoption survey and CMS Hospital Compare data set for the year 2009	To investigate whether there is an association between clinical decision support system use and quality disparities in pneumonia process indicators between rural and urban hospitals	Percentage of hospitals meeting quality requirements and pneumonia process composite scores		✓
Patterson et al [60]	*Applied Clinical Informatics*	Data used from the AHA Health IT survey and Medicare Part A claims (n=52,048 Medicare beneficiaries discharged for heart failure anytime during the calendar year 2008)	To compare 30 days all-cause readmission rates for Medicare patients with health failure discharged from hospitals with fully implemented comprehensive EHR vs without it	30-day all-cause readmission rates		✓

Study	Journal or conference	Study period or data set	Objective	Outcome measures	Financial (n=21)	Clinical (n=54)
Persell et al [61]	*Medical Care*	Time series analysis at a large internal medicine practice from February 1, 2007, to February 1, 2009 (n=12,299 patients eligible at the beginning of the intervention)	To implement and analyze a multifaceted quality improvement intervention using EHRs as tools for improving performance	Quality measures pertaining to coronary heart disease, health failure, diabetes mellitus, and prevention		✓
Radley et al [62]	*Journal of American Medical Informatics Association*	Systematic literature review and random-effects meta-analytic techniques, American Society of Health System Pharmacists Annual survey (2007), AHA Annual Survey (2007), and AHA EHR Adoption Database supplement (2008)	To analyze the adoption of CPOE systems on the reduction in medication errors in hospitals	Likelihood of medication errors		✓
Rao et al [63]	*Journal of American Medical Informatics Association*	Mailed surveys to a nationally representative random sample of practicing physicians from the Physician Masterfile of the American Medical Association (n=2769)	To analyze variation in the adoption of EHR functionalities and their use patterns, barriers to adoption, and perceived benefits by physician practice size	Physician perceptions of quality of clinical decision, quality of communication with patients and other providers, delivery of preventive or chronic care that met the guidelines, avoiding medication errors and prescription refills		✓
Risko et al [64]	*Healthcare*	Patient processing metrics (n=374 observations) were collected for ED physicians (34 physicians) at 2 hospitals for 7 months before and 10 months after EHR implementation	To analyze the impact of EHR implementation on ED physician efficiency and patient throughput	Patient workup times and LOS		✓
Ryan et al [65]	*Medical Care*	Data collected from 143 practices with EHR implementation (2009-2011)	To analyze whether EHR implementation and complementary interventions, such as clinical decision support, technical assistance, and financial incentives improved, the quality of care provided	Quality of care was analyzed from 8 separate indicators; 4 cardiovascular measures (smoking cessation intervention, BP control, cholesterol control, and aspirin or antithrombotic treatment) and 4 additional clinically important measures (BMI measurement, HbA$_{1c}$ control, pneumococcal vaccine, and asthma control)		✓
Schreiber and Shaha [66]	*Journal of Innovation in Health Informatics*	Data collected from a community hospital for 5 years after CPOE adoption	To evaluate whether an increase in adoption of CPOE leads to a decrease in LOS	LOS and cost measured by LOS	✓	✓
Scott et al [67]	*The Journal of Bone and Joint Surgery*	Data collected from an outpatient adult reconstruction clinic (n=143 patients) before implementing the hospital system–wide EMR system and 2 months, 6 months, and 2 years after implementation	To evaluate the impact of EMR implementation using advanced cost-accounting methods on orthopedic surgeons in an outpatient setting	Labor cost, documentation time for providers, and time spent interacting with patients	✓	✓

Study	Journal or conference	Study period or data set	Objective	Outcome measures	Financial (n=21)	Clinical (n=54)
Shen et al [68]	*International Journal of Healthcare Technology and Management*	National Inpatient Sample and AHA EHR implementation survey for the year 2009	To examine how EHR adoption affected the cost of care and quality outcomes in an acute care hospital setting	Cost of care for the 8 quality indicators (cardiovascular and cerebrovascular) and quality indicators for 5 cardiovascular and 3 cerebrovascular conditions and procedures	✓	✓
Silow-Carroll et al [69]	*Issue Brief (Commonwealth Fund)*	Interviews with individuals in the 9 hospitals that implemented a comprehensive EHR system	To analyze the experience of 9 hospitals in using EHR to improve quality and efficiency	Communication among providers, care coordination, patient engagement, and medical errors		✓
Singh et al [70]	*Journal of American Medical Association Ophthalmology*	Retrospective case-control study comparing the pre- (n=13,969 patient encounters) and post-EHR (n=14,191 patient encounters) implementation periods at an eye institute	To evaluate the impact of EHR system implementation from clinical and economic perspectives at a large multidisciplinary ophthalmic practice	Net revenue, revenue to volume ratio, capital and implementation costs, EHR incentive payments received, patient volume, diagnostic and procedure volume, and coding volumes	✓	✓
Sockolow et al [30]	*Applied Clinical Informatics*	Pre- and postobservational mixed methods study, Philadelphia-based homecare agency with 137 clinicians—data included clinician EHR documentation completion, EHR use data, Medicare billing data, an EHR Nurse Satisfaction survey, clinician observations, clinician interviews, and patient outcomes	To compare workflows, financial billing, and patient outcomes before and after implementation to analyze the effect of a homecare point of care EHR	Number of days required to create a financial reimbursement bill, productivity, behavioral outcomes, and clinicians' perceptions of patient safety	✓	✓
Thirukumaran et al [71]	*Health Services Research*	Data collected from the SCIP Core Measure data set from the CMS Hospital Inpatient Quality Reporting (n=1816) program (March 2010 to February 2012)	To evaluate the effect of EHR placement on SCIP measures in a tertiary care teaching hospital	SCIP scores		✓
Tidwell et al [72]	*Obstetrics and Gynecology*	Data collected from an obstetrics and gynecology practice comprising 6 physicians and 6 midwives with 150 daily visits	To evaluate whether a low-cost electronic practice management system (EHR) can improve care coordination and financial measures	Net profit, days in accounts receivables, patient visits, no-show rate, and quality data gathering	✓	✓
Varpio et al [73]	*Medical Education*	A 2-phase longitudinal study; data collected through field observations (146 hours with 300 providers, 22 patients, and 32 patient family members), think-aloud (n=13) and think-after (n=11) sessions, interviews (n=39) and document retrieval (n=392)	To evaluate the impact of adopting EHR on clinician experience	Clinician experience was measured in terms of cognitive workload, clinical reasoning support mechanisms, and knowledge about the patient		✓

Study	Journal or conference	Study period or data set	Objective	Outcome measures	Financial (n=21)	Clinical (n=54)
Walker-Czyz et al [74]	*Journal of Nursing Administration*	Data for a quantitative, retrospective analysis collected from urban hospitals (431 beds) with 10 medical-surgical units and 2 critical care units	To evaluate how an integrated EHR innovation adoption affects cost, nurse satisfaction, and nursing care delivered in terms of quality	Cost (nurse hours per patient day, nurse turnover, and nurse overtime), quality nursing care outcomes (hospital-acquired falls and pressure ulcers, ventilator-associated pneumonia, central line–associated bloodstream infections, and catheter-associated urinary tract infections)	✓	✓
Wang et al [75]	*Preventing Chronic Disease*	Clinical quality measure performance data collected from 151 primary care practices that implemented EHR (October 2009 to October 2011)	To analyze how clinical quality measures for independent primary care practices improve as a result of EHR use and technical support from a local public health agency	4 key quality measures: antithrombotic therapy, BP control, HbA_{1c} testing, and smoking cessation intervention		✓
Wang et al [26]	*International Journal of Accounting Information Systems*	Definitive health care data set for hospital-level data for the years 2011 to 2016 (n=3266 observations)	To evaluate how HIT expenses and intermediate business processes affect hospital financial performance and productivity	Return on assets, productivity ([net revenue, 1 million]), and number of staff beds)	✓	✓
Xiao et al [76]	*Perspectives in Health Information Management*	Charts were reviewed to collect data from a large tertiary public medical center (3 years before and 3 years after EHR implementation in July 2009)	To describe how electronic charting implementation in a large public outpatient clinic improves clinical documentation	Note completion and documentation of medication		✓
Yeung [16]	*International Journal of Medical Informatics*	433 local health departments’ population-based data for 433 counties	To determine the impact of the adoption of EHR and health information exchange changes by local health departments on population health	The health of a population at the county level, as measured by health outcomes such as premature death and health-related quality of life		✓
Wani and Malhotra [77]	*Journal of Operations Management*	Acute care hospitals in California	To analyze the impact of EHR adoption in terms of full adoption vs meaningful assimilation on clinical outcomes	LOS and readmission rates		✓

Study	Journal or conference	Study period or data set	Objective	Outcome measures	Financial (n=21)	Clinical (n=54)
Zhou et al [78]	*Journal of the American Medical Informatics Association*	To evaluate the extent of EHR use and how the quality of care delivered in ambulatory care practices varied according to the duration of EHR availability	Quality measures are aggregated into 6 clinical categories (asthma care, behavioral and mental health, cancer screening, diabetes care, well-child and adolescent visits, and women's health screenings)	Quality measures aggregated into 6 clinical categories (asthma care, behavioral and mental health, cancer screening, diabetes care, well child and adolescent visit, women's health screenings)	✓	

[a] EHR: electronic health record.

[b] CMS: Centers for Medicare and Medicaid Services.

[c] ✓: indicates that the outcome was discussed in the study.

[d] HIT: health IT.

[e] AHRQ: Agency for Healthcare Research and Quality.

[f] EMR: electronic medical record.

[g] SCIP: Surgical Care Improvement Project.

[h] HIMSS: Healthcare Information and Management Systems Society.

[i] HbA_{1c}: hemoglobin A_{1c}.

[j] AHA: American Hospital Association.

[k] ED: emergency department.

[l] LOS: length of stay.

[m] MICU: medical intensive care unit.

[n] CPOE: certified provider order entry.

[o] BP: blood pressure.

Most of the studies included in this review of the literature had financial outcome measures that demonstrated some form of improvement. One of the studies reported that costs that increased during the implementation period were equivalent to the preimplementation level after 6 months [67]. Hepp et al [17] found that the certified physician order entry (CPOE) system (part of the EHR system) generated lower costs in addition to improving medication safety. A few other studies also confirmed that patients in facilities with EHR systems incurred lower costs than those in facilities without an EHR system [54,68,69].

In terms of mixed financial outcomes, the analysis of Adler-Milstein et al [18] exhibited that greater EHR adoption did not improve financial efficiency (measured by the ratio of a hospital's total expenditures to adjusted patient days) for nonfederal acute care hospitals immediately after the adoption of EHR; however, the results from this study reported improvements in financial efficiency for the years 2010 and 2011 compared with the years 2008 and 2009 [18].

Regarding the reimbursement measure, EHR systems were thought to be responsible for significant improvements in the timeliness of clinical documentation and billing for reimbursement [30,41,76]. The analysis of Cheriff et al [37] documented that physicians who adopted EHRs in a large academic multispecialty physician group captured higher average monthly charges than before the use of EHRs. Similarly, another study reported that the introduction of EHRs was associated with an increase in average per-patient charge and an increase in average per-patient collection [44].

In terms of revenues, profit margins, and return on assets, revenues were reported to have increased in conjunction with EHR adoption [29,51]. A few studies reported improved financial performance concerning savings [42], net profit, and days in account receivables [72] as a result of EHR adoption. One of the studies examined the association among HIT expenses, hospital financial performance, and productivity, with EHR adoption being an intermediate variable. This study indicated a direct and positive association between HIT investment and positive financial performance regarding return on assets [26].

By contrast, a set of results from a survey of ophthalmologists indicated increasing costs and decreasing revenue and productivity with the adoption of EHRs [28]. Other studies have similarly reported findings in terms of a decrease in revenue [54,70] and an increase in cost [29] as a result of EHR adoption. Dandu et al [41] did not provide any statistically significant evidence to report a direct association between EHR adoption and higher-level billing [41]. Similarly, Mirani and Harpalani [27] did not provide any statistically significant evidence to report a direct association between EHR adoption and revenue. Findings from Collum et al [8] suggested that alterations in the level of EHR adoption were not related to increases in revenue and the reduction of operating margins.

Clinical Outcomes

Of the 58 reviewed studies, 55 (95%) contained segments that were coded under the category of "Value-Clinical Outcomes." The differing measures for clinical outcomes in these studies were productivity [26,28,30], workflow inefficiency, medical errors, patient safety [3], patient satisfaction, clinical volume, readmission rates, patient LOS [27], and quality indicators at the individual patient level. The different measures of clinical outcomes are listed and described in depth in Table 2. The studies detailed both positive (33/58, 57%), negative (16/58, 28%), and no (7/58, 12%) association relationships between EHR adoption and clinical outcomes. Similar to financial outcomes, an overlap of both positive and negative impacts pertaining to EHR adoption on clinical outcomes was observed in some of the studies.

Most of the clinical outcome measures involved in this review exhibited some form of improvement. The Hessels et al [50] study reported a statistically significant association between EHR adoption and LOS. A significant reduction of LOS in emergency departments [42] and medical errors in emergency and critical care departments [48,49], as well as inpatient acute care settings [62], were indicated as a result of EHR adoption. The rising and falling CPOE rates were also determined to be in correlation with the increase and decrease in LOS [66].

In connection with workflow efficiency and productivity, EHR use was reportedly helpful in improving the promptness of clinical documentation [30], enhancing productivity and efficiency in the workloads of primary care physicians [32], and increasing productivity [37]. Furthermore, EHR was found to be responsible for an increase in patient visits (which results in increased revenue), a decrease in no-show rates (also increasing revenue), and improved care coordination [72]. There was statistically significant progress in terms of completion rates of assessments [29,54], better documentation of medication, patients' vital signs and pain scores [40], and improved clinical documentation [53,56,76] as a result of EHR adoption.

For the category of patient satisfaction, physicians recognized electronic communication permitted through EHR as a secure and efficient way of communicating with patients, resulting in improvements in patient satisfaction [34]. A study discovered evidence that higher levels of EHR adoption were positively associated with performance and patient satisfaction. This study detected improvements in performance and patient satisfaction for the years 2010 and 2011 compared with the years 2008 and 2009 [18].

With regard to patient safety and medical errors, surgical IT systems (as a subset of EHR systems) positively affected levels of patient safety, compliance, and quality and process measures for patients undergoing surgical procedures in hospitals [31,35]. Outside of surgical IT systems, clinical decision support has also been shown to address other areas of patient safety [59]. For example, adverse drug events decreased by 20% [47], and CPOE was reported to provide exceptional value by improving medication safety in a cost-effective manner [17].

Indicators of quality at the individual patient level, such as rates of antithrombotic therapy and nicotine use documentation, increased immediately following EHR adoption [58]. Similarly, another study reported improvements in antibiotic therapy, blood pressure control, hemoglobin A_{1c} testing, and smoking cessation interventions because of EHR systems [75].

In contrast, for productivity and workload efficiency, the results of a survey indicated that physicians perceived that EHR adoption harmed productivity and increased their workload [28,34,45]. EHR implementation was reportedly associated with increased documentation effort and time, with little to no increase in clinical volume and little to no or perhaps a negative impact on clinical and surgical volume [38,39,41]. Increased documentation time because of EHR adoption resulted in a decrease in the time spent reviewing patient records and performing physical examinations [67]. The results from one of the studies did not identify any differences in productivity (total visit volume) resulting from EHR adoption [70]; however, 3% (2/58) of other studies detailed a decrease in productivity immediately following the adoption of EHR [51,64]. Another example includes significant and consistent decreases in patient volume spanning 4 years after EHR adoption in an academic outpatient ophthalmology practice [55]. EHR systems were said to increase the number of missed assessments, decrease the timely completion rate of assessments, and negatively affect the productivity of clinicians [54]. A study reported that physicians were mostly checking boxes to complete the EHR data process instead of developing or using investigative strategies, which are common among diagnosticians [73].

Considering the patient satisfaction, quality, safety, LOS, and readmission rate perspectives, EHR use resulted in lower patient satisfaction [79] and quality of care [71] for a few years following the adoption of EHRs. In addition, EHR use was associated with an increase in hospital-acquired conditions during EHR implementation [74]. No relationship was found to exist between practice size and the impact of EHR on the quality of patient care from the perspectives of physicians [63]. Some studies reported no association between EHR adoption and improvement in the quality of care provided [36,52,68,78], readmission rates [60], and LOS [48]. Findings from another study that examined physician perceptions of EHRs indicated that physicians believed that EHRs could create new opportunities for error [3].

The Intersection of Financial and Clinical Outcomes

Having reported on studies that examined financial and clinical outcomes as individual factors, we now report on studies that examined *both* financial and clinical outcomes.

Overall, 9% (5/58) of studies surveyed for this review of the literature reported on the intersection of financial and clinical outcomes. To further investigate this intersection, the category "Value–Intersection of Financial and Clinical Outcomes" was generated. Furthermore, 80% (4/5) of these studies specified a positive association between EHR adoption and financial and clinical outcomes.

In terms of the financial outcomes, hospitals that had adopted EHR selectively increased the efficiency of their turnover rate of Medicare patients to receive higher MU incentives [27]. These findings point toward the impact of EHR adoption on a patient's stay duration on average (clinical outcome), which, in turn, affects their compensation because of the loss of patient days (financial outcome) from CMS. EHR adoption was associated with enabling the prioritization of improvements in clinical documentation time to improve agency cash flow [30]. EHR use was thought to contribute to shortened emergency department LOS, which led to a positive impact in terms of CMS compensation [42]. Similarly, CPOE, a subset of EHR, was said to be an independent factor in the impact of LOS; therefore, it indirectly contributed to lower costs [66]. By contrast, 20% (1/5) of the studies reported that EHR adoption required a learning period, where increased medical assistant time, patient time, and physician documentation time incurred additional costs [67].

Discussion

Principal Findings

The primary goal of this literature review was to substantiate how EHR value is described concerning 2 different outcome categories, financial and clinical outcomes, and to further the exploration of the impact of EHR adoption on these 2 outcome categories. Subsequently, this review incorporated studies that described relationships between EHR adoption along with financial and clinical outcomes with a priori categories (financial outcomes and clinical outcomes) and with an additional category that included the intersection of financial and clinical outcomes. This review of the literature included a total of 58 studies.

Overall, 76% (16/21) of the studies that discussed the financial outcomes of EHR adoption presented a positive relationship between EHR adoption and financial outcomes. These studies observed changes in financial outcomes in terms of profit ratios, costs, revenues, reimbursements, and return on assets. Consistent with the literature, value realization, especially in terms of financial outcomes, is lagging as it involves a large upfront cost [18].

Regarding clinical outcomes, 76% (35/58) of the studies that examined the clinical outcomes of EHR adoption indicated a positive relationship between EHR adoption and clinical outcomes in terms of LOS, readmission rates, patient satisfaction, medical errors, patient safety, user productivity, and quality indicators at individual patient levels. Similar to financial outcomes, value realization regarding clinical outcomes also improved over time. For instance, clinical outcome measures such as rates of hemoglobin A_{1c} testing, recorded BMI, and cholesterol testing decreased before rebounding, following the adoption of EHR [57].

Of the 58 studies in this review of the literature, 5 (9%) studies highlighted the intersection of financial and clinical outcomes. EHR adoption allowed for improvements in clinical documentation time and LOS and sequentially reduced overall costs and improved reimbursement [27,30,42,66]. EHR adoption was also responsible for an increase in personnel costs in association with the new technology’s initial steep learning curve [67]. Overall, these studies indicated interdependence between financial and clinical outcomes, in essence, how one was associated with the other in some form.

This review of the literature discovered some studies with contradictory findings. For example, financial outcomes such as profit margins, return on assets, and costs were some of the measures that reported contradictory findings. A potential reason could be that the studies that reported an inverse relationship reviewed these measures right after the adoption, as opposed to studies that reported it after a longer period. Organizational performance measures such as return on assets, ROI, and return on equity could be examined to explore the cyclical relationship between IT inputs and productivity [80]. Future research may be required to investigate the trajectory and extent of the relationship between IT investments and reinvestments, such as EHR adoption or readoption, and clinical outcomes to further expand upon this question.

Limitations

The comprehensive findings of this literature review should be considered along with the limitations. Concerning the searched databases, PubMed, Scopus, and Embase—the primary health services and HIT databases—were used. It is possible that studies on the value of EHRs were published outside of health-focused journals and if so, may not have been included in this literature review. Another limitation of this review involves the keywords used in the selection criteria of the article search process. It is possible that the used keywords were not exhaustive, and studies could have been overlooked. Finally, this review included English-only studies that were conducted in the United States. It is possible that other countries with EHRs may have had an experiential understanding that could have contributed to this review. To mitigate bias, manual screening of all the references of included studies was conducted.

Conclusions

This review of the literature reports on the individual and collective value of EHRs from a financial and clinical outcomes perspective. The collective perspective examined the intersection of financial and clinical outcomes, suggesting a reversal of the current understanding of how IT investments could generate productivity improvements, and prompted a new question to be asked about whether an increase in productivity could potentially lead to more IT investments.

Conflicts of Interest

None declared.

References

1. Electronic Health Records. Healthcare Information and Management Systems Society. 2011. URL: https://www.himss.org/library/ehr [accessed 2022-09-12]
2. Garrett P, Seidman J. EMR vs EHR – What is the Difference? Health IT Buzz. 2011 Jan 4. URL: https://www.healthit.gov/buzz-blog/electronic-health-and-medical-records/emr-vs-ehr-difference [accessed 2022-09-12]

3. Love JS, Wright A, Simon SR, Jenter CA, Soran CS, Volk LA, et al. Are physicians' perceptions of healthcare quality and practice satisfaction affected by errors associated with electronic health record use? J Am Med Inform Assoc 2012;19(4):610-614 [FREE Full text] [doi: 10.1136/amiajnl-2011-000544] [Medline: 22199017]
4. Redd TK, Read-Brown S, Choi D, Yackel TR, Tu DC, Chiang MF. Electronic health record impact on productivity and efficiency in an academic pediatric ophthalmology practice. J AAPOS 2014 Dec;18(6):584-589 [FREE Full text] [doi: 10.1016/j.jaapos.2014.08.002] [Medline: 25456030]
5. McAlearney AS, Sieck C, Hefner J, Robbins J, Huerta TR. Facilitating ambulatory electronic health record system implementation: evidence from a qualitative study. Biomed Res Int 2013;2013:629574 [FREE Full text] [doi: 10.1155/2013/629574] [Medline: 24228257]
6. Adler-Milstein J, Green CE, Bates DW. A survey analysis suggests that electronic health records will yield revenue gains for some practices and losses for many. Health Aff (Millwood) 2013 Mar;32(3):562-570. [doi: 10.1377/hlthaff.2012.0306] [Medline: 23459736]
7. Medicare and Medicaid Promoting Interoperability Program Basics. Centers for Medicare & Medicaid Services. 2018. URL: https://www.cms.gov [accessed 2019-01-16]
8. Collum TH, Menachemi N, Sen B. Does electronic health record use improve hospital financial performance? Evidence from panel data. Health Care Manage Rev 2016;41(3):267-274. [doi: 10.1097/HMR.0000000000000068] [Medline: 26052785]
9. Jang Y, Lortie MA, Sanche S. Return on investment in electronic health records in primary care practices: a mixed-methods study. JMIR Med Inform 2014 Sep 29;2(2):e25 [FREE Full text] [doi: 10.2196/medinform.3631] [Medline: 25600508]
10. Pine R, Tart K. Return on investment: benefits and challenges of baccalaureate nurse residency program. Nurs Econ 2007;25(1):13-19. [Medline: 17402673]
11. Peterson LT, Ford EW, Eberhardt J, Huerta TR, Menachemi N. Assessing differences between physicians' realized and anticipated gains from electronic health record adoption. J Med Syst 2011 Apr;35(2):151-161. [doi: 10.1007/s10916-009-9352-z] [Medline: 20703574]
12. value definition. Oxford Dictionary. 2019. URL: https://en.oxforddictionaries.com/definition/value [accessed 2019-01-16]
13. Payne TH, Bates DW, Berner ES, Bernstam EV, Covvey HD, Frisse ME, et al. Healthcare information technology and economics. J Am Med Inform Assoc 2013;20(2):212-217 [FREE Full text] [doi: 10.1136/amiajnl-2012-000821] [Medline: 22781191]
14. Feldman SS, Horan TA. Collaboration in electronic medical evidence development: a case study of the Social Security Administration's MEGAHIT System. Int J Med Inform 2011 Aug;80(8):e127-e140. [doi: 10.1016/j.ijmedinf.2011.01.012] [Medline: 21333588]
15. Riskin L, Koppel R, Riskin D. Re-examining health IT policy: what will it take to derive value from our investment? J Am Med Inform Assoc 2015 Mar;22(2):459-464. [doi: 10.1136/amiajnl-2014-003065] [Medline: 25326600]
16. Yeung T. Local health department adoption of electronic health records and health information exchanges and its impact on population health. Int J Med Inform 2019 Aug;128:1-6. [doi: 10.1016/j.ijmedinf.2019.04.011] [Medline: 31160006]
17. Hepp Z, Forrester SH, Roth J, Wirtz HS, Devine EB. Cost-effectiveness of a computerized provider order entry system in improving medication safety: a case study in ambulatory care. Value Health 2013 May 1;16(3):A205-A206. [doi: 10.1016/j.jval.2013.03.1038]
18. Adler-Milstein J, Everson J, Lee SD. EHR adoption and hospital performance: time-related effects. Health Serv Res 2015 Dec;50(6):1751-1771 [FREE Full text] [doi: 10.1111/1475-6773.12406] [Medline: 26473506]
19. Shah GH, Leider JP, Castrucci BC, Williams KS, Luo H. Characteristics of local health departments associated with implementation of electronic health records and other informatics systems. Public Health Rep 2016;131(2):272-282 [FREE Full text] [doi: 10.1177/003335491613100211] [Medline: 26957662]
20. World Health Organization. 2019. URL: https://www.who.int/en/ [accessed 2022-09-12]
21. Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM. Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. Rockville, MD, USA: Agency for Healthcare Research and Quality; 2013.
22. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 2009 Jul 21;6(7):e1000100 [FREE Full text] [doi: 10.1371/journal.pmed.1000100] [Medline: 19621070]
23. Fond G, Hamdani N, Kapczinski F, Boukouaci W, Drancourt N, Dargel A, et al. Effectiveness and tolerance of anti-inflammatory drugs' add-on therapy in major mental disorders: a systematic qualitative review. Acta Psychiatr Scand 2014 Mar;129(3):163-179. [doi: 10.1111/acps.12211] [Medline: 24215721]
24. Pichler G, Cheung PY, Aziz K, Urlesberger B, Schmölzer GM. How to monitor the brain during immediate neonatal transition and resuscitation? A systematic qualitative review of the literature. Neonatology 2014;105(3):205-210 [FREE Full text] [doi: 10.1159/000357162] [Medline: 24481411]
25. Friese S. Qualitative Data Analysis with ATLAS.ti. 2nd edition. Thousand Oaks, CA, USA: Sage Publications; 2014.
26. Wang T, Wang Y, McLeod A. Do health information technology investments impact hospital financial performance and productivity? Int J Account Inf Syst 2018 Mar;28:1-13. [doi: 10.1016/j.accinf.2017.12.002]

27. Mirani R, Harpalani A. Business benefits or incentive maximization? Impacts of the medicare EHR incentive program at acute care hospitals. ACM Trans Manage Inf Syst 2013 Dec;4(4):1-19. [doi: 10.1145/2543900]
28. Lim MC, Boland MV, McCannel CA, Saini A, Chiang MF, Epley KD, et al. Adoption of electronic health records and perceptions of financial and clinical outcomes among ophthalmologists in the United States. JAMA Ophthalmol 2018 Feb 01;136(2):164-170 [FREE Full text] [doi: 10.1001/jamaophthalmol.2017.5978] [Medline: 29285542]
29. Brown Jr LS, Kritz S, Lin M, Zavala R. Evaluation of an electronic medical record system at an opioid agonist treatment program. J Addict Med 2014;8(2):96-101 [FREE Full text] [doi: 10.1097/ADM.0000000000000018] [Medline: 24562402]
30. Sockolow PS, Bowles KH, Adelsberger MC, Chittams JL, Liao C. Impact of homecare electronic health record on timeliness of clinical documentation, reimbursement, and patient outcomes. Appl Clin Inform 2014 Apr 30;5(2):445-462 [FREE Full text] [doi: 10.4338/ACI-2013-12-RA-0106] [Medline: 25024760]
31. Appari A, Johnson EM, Anthony DL. Information technology and hospital patient safety: a cross-sectional study of US acute care hospitals. Am J Manag Care 2014 Nov;20(11 Spec No. 17):eSP39-eSP47 [FREE Full text] [Medline: 25811818]
32. Bae J, Encinosa WE. National estimates of the impact of electronic health records on the workload of primary care physicians. BMC Health Serv Res 2016 May 10;16:172 [FREE Full text] [doi: 10.1186/s12913-016-1422-6] [Medline: 27160147]
33. Behkami NA, Dorr DA, Morrice S. A business case for HIT adoption: effects of "meaningful use" EHR financial incentives on clinic revenue. Stud Health Technol Inform 2010;160(Pt 1):779-783 [FREE Full text] [Medline: 20841792]
34. Bishop TF, Press MJ, Mendelsohn JL, Casalino LP. Electronic communication improves access, but barriers to its widespread adoption remain. Health Aff (Millwood) 2013 Aug;32(8):1361-1367 [FREE Full text] [doi: 10.1377/hlthaff.2012.1151] [Medline: 23918479]
35. Bucher BT, Swords DS, Robinson J, Jackson GP, Finlayson SR. Advanced electronic health record adoption improves hospital compliance with surgical care improvement project core measures. J Am Coll Surgeons 2016 Oct;223(4):e33. [doi: 10.1016/j.jamcollsurg.2016.08.088]
36. Burke HB, Becher DA, Hoang A, Gimbel RW. The adoption of an electronic health record did not improve A1c values in type 2 diabetes. J Innov Health Inform 2016 Apr 15;23(1):144 [FREE Full text] [doi: 10.14236/jhi.v23i1.144] [Medline: 27348484]
37. Cheriff AD, Kapur AG, Qiu M, Cole CL. Physician productivity and the ambulatory EHR in a large academic multi-specialty physician group. Int J Med Inform 2010 Jul;79(7):492-500. [doi: 10.1016/j.ijmedinf.2010.04.006] [Medline: 20478738]
38. Chiang MF, Read-Brown S, Tu DC, Beaudet K, Yackel TR. Electronic health record implementation in pediatric ophthalmology: impact on volume and time. J AAPOS 2013 Feb 1;17(1):e3. [doi: 10.1016/j.jaapos.2012.12.011]
39. Chiang MF, Read-Brown S, Tu DC, Choi D, Sanders DS, Hwang TS, et al. Evaluation of electronic health record implementation in ophthalmology at an academic medical center (an American Ophthalmological Society thesis). Trans Am Ophthalmol Soc 2013 Sep;111:70-92 [FREE Full text] [Medline: 24167326]
40. Choi CK, Saberito D, Tyagaraj C, Tyagaraj K. Organizational performance and regulatory compliance as measured by clinical pertinence indicators before and after implementation of Anesthesia Information Management System (AIMS). J Med Syst 2014 Jan;38(1):5. [doi: 10.1007/s10916-013-0005-x] [Medline: 24424430]
41. Dandu N, Zmistowski B, Chen AF, Chapman T, Howley M. How are electronic health records associated with provider productivity and billing in orthopaedic surgery? Clin Orthop Relat Res 2019 Nov;477(11):2443-2451 [FREE Full text] [doi: 10.1097/CORR.0000000000000896] [Medline: 31389875]
42. Daniel GW, Ewen E, Willey VJ, Reese Iv CL, Shirazi F, Malone DC. Efficiency and economic benefits of a payer-based electronic health record in an emergency department. Acad Emerg Med 2010 Aug;17(8):824-833 [FREE Full text] [doi: 10.1111/j.1553-2712.2010.00816.x] [Medline: 20670319]
43. Deily ME, Hu T, Terrizzi S, Chou SY, Meyerhoefer CD. The impact of health information technology adoption by outpatient facilities on pregnancy outcomes. Health Serv Res 2013 Feb;48(1):70-94 [FREE Full text] [doi: 10.1111/j.1475-6773.2012.01441.x] [Medline: 22742682]
44. Edwardson N, Kash BA, Janakiraman R. Measuring the impact of electronic health record adoption on charge capture. Med Care Res Rev 2017 Oct;74(5):582-594. [doi: 10.1177/1077558716659408] [Medline: 27416948]
45. Ehrlich JR, Michelotti M, Blachley TS, Zheng K, Couper MP, Greenberg GM, et al. A two-year longitudinal assessment of ophthalmologists' perceptions after implementing an electronic health record system. Appl Clin Inform 2016 Oct 12;7(4):930-945 [FREE Full text] [doi: 10.4338/ACI-2016-05-RA-0075] [Medline: 27730248]
46. Flatow VH, Ibragimova N, Divino CM, Eshak DS, Twohig BC, Bassily-Marcus AM, et al. Quality outcomes in the surgical intensive care unit after electronic health record implementation. Appl Clin Inform 2015 Oct 7;6(4):611-618 [FREE Full text] [doi: 10.4338/ACI-2015-04-RA-0044] [Medline: 26767058]
47. Furukawa MF, Spector WD, Rhona Limcangco MR, Encinosa WE. Meaningful use of health information technology and declines in in-hospital adverse drug events. J Am Med Inform Assoc 2017 Jul 01;24(4):729-736 [FREE Full text] [doi: 10.1093/jamia/ocw183] [Medline: 28339642]
48. Han JE, Rabinovich M, Abraham P, Satyanarayana P, Liao TV, Udoji TN, et al. Effect of electronic health record implementation in critical care on survival and medication errors. Am J Med Sci 2016 Jun;351(6):576-581. [doi: 10.1016/j.amjms.2016.01.026] [Medline: 27238919]

49. Herasevich V, Tsapenko M, Kojicic M, Ahmed A, Kashyap R, Venkata C, et al. Limiting ventilator-induced lung injury through individual electronic medical record surveillance. Crit Care Med 2011 Jan;39(1):34-39. [doi: 10.1097/CCM.0b013e3181fa4184] [Medline: 20959788]
50. Hessels A, Flynn L, Cimiotti JP, Bakken S, Gershon R. Impact of heath information technology on the quality of patient care. Online J Nurs Inform 2015;19:1 [FREE Full text] [Medline: 27570443]
51. Howley MJ, Chou EY, Hansen N, Dalrymple PW. The long-term financial impact of electronic health record implementation. J Am Med Inform Assoc 2015 Mar;22(2):443-452. [doi: 10.1136/amiajnl-2014-002686] [Medline: 25164255]
52. Jones SS, Adams JL, Schneider EC, Ringel JS, McGlynn EA. Electronic health record adoption and quality improvement in US hospitals. Am J Manag Care 2010 Dec;16(12 Suppl HIT):SP64-SP71 [FREE Full text] [Medline: 21314225]
53. Katzer R, Barton DJ, Adelman S, Clark S, Seaman EL, Hudson KB. Impact of implementing an EMR on physical exam documentation by ambulance personnel. Appl Clin Inform 2012 Jul 25;3(3):301-308 [FREE Full text] [doi: 10.4338/ACI-2012-03-RA-0008] [Medline: 23646077]
54. Kritz S, Brown LS, Chu M, John-Hull C, Madray C, Zavala R, et al. Electronic medical record system at an opioid agonist treatment programme: study design, pre-implementation results and post-implementation trends. J Eval Clin Pract 2012 Aug;18(4):739-745 [FREE Full text] [doi: 10.1111/j.1365-2753.2011.01664.x] [Medline: 21414112]
55. Lam JG, Lee BS, Chen PP. The effect of electronic health records adoption on patient visit volume at an academic ophthalmology department. BMC Health Serv Res 2016 Jan 13;16:7 [FREE Full text] [doi: 10.1186/s12913-015-1255-8] [Medline: 26762304]
56. Lowe JR, Raugi GJ, Reiber GE, Whitney JD. Does incorporation of a clinical support template in the electronic medical record improve capture of wound care data in a cohort of veterans with diabetic foot ulcers? J Wound Ostomy Continence Nurs 2013;40(2):157-162 [FREE Full text] [doi: 10.1097/WON.0b013e318283bcd8] [Medline: 23466720]
57. McCullough JS, Casey M, Moscovice I, Prasad S. The effect of health information technology on quality in U.S. hospitals. Health Aff (Millwood) 2010 Apr;29(4):647-654. [doi: 10.1377/hlthaff.2010.0155] [Medline: 20368594]
58. McCullough CM, Wang JJ, Parsons AS, Shih SC. Quality measure performance in small practices before and after electronic health record adoption. EGEMS (Wash DC) 2015 Jan 6;3(1):1131 [FREE Full text] [doi: 10.13063/2327-9214.1131] [Medline: 25848635]
59. Mitchell J, Probst J, Brock-Martin A, Bennett K, Glover S, Hardin J. Association between clinical decision support system use and rural quality disparities in the treatment of pneumonia. J Rural Health 2014;30(2):186-195. [doi: 10.1111/jrh.12043] [Medline: 24689543]
60. Patterson ME, Marken P, Zhong Y, Simon SD, Ketcherside W. Comprehensive electronic medical record implementation levels not associated with 30-day all-cause readmissions within Medicare beneficiaries with heart failure. Appl Clin Inform 2014 Jul 30;5(3):670-684 [FREE Full text] [doi: 10.4338/ACI-2014-01-RA-0008] [Medline: 25298808]
61. Persell SD, Kaiser D, Dolan NC, Andrews B, Levi S, Khandekar J, et al. Changes in performance after implementation of a multifaceted electronic-health-record-based quality improvement system. Med Care 2011 Feb;49(2):117-125. [doi: 10.1097/MLR.0b013e318202913d] [Medline: 21178789]
62. Radley DC, Wasserman MR, Olsho LE, Shoemaker SJ, Spranca MD, Bradshaw B. Reduction in medication errors in hospitals due to adoption of computerized provider order entry systems. J Am Med Inform Assoc 2013 May 01;20(3):470-476 [FREE Full text] [doi: 10.1136/amiajnl-2012-001241] [Medline: 23425440]
63. Rao SR, Desroches CM, Donelan K, Campbell EG, Miralles PD, Jha AK. Electronic health records in small physician practices: availability, use, and perceived benefits. J Am Med Inform Assoc 2011 May 01;18(3):271-275 [FREE Full text] [doi: 10.1136/amiajnl-2010-000010] [Medline: 21486885]
64. Risko N, Anderson D, Golden B, Wasil E, Barrueto F, Pimentel L, et al. The impact of electronic health record implementation on emergency physician efficiency and patient throughput. Healthc (Amst) 2014 Sep;2(3):201-204. [doi: 10.1016/j.hjdsi.2014.06.003] [Medline: 26250507]
65. Ryan AM, McCullough CM, Shih SC, Wang JJ, Ryan MS, Casalino LP. The intended and unintended consequences of quality improvement interventions for small practices in a community-based electronic health record implementation project. Med Care 2014 Sep;52(9):826-832. [doi: 10.1097/MLR.0000000000000186] [Medline: 25100231]
66. Schreiber R, Shaha SH. Computerised provider order entry adoption rates favourably impact length of stay. J Innov Health Inform 2016 Apr 18;23(1):166 [FREE Full text] [doi: 10.14236/jhi.v23i1.166] [Medline: 27348485]
67. Scott DJ, Labro E, Penrose CT, Bolognesi MP, Wellman SS, Mather 3rd RC. The impact of electronic medical record implementation on labor cost and productivity at an outpatient orthopaedic clinic. J Bone Joint Surg Am 2018 Sep 19;100(18):1549-1556. [doi: 10.2106/JBJS.17.01339] [Medline: 30234619]
68. Shen JJ, Cochran CR, Neish S, Moseley CB, Mukalian R. Level of EHR adoption and quality and cost of care - evidence from vascular conditions and procedures. Int J Healthcare Technol Manag 2015 Jul 13;15(1):4-21. [doi: 10.1504/ijhtm.2015.070514]
69. Silow-Carroll S, Edwards JN, Rodin D. Using electronic health records to improve quality and efficiency: the experiences of leading hospitals. Issue Brief (Commonw Fund) 2012 Jul;17:1-40. [Medline: 22826903]

70. Singh RP, Bedi R, Li A, Kulkarni S, Rodstrom T, Altus G, et al. The practice impact of electronic health record system implementation within a large multispecialty ophthalmic practice. JAMA Ophthalmol 2015 Jun;133(6):668-674. [doi: 10.1001/jamaophthalmol.2015.0457] [Medline: 25880083]
71. Thirukumaran CP, Dolan JG, Reagan Webster P, Panzer RJ, Friedman B. The impact of electronic health record implementation and use on performance of the Surgical Care Improvement Project measures. Health Serv Res 2015 Feb;50(1):273-289 [FREE Full text] [doi: 10.1111/1475-6773.12191] [Medline: 24965357]
72. Tidwell C, Wolfberg A, Corrigan H. Leveraging technology to improve finances and care coordination in a small rural practice. Obstetrics Gynecology 2016 May;127:54S-55S. [doi: 10.1097/01.aog.0000483870.50355.52]
73. Varpio L, Day K, Elliot-Miller P, King JW, Kuziemsky C, Parush A, et al. The impact of adopting EHRs: how losing connectivity affects clinical reasoning. Med Educ 2015 May;49(5):476-486. [doi: 10.1111/medu.12665] [Medline: 25924123]
74. Walker-Czyz A. The impact of an integrated electronic health record adoption on nursing care quality. J Nurs Adm 2016;46(7-8):366-372. [doi: 10.1097/NNA.0000000000000360] [Medline: 27379908]
75. Wang JJ, Sebek KM, McCullough CM, Amirfar SJ, Parsons AS, Singer J, et al. Sustained improvement in clinical preventive service delivery among independent primary care practices after implementing electronic health record systems. Prev Chronic Dis 2013 Aug 01;10:E130 [FREE Full text] [doi: 10.5888/pcd10.120341] [Medline: 23906330]
76. Xiao AQ, Acosta FX. Implementation and impact of psychiatric electronic medical records in a public medical center. Perspect Health Inf Manag 2016 Oct 1;13(Fall):1e [FREE Full text] [Medline: 27843422]
77. Wani D, Malhotra M. Does the meaningful use of electronic health records improve patient outcomes? J Oper Manag 2018 Jun 22;60(1):1-18. [doi: 10.1016/j.jom.2018.06.003]
78. Zhou L, Soran CS, Jenter CA, Volk LA, Orav EJ, Bates DW, et al. The relationship between electronic health record use and quality of care over time. J Am Med Inform Assoc 2009;16(4):457-464 [FREE Full text] [doi: 10.1197/jamia.M3128] [Medline: 19390094]
79. Marmor RA, Clay B, Millen M, Savides TJ, Longhurst CA. The impact of physician EHR usage on patient satisfaction. Appl Clin Inform 2018 Jan;9(1):11-14 [FREE Full text] [doi: 10.1055/s-0037-1620263] [Medline: 29298451]
80. Baker J, Song J, Jones DR. Closing the loop: empirical evidence for a positive feedback model of IT business value creation. J Strategic Inf Syst 2017 Jun;26(2):142-160. [doi: 10.1016/j.jsis.2016.12.001]

Abbreviations

CMS: Center for Medicare and Medicaid Services
CPOE: certified physician order entry
EHR: electronic health record
HIT: health IT
HITECH: Health Information Technology for Economic and Clinical Health
LOS: length of stay
MU: meaningful use
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
ROI: return on investment

Edited by C Lovis; submitted 14.02.22; peer-reviewed by M Pinto da Costa, C Ta; comments to author 27.03.22; revised version received 10.05.22; accepted 31.07.22; published 27.09.22.

Please cite as:
Modi S, Feldman SS
The Value of Electronic Health Records Since the Health Information Technology for Economic and Clinical Health Act: Systematic Review
JMIR Med Inform 2022;10(9):e37283
URL: https://medinform.jmir.org/2022/9/e37283
doi:10.2196/37283
PMID:36166286

Corrigenda and Addenda

Correction: Identifying Patients Who Meet Criteria for Genetic Testing of Hereditary Cancers Based on Structured and Unstructured Family Health History Data in the Electronic Health Record: Natural Language Processing Approach

Jianlin Shi[1,2,3], MS, MD, PhD; Keaton L Morgan[3,4], MS, MD; Richard L Bradshaw[3], MS, PhD; Se-Hee Jung[3,5], BSN; Wendy Kohlmann[6,7], MS; Kimberly A Kaphingst[7,8], SCD; Kensaku Kawamoto[3], MPH, MD, PhD; Guilherme Del Fiol[3], MD, PhD

[1]Veterans Affairs Informatics and Computing Infrastructure, Department of Veterans Affairs Salt Lake City Health Care System, Salt Lake City, UT, United States

[2]Division of Epidemiology, Department of Internal Medicine, School of Medicine, University of Utah, Salt Lake City, UT, United States

[3]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, United States

[4]Department of Emergency Medicine, University of Utah, Salt Lake City, UT, United States

[5]College of Nursing, University of Utah, Salt Lake City, UT, United States

[6]Department of Population Health Sciences, University of Utah, Salt Lake City, UT, United States

[7]Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, United States

[8]Department of Communication, University of Utah, Salt Lake City, UT, United States

Corresponding Author:
Guilherme Del Fiol, MD, PhD
Department of Biomedical Informatics
University of Utah
421 Wakara Way
Ste 140
Salt Lake City, UT, 84108-3514
United States
Phone: 1 801 581 4080
Fax: 1 801 581 4297
Email: guilherme.delfiol@utah.edu

Related Article:

Correction of: https://medinform.jmir.org/2022/8/e37842

(JMIR Med Inform 2022;10(9):e42533) doi:10.2196/42533

In “Identifying Patients Who Meet Criteria for Genetic Testing of Hereditary Cancers Based on Structured and Unstructured Family Health History Data in the Electronic Health Record: Natural Language Processing Approach” (JMIR Med Inform 2022;10(8):e37842), the authors noted the following corrections:

(1) In the originally published paper, the following sentence was present in the *Methods* section of the *Abstract*:

> *Algorithms were developed based on the National Comprehensive Cancer Network (NCCN) guidelines for genetic testing for hereditary breast or ovarian and colorectal cancers.*

This has been changed to:

> *Algorithms were developed based on the National Comprehensive Cancer Network (NCCN) guidelines for genetic testing for hereditary breast, ovarian, pancreatic, and colorectal cancers.*

(2) Textbox 1 has been revised for clarity and accuracy, and to comply with the citation guidelines of the National Comprehensive Cancer Network (NCCN).

(3) In the originally published paper, the following sentence was present in the *Background* section:

> *The National Comprehensive Cancer Network (NCCN) has published a set of evidence-based guidelines for genetic testing of hereditary cancers, including breast, ovarian, and colorectal cancers.*

This has been changed to:

> *The National Comprehensive Cancer Network (NCCN) has published a set of evidence-based*

> *guidelines for genetic testing of hereditary cancers, including breast, ovarian, pancreatic, and colorectal cancers.*

(4) References [6] and [7] have been updated to the following with NCCN's permission and disclaimer statements:

> *6. Daly MB, Pilarski R, Yurgelun MB, Berry MP, Buys SS, Dickson P, et al. NCCN guidelines insights: genetic/familial high-risk assessment: breast, ovarian, and pancreatic, version 1.2020. J Natl Compr Canc Netw 2020 Apr;18(4):380-391 [Referenced with permission from the National Comprehensive Cancer Network, Inc. 2020]. [doi: 10.6004/jnccn.2020.0017] [Medline: 32259785]*
>
> *7. Gupta S, Provenzale D, Llor X, Halverson AL, Grady W, Chung DC, et al. NCCN guidelines insights: genetic/familial high-risk assessment: colorectal, version 2.2019. J Natl Compr Canc Netw 2019 Sep 01;17(9):1032-1041 [Referenced with permission from the National Comprehensive Cancer Network, Inc. 2020]. [doi: 10.6004/jnccn.2019.0044] [Medline: 31487681]*

Textbox 1. Excerpt of National Comprehensive Cancer Network (NCCN) criteria for unaffected individuals' family history–based genetic testing of breast, ovarian, pancreatic, and colorectal cancers (referenced with permission).

Breast or ovarian cancer:

1. First- or second-degree relative with breast cancer at age ≤45 years
2. First- or second-degree relative with ovarian cancer
3. First-degree relative with pancreatic cancer
4. Breast cancer in a male relative
5. Three or more first- or second-degree relatives with breast or prostate cancer on the same side of the family
6. Ashkenazi Jewish and any breast or prostate cancer in any relative at any age
7. BRCA1/2, CHEK2, ATM, PALB2, TP53, PTEN, or CDH1 genes, Cowden Syndrome, Li-Fraumeni Syndrome in any relative at any age

Colorectal cancer:

1. MLH1, MSH2, PMS2, MSH6, EPCAM, MYH, or MUTYH genes, Lynch syndrome, familial adenomatous polyposis (FAP), adenomatous polyposis coli (APC), serrated polyposis or polyposis discovered in the coded family history
2. First-degree relative with colon cancer at ≤50 years
3. First-degree relative with endometrial cancer at ≤50 years
4. Three or more first- or second-degree relatives with Lynch syndrome, HNPCC, colon cancer, endometrial, uterine, ovarian, stomach, gastric, small bowel, small intestine, kidney, ureteral, bladder, urethra, brain, pancreas, also all on the same side of the family

The correction will appear in the online version of the paper on the JMIR Publications website on September 13, 2022, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

References

6. Daly MB, Pilarski R, Yurgelun MB, Berry MP, Buys SS, Dickson P, et al. NCCN guidelines insights: genetic/familial high-risk assessment: breast, ovarian, and pancreatic, version 1.2020. J Natl Compr Canc Netw 2020 Apr;18(4):380-391 [Referenced with permission from the National Comprehensive Cancer Network, Inc. 2020]. [doi: 10.6004/jnccn.2020.0017] [Medline: 32259785]
7. Gupta S, Provenzale D, Llor X, Halverson AL, Grady W, Chung DC, et al. NCCN guidelines insights: genetic/familial high-risk assessment: colorectal, version 2.2019. J Natl Compr Canc Netw 2019 Sep 01;17(9):1032-1041 [Referenced with permission from the National Comprehensive Cancer Network, Inc. 2020]. [doi: 10.6004/jnccn.2019.0044] [Medline: 31487681]

Submitted 07.09.22; this is a non–peer-reviewed article;accepted 07.09.22; published 13.09.22.

Please cite as:
Shi J, Morgan KL, Bradshaw RL, Jung SH, Kohlmann W, Kaphingst KA, Kawamoto K, Fiol GD
Correction: Identifying Patients Who Meet Criteria for Genetic Testing of Hereditary Cancers Based on Structured and Unstructured Family Health History Data in the Electronic Health Record: Natural Language Processing Approach
JMIR Med Inform 2022;10(9):e42533
URL: https://medinform.jmir.org/2022/9/e42533
doi:10.2196/42533
PMID:36099593

Original Paper

Identifying the Perceived Severity of Patient-Generated Telemedical Queries Regarding COVID: Developing and Evaluating a Transfer Learning–Based Solution

Joseph Gatto[1], BA; Parker Seegmiller[1], BSc; Garrett Johnston[1], BA; Sarah Masud Preum[1], BSc, MSc, PhD

Department of Computer Science, Dartmouth College, Hanover, NH, United States

Corresponding Author:
Joseph Gatto, BA
Department of Computer Science
Dartmouth College
15 Thayer Drive
Hanover, NH, 03755
United States
Phone: 1 603 646 1110
Email: joseph.m.gatto.gr@dartmouth.edu

Abstract

Background: Triage of textual telemedical queries is a safety-critical task for medical service providers with limited remote health resources. The prioritization of patient queries containing medically severe text is necessary to optimize resource usage and provide care to those with time-sensitive needs.

Objective: We aim to evaluate the effectiveness of transfer learning solutions on the task of telemedical triage and provide a thorough error analysis, identifying telemedical queries that challenge state-of-the-art natural language processing (NLP) systems. Additionally, we aim to provide a publicly available telemedical query data set with labels for severity classification for telemedical triage of respiratory issues.

Methods: We annotated 573 medical queries from 3 online health platforms: HealthTap, HealthcareMagic, and iCliniq. We then evaluated 6 transfer learning solutions utilizing various text-embedding strategies. Specifically, we first established a baseline using a lexical classification model with term frequency–inverse document frequency (TF-IDF) features. Next, we investigated the effectiveness of global vectors for text representation (GloVe), a pretrained word-embedding method. We evaluated the performance of GloVe embeddings in the context of support vector machines (SVMs), bidirectional long short-term memory (bi-LSTM) networks, and hierarchical attention networks (HANs). Finally, we evaluated the performance of contextual text embeddings using transformer-based architectures. Specifically, we evaluated bidirectional encoder representation from transformers (BERT), Bio+Clinical-BERT, and Sentence-BERT (SBERT) on the telemedical triage task.

Results: We found that a simple lexical model achieved a mean F1 score of 0.865 (SD 0.048) on the telemedical triage task. GloVe-based models using SVMs, HANs, and bi-LSTMs achieved a 0.8-, 1.5-, and 2.1-point increase in the F1 score, respectively. Transformer-based models, such as BERT, Bio+Clinical-BERT, and SBERT, achieved a mean F1 score of 0.914 (SD 0.034), 0.904 (SD 0.041), and 0.917 (SD 0.037), respectively. The highest-performing model, SBERT, provided a statistically significant improvement compared to all GloVe-based and lexical baselines. However, no statistical significance was found when comparing transformer-based models. Furthermore, our error analysis revealed highly challenging query types, including those with complex negations, temporal relationships, and patient intents.

Conclusions: We showed that state-of-the-art transfer learning techniques work well on the telemedical triage task, providing significant performance increase over lexical models. Additionally, we released a public telemedical triage data set using labeled questions from online medical question-and-answer (Q&A) platforms. Our analysis highlights various avenues for future works that explicitly model such query challenges.

(JMIR Med Inform 2022;10(9):e37770) doi:10.2196/37770

KEYWORDS

natural language processing; transfer learning; telemedicine triage; COVID-19; health resource; health care; patient query; learning solution; telemedical; lexical model; machine learning; BERT; telehealth

Introduction

Background

The COVID-19 pandemic has led to an increased demand for telemedicine services [1]. Projections state that up to 50% of consultations could be performed through telehealth by 2025 for certain demographic groups [2]. The general demographic makeup of telemedicine patients, 1 study finds, is most often White English-speaking females using private medical insurance, with minority groups using significantly less telemedical services [3]. Patients use these services to communicate with a diverse set of medical specialists, including dentists, rheumatologists, and prenatal care specialists, among many others, all with high levels of satisfaction [2]. These studies, however, examine patients who utilize telemedicine services to interact with their existing care providers in a remote setting. Recently, there has been a rise in affordable and accessible telemedicine platforms that connect anyone with an internet connection to licensed medical professionals worldwide. Such platforms include *HealthTap* [4], *iCliniq*, and *HealthcareMagic*. HealthTap, for example, is a Health Insurance Portability and Accountability Act (HIPAA)–certified website that provides online users with access to a qualified doctor with an active US medical license [4]. These platforms are easy to access and provide greater accessibility to professional medical consultation. However, such ease and accessibility can cause these services to be flooded with questions that are *not* medically severe or relevant. This is a safety-critical problem, as an abundance of nonsevere medical queries will hinder the speed at which medical professionals can respond to time-sensitive issues. Evidence of this phenomenon was observed with COVID-19 hotlines, where confusion about coronavirus caused long wait times, with Margolius et al [5] finding that of the 12,512 calls made to their triage system between March 13 and April 20, 2020, “52% were not COVID-19 related or required no additional care.” Large numbers of nonsevere telemedical queries not only are dangerous but also may unnecessarily increase health care spending, as telemedical service convenience may encourage patients to inquire about negligible health concerns [6].

This situation necessitates a system for prioritizing which queries require immediate care. To address this problem, we examined data from 3 telemedicine platforms: HealthTap, iCliniq, and HealthcareMagic. These platforms facilitate written medical queries to be answered remotely by licensed doctors. Our goal was to optimize the time spent by health care workers by ranking patient queries by severity so that potentially severe queries are answered first. In this study, a query was deemed severe when a patient had at least 1 active COVID-19– or pneumonia-related symptom. Nonsevere queries, however, were from patients with no active symptoms who submitted general information requests, nonsensical text, or extremely vague questions. Telemedical triage conserves the limited amount of professional health care provider resources available to telemedicine platforms by prioritizing severe queries, encouraging remote medical care to be provided to those most desperately in need.

In this work, we examined telemedical triage through the lens of online medical question-and-answer (Q&A) forums. Specifically, we formulated triage as a binary text classification problem, where we aimed to classify medical queries [7] as either severe or not severe. We noted that this formulation does not capture the full spectrum of severity but is useful for lowering the priority of queries with no medical urgency. To do so, we introduced an extension to the publicly available data set *COVID-Dialogue* [8], which contains 603 doctor-patient conversations extracted from HealthTap, iCliniq, and HealthcareMagic, with labels for severe query classification. Given the limited number of available samples, we then investigated various transfer learning approaches to text classification and contrasted them with a lexical approach. Specifically, we explored the applicability of different embedding methods, such as global vectors for text representation (GloVe) and transformers, pretrained using both general and medical texts, in comparison to term frequency–inverse document frequency (TF-IDF) features. Our experiments showed that transformer-based solutions are a superior transfer learning approach for identifying severe medical queries. Additionally, we found that pretraining on medical texts provides no benefit when classifying our telemedical triage data set. Finally, we provided an in-depth error analysis of sentence bidirectional encoder representation from transformers (sentence-BERT or SBERT) [9], identifying challenging patient query patterns that motivate future work on telemedical triage. Specifically, we noticed difficulties in modeling the negation, temporality, and intent of symptom mentions in patient-generated textual queries.

Our contributions are as follows:

- We established baseline results across 6 relevant natural language processing (NLP) models on the telemedical triage task—identifying optimal pretraining strategies for query ranking according to severity. We identified contextual embedding models to work best for triage, with all transformer-based approaches achieving statistically significant improvements over both lexical and word embedding–based approaches. We found no benefit of pretraining transformer models with clinical text.
- We provided a thorough error analysis of SBERT and identified several medical query types that pose difficulties to NLP systems—where the core challenge is identified as the modeling of complex symptom presentations.
- To the best of our knowledge, we have provided the first publicly available telemedical triage classification data set using real samples from online telemedical services. All code and data for this study have been made publicly available [7].

Related Works

Transfer Learning for Medical Text Classification

The need for data privacy and patient anonymity makes large-scale collection and labeling of health care texts extremely difficult. This has motivated the use of transfer learning in medical NLP to alleviate the challenges in resource-constrained modeling. In recent years, transfer learning has benefited greatly from leveraging large amounts of unlabeled text to train transformer-based models [10] using a masked language modeling objective. This framework allows common linguistic pattern understanding to be transferred to other downstream tasks, reducing the need for large amounts of labeled data to solve various low-resource problems.

Bidirectional encoder representation from transformers (BERT) [11] is a popular transformer-based model that has been pretrained on general text, such as the Wikipedia and Brown corpus. Medical inference tasks that require domain-specific linguistic knowledge, such as medical natural language inference [12] or medical concept extraction [13], have been shown to benefit significantly from pretraining on medical-specific text [14]. In this study, we explored transfer learning using both general and medical text as pretraining methods. Analysis of optimal pretraining strategies for telemedical triage is of interest as patient queries are typically composed of common language but often interwoven with complex medical terminology.

Machine Learning for Triage

The COVID-19 pandemic overwhelmed the US health care system, bringing about a demand for machine learning solutions for the telemedicine triage problem. For example, Lai et al [15] found that COVID-19 hotlines became overwhelmed with calls, prompting the development of an artificial intelligence (AI) system of patient ranking. Lai et al [15] used an AI chatbot to prescreen callers by asking questions that reveal whether a patient has COVID-19 symptoms. The resulting information is fed to a logic-based inference model, which determines whether further consultation from the hotline's health care workers is required.

Hospital emergency departments (EDs) have similarly become overwhelmed with patients [16], causing dangerous treatment delays for those coming to the ED with an urgent need. Yao et al [16] trained a deep learning model to predict which patients will eventually require hospitalization by using incoming patient emergency medical records. This allows for an automated system of patient care prioritization that does not depend on nursing resources. Similar work was done by Gligorijevic et al [17], where a deep attention model was used to triage patients using multimodal electronic health record (EHR) data, where triage was formulated as a classification problem based on the Emergency Severity Index [18].

Unlike the aforementioned work, we viewed triage solely through the lens of textual queries—specifically those submitted by patients to telemedicine platforms. With rising demand for textual medical support, through either public medical Q&A platforms, such as HealthTap, or private doctor-patient messaging apps, we foresee a growing need for NLP solutions for the triage problem over free-text patient queries.

Medical Risk Identification

A similar work for triage of telemedicine platform messages was performed by Si et al [19], who classified doctor-patient messages based on their urgency by using data collected from adults at a university hospital. The data are unfortunately not public. Additionally, the message content contains queries about cardiology from patients to their existing cardiologists. This is in stark contrast to our data set, which is public, has a different label space, and contains samples about respiratory illnesses from patients to doctors they have never spoken to before (and thus are unable to make decisions using prior knowledge of medical history). Both our work and that of Si et al [19], however, explore BERT-based solutions to triage.

Additional, similar work resides in the realm of medical risk identification from text. For example, Fu et al [20] introduced a knowledge graph–based distant supervision approach to suicide risk prediction from social media posts, Wang et al [21] explored transformer-based solutions for depression risk prediction from social media data, and Klein et al [22] applied a BERT-based classifier to identify potential COVID-19 cases from tweets.

This work is similar in that we explored BERT-based solutions for medical risk identification. However, unlike social media data, medical queries submitted to telemedicine platforms often contain complex clinical terminology. Furthermore, the telemedical services through which doctors interact with patients contain less restrictive character limits, requiring modeling of long-range textual dependencies. Finally, social media–based studies have the luxury of large-scale data mining. In this study, we operated in an extremely resource-constrained data setting, which challenged our capacity to model and understand medical query text.

Methods

Data Set

In this study, we utilized the publicly available COVID-Dialogue data set [8]. This data set contains 603 anonymized patient queries extracted from 3 telemedicine platforms, namely HealthTap, iCliniq, and HealthcareMagic. The original data set was collected with the intention to facilitate better AI dialogue systems during the COVID-19 pandemic. Thus, each of the 603 doctor-patient conversations includes the full patient query, a summarized patient query, and the doctor's response. The data set was not curated for text classification; thus, after filtering samples unusable in our classification setting (ie, duplicate, non-English, and out-of-scope entries—where "out of scope" is defined as entries not regarding COVID-19 or pneumonia symptoms), we annotated 573 (95%) samples. All multiturn dialogues were truncated to the initial patient utterance, and no doctor responses were used in our pipeline.

Each sample in the COVID-Dialogue data set contains queries regarding either COVID-19 or related pneumonia symptoms. Each sample includes no patient demographics or medical history; thus, severity was detected solely using a single free-text inquiry. Table 1 provides an example from each class in our labeled data set. The general goal of our labeling schema was to prioritize those with active symptoms and reduce the priority of the hundreds of samples that exhibit no medically severe text. Our final data set contained 314 (55%) severe samples and 259 (45%) nonsevere samples.

Table 1. Samples from the COVID-Dialogue data set with our introduced severity label. Nonsevere samples are often irrelevant queries or from patients with little to no symptoms. Severe samples always contain patients with active symptoms that may require medical attention.

Patient query	Ground truth label
"Should I shave my beard to reduce my chances of contracting coronavirus/covid-19?"	Not severe
"My daughter is 11 years old she has has pneumonia she has been sick since January 3rd symptoms keep changing. she is up at night itching all over her upper torso, head, and ears. She has major headache and abdominal pain."	Severe

Ethical Considerations

Given that the data are publicly available, no Institutional Review Board approval was required for this study.

The data set used for the development and evaluation of the solution is anonymous and does not reveal the identity of doctors and patients. No demographic information is available for this data set.

We consulted 3 professional health care providers about the real-world implications of such a telemedicine triage system. We consulted Drs Timothy E. Burdick, Stephen K. Liu, and Jiazuo H. Feng. All of them serve as primary care providers at a local teaching hospital. An interesting question regarding the ethical use of future telemedical triage systems is whether to include demographic, socioeconomic, physiological, or other EHR information in future medical triage systems, given such information is available. Although demographic or past medical history (eg, age of the patient, pre-existing conditions) might be relevant to determine the actual severity of the patient's query, sucn information can also introduce bias. Related works on telemedical triage, such as Si et al [19], similarly propose the use of demographic information in future work. Determining the fairness and equity of such systems would require additional exploration with additional ground truth from the user that is available in the EHR data, including but not limited to emergency visits, urgent care visits, and scheduling new appointments immediately after receiving a response from the care provider. This is out of scope of this paper. However, we are currently designing a study to investigate this question by using EHR data collected from a local hospital, as mentioned later.

Data Sources

Next, we describe the sources used in the collection of the COVID-Dialogue data set [8], which was publicly released in March 2020. Samples in this data set were collected between February 7 and March 25, 2020.

HealthTap

Founded in 2010, HealthTap is a telemedicine platform that remotely connects patients with US licensed medical professionals for a variety of services, including virtual consultations and doctor-patient Q&A. According to Dahl [23], patients have had close to 1 billion questions answered on HealthTap. Additionally, HealthTap accepts over 100 insurance plans and employs doctors from over 140 specialties. HealthTap data in the COVID-Dialogue data set were collected from its medical Q&A forum.

iCliniq

iCliniq is a virtual hospital providing video, voice, and text chat medical services to patients worldwide. iCliniq works with more than 3500 licensed doctors internationally, covering over 80 medical specialties. Samples from iCliniq were drawn from its medical Q&A forum.

HealthcareMagic

Unlike Healthtap and iCliniq, HealthcareMagic is strictly an online medical Q&A forum. With over 18,000 doctors across 78 medical specialties, 1.7 million questions have been answered on HealthcareMagic.

Annotation Details

Each sample in our data set was annotated by 3 of the authors as either severe or nonsevere. Use of authors as annotators for small-scale medical web information has been successful in other studies [24,25]. Each annotator has a college degree and an adequate level of health literacy and has invested significant time to educate themselves on the potential symptoms associated with the 2 illnesses observed in this data set. We noted that the use of nonmedical professionals limited the degree of granularity with which we could label this data set. However, the annotators carefully reviewed the response to the original query to determine potential severe queries. In addition, the annotators observed that there were a lot of irrelevant samples compared to those exhibiting significant symptoms. For example, the following query can be safely annotated as "not severe" since it does not warrant significant medical knowledge to answer "Where can I get a COVID-19 test?" This question can be answered using Google Search for most parts of the United States. We noted, however, that assuming internet search availability might bias annotation against those in rural, remote areas without reliable access to the internet and familiarity with a web search for health. However, those without access to Google Search or an intent to use a web search for health issues would be also less likely to rely on telemedicine services [2].

We additionally noted that there might be some samples where the perceived severity based on the query and the response from the medical professional would be different from the actual severity of the condition of the patient. However, since we do not have any ground truth from the actual user, such cases cannot be resolved. This motivated us to pursue future work in this direction by utilizing our collaboration with doctors in a local hospital, as reported in the Future Work section. In addition, we performed a thorough error analysis of the performance of our proposed solution and illustrated its strengths and limitations with respect to this annotated data set. The final annotation for each sample was the majority vote label from the 3 annotators. The interannotator agreement across all samples was 82%. Next, we detail the annotation schema for both nonsevere and severe samples.

Not Severe

The guiding principle behind a *nonsevere* annotation was a patient query that did not indicate an active symptom, an immediate need for diagnosis, or an immediate need for a medical response. This included queries that were unspecific or speculative. Examples selected from the data set and their nonsevere annotation rationale are listed next.

> *Where can I get a covid test?*

This query does not indicate immediate danger, a need for diagnosis, or a need for a medical response. This query can also be served by Google Search and thus does not need feedback from a medical professional.

> *Will I have to be hospitalised if I get the virus, I have type 1 diabetes.*

Although this query is medically valid and deserves a response, the need is not deemed immediate as the patient has no active symptoms.

Severe

A *severe* annotation was given to a patient who indicated an active symptom that may present danger to the patient, an immediate need for diagnosis, or an immediate need for a medical response. This included queries in which a patient listed current symptoms or demonstrated a need for actionable doctor advice. Examples selected from the data set and their severe annotation rationale are listed next.

> *My son is not feeling well. He has a very snotty nose, sore throat, occasional flemmy cough, uneasy stomach. He had a headache last night. No fever. Is it a common cold or must he be checked for Covid 19. Not travelled or been in contact with anyone?*

This query describes symptoms that are consistent with COVID-19 and demonstrate a sufficient need for medical advice.

> *Preauricular lymph node on left side very tender, scalp on left side of head tender and hurts to touch, superficial parotid lymph node area on left side swollen and tender. Pain behind both ears. No injury. Came on suddenly, has been 1 day. Temp 100.1°.*

This query contains a clear, immediate danger to the patient and requires a medical response.

Transfer Learning Methods

Bidirectional Encoder Representation From Transformers

BERT is a state-of-the-art transformer-based model that leverages unlabeled text to produce contextualized language representations [11]. In this study, we used standard BERT for the text classification pipeline outlined by Devlin et al [11], where we first generated contextualized text features using a pretrained BERT model, followed by feeding the special CLS token to a linear classification head, which outputs the final query label. We explored BERT for telemedical triage, as BERT has been shown to be successful in related tasks, such as depression risk prediction [21], suicide risk prediction [20], and COVID-19 case identification [22].

Bio+Clinical-BERT

The Bio+Clinical-BERT architecture is the same as that of BERT but with weights pretrained on medical texts. Specifically, this pretraining procedure first takes the BioBERT model [26], which is BERT fine-tuned on biomedical research text collected from PubMed. Next, BioBERT is fine-tuned on clinical notes from the Medical Information Mart for Intensive Care (MIMIC)-III database [27], producing Bio+Clinical-BERT [14]. The combination of BERT, biomedical research texts, and clinical notes was shown to significantly outperform BERT on the medical natural language inference task [12]. Thus, we use Bio+Clinical-BERT as our medically informed architectural baseline for the patient query task. This is our only transfer learning approach leveraging knowledge from medical texts.

SBERT With Triplet Loss

We also explored the effectiveness of SBERT [9] for telemedical triage. Unlike BERT, which learns to output a contextualized embedding for *every* input token, SBERT produces a single embedding for a given input. SBERT has proved effective in adjacent medical NLP tasks, such as COVID-19 vaccine sentiment analysis [28] and COVID-19 misinformation detection [29]. In this study, we used SBERT as it permits both text classification and useful methods of embedding interpretability.

To perform text classification with SBERT, we first fine-tuned an SBERT model to minimize the following triplet loss function:

where A is an anchor sample, P is a positive sample (same class as A), N is a negative sample (opposite class of A) and d is the cosine-similarity distance function. This objective can be interpreted as learning to push query embeddings from the same class close together in embedding space, while pushing samples from opposite classes further apart. The margin parameter α influences the distance between positive and negative pairs in embedding space. To generate a training triple, a given sample was randomly paired with a sample from the same and opposite classes. This process was repeated 10 times per sample, generating 4580 training triplets.

Using the embeddings from the fine-tuned SBERT model, we then trained a K-nearest neighbor (KNN) classifier using the Scikit-Learn package [30]. Specifically, we set the number of neighbors K=10, otherwise using the default parameters provided by Scikit-Learn (which uses the Minkowski distance metric with p=2). The KNN was trained using the same training set as all other experiments and then used to label the test set queries based on their relationship to the training samples in the embedding space.

Baseline Experiments

For TF-IDF+SVM, we fed the TF-IDF [31] feature vector from the patient query to the support vector machine (SVM) classifier [32]. This baseline examined the effectiveness of a simple lexical model on telemedical triage.

For GloVE+SVM, we obtained the pretrained GloVe [33] embedding for each word in the patient query and fed the mean vector to the SVM classifier. This baseline tested the performance of transfer learning without the use of contextual modeling.

A 2-layer bidirectional long short-term memory (bi-LSTM) model was trained on GloVe embeddings for classification. The bi-LSTM model examined the effectiveness of contextual sequence modeling on pretrained word embeddings. Bi-LSTM models have been shown to be effective in a variety of clinical text prediction tasks [34].

The hierarchical attention network (HAN) [35] for text classification mimics the natural language hierarchy by modeling attention at the sentence and word level for document classification. HAN word embeddings in this experiment were initialized using GloVe. Other prior works have established a HAN to be an effective classifier for medical text [36].

Evaluation Setting

For each experiment, we reported the weighted mean F1, precision, and recall scores over a 5-fold cross-validation. Additionally, we reported the 95% CI for the reported mean. Finally, we conducted statistical significance testing using the McNemar test [37] for each model with respect to our top-performing approach, SBERT. Each train and test split contained approximately 458 and 115 samples, respectively.

Results

In this section, we present our results of the telemedical triage task across various NLP baselines. Our goal is to answer the following research questions (RQs):

- How effective are transfer learning models for telemedical triage for COVID-19–related queries when compared to other text classification models?
- What types of health queries challenge state-of-the-art NLP systems?

Analysis

Our results showed that telemedical triage benefits greatly from transfer learning as our lowest-performing model, TF-IDF+SVM, used no transfer learning. TF-IDF features achieved a reasonable mean F1 score of 0.865 (SD 0.048). However, we found a 0.8-, 1.5-, and 2.1-point increase in the F1 score by applying GloVE-based models, such as GloVe+SVM, HAN, and bi-LSTM models, respectively. Generally, light-weight modeling options, such as TF-IDF and GloVe, report reasonable F1 scores and are thus viable solutions in cases with limited computational resources.

RQ1: How Effective Are Transfer-Learning Models on Telemedical Triage ?

We found transformer-based models to be the superior method of transfer learning, with BERT, Bio+Clinical BERT, and SBERT achieving mean F1 scores of 0.914 (SD 0.034), 0.904 (SD 0.041), and 0.917 (SD 0.037), respectively (Table 2). We noted that Bio+Clinical-BERT did not outperform the BERT baseline. This is likely due to the difference in language found in BERT versus Bio+Clinical-BERT training data. Clinical notes used to train Bio+Clinical-BERT are written by medical practitioners and thus far more technical than queries written by patients. Thus, although telemedical query texts do contain medical terminology, clinical note pretraining is not helpful in this setting.

Our results showed that SBERT, on average, is the best predictor of query severity, producing both the highest average F1, precision, and recall scores compared to other approaches. A higher recall is particularly important in the realm of triage as reducing false negatives is more important in such a safety-critical task.

Using the McNemar test for statistical significance, we found SBERT to perform significantly better than TF-IDF+SVM (*P*<.001), GloVE+SVM (*P*=.001), bi-LSTM (*P*=.03), and HAN (*P*=.001). However, tests for statistical significance failed when comparing SBERT with other transformer-based models, such as BERT (*P*=.81) and Clinical-BERT (*P*=.22). Thus, the difference in predictive distributions between transformer-based approaches is insignificant, with all being valid options for transfer learning solutions to telemedical triage.

Although the performance of the transformer-based models was high (all F1 scores>0.9), it is important to note that the general problem of telemedical triage is far from solved. This study looked at triage through the narrow lens of COVID-19– and pneumonia-related queries; diseases involving nonrespiratory complications would not be recognized by this system.

Table 2. Results displaying triage performance for all models. Each metric is the mean result over 5-fold cross-validation, surrounded by the 95% CI computed using the metric score for each validation fold.

Model	F1 score, mean (SD)	Precision, mean (SD)	Recall, mean (SD)
TF-IDF[a]+SVM[b]	0.865 (0.048)	0.871 (0.043)	0.865 (0.048)
GloVe[c]+SVM	0.873 (0.036)	0.878 (0.030)	0.874 (0.035)
Bi-LSTM[d]	0.886 (0.051)	0.880 (0.049)	0.879 (0.052)
HAN[e]	0.880 (0.035)	0.890 (0.031)	0.880 (0.033)
BERT[f]	0.914 (0.034)	0.917 (0.033)	0.914 (0.034)
Bio+Clinical-BERT	0.904 (0.041)	0.905 (0.040)	0.904 (0.041)
SBERT[g]	0.917 (0.037)	0.920 (0.034)	0.918 (0.036)

[a] TF-IDF: term frequency–inverse document frequency.

[b] SVM: support vector machine.

[c] GloVe: global vectors for text representation.

[d] Bi-LSTM: bidirectional long short-term memory.

[e] HAN: hierarchical attention network.

[f] BERT: bidirectional encoder representation from transformers.

[g] SBERT: sentence bidirectional encoder representation from transformers.

RQ2: What Types of Telemedical Queries Challenge State-of-the-Art NLP Systems?

In the previous section, we identified transformer-based models to be the most effective form of pretraining for triage. To identify telemedical queries that are difficult to triage, we investigated SBERT, as this architecture outputs a single embedding for an entire query, which is useful for interpretability.

We first visualized the SBERT embedding for each test sample in 1 of our test splits using t-distributed stochastic neighbor embedding (t-SNE) [38], projecting the 768D SBERT embeddings for each sample down to a 2D space that aims to preserve the embedding distance found in higher dimensions.

Figure 1 visualizes the projected embeddings of our test queries before and after SBERT was fine-tuned using triplet loss. We found that SBERT learned meaningful clusters that largely separated severe from nonsevere samples. Using K-means clustering, we produced Figure 2, where the convex hull of each cluster is highlighted.

We were interested in analyzing patient queries that did not fall into the correct cluster. Textbox 1 highlights all the false positives, as determined by the K-means clusters shown in Figure 2. Specifically, these are the nonsevere (blue) samples in Figure 2 that appear in the Severe cluster.

A common theme among the false positives was that *all samples mentioned a symptom or disease*. SBERT was thus dependent on symptom interpretation for proper query embedding and likely misinterpreted how some symptoms were being presented. A qualitative analysis of Textbox 1 highlighted the following challenges:

- Symptom negation: Samples 1 and 4 highlight how negated symptoms may confuse telemedical triage symptoms. When analyzing sample 1, for example, a triage system must understand that the mentions of dry cough, fever, and sore throat are to highlight their absence and are not indicative of severity.
- Symptom temporality: Samples 4, 5, and 11 all have symptom mentions amidst complex temporal relationships. Automatic triage systems must be able to identify that not all symptom mentions are active, while highlighting what symptoms pertain to a given query.
- Ambiguous questions: Samples 2, 3, and 5 highlight a tricky subset of what we call "ambiguous questions," where symptom mentions may occur but the purpose of the query is unclear or the proposed question is difficult to answer. Such samples were marked as not severe by the annotators.
- General queries: Samples 6, 7, 9, and 10 contain symptom mentions within the context of a general, nonsevere query. For example, the purpose of query 7, which came from a patient with pneumonia, was to obtain more information about how pneumonia manifests in the lung. This was not deemed severe by the annotators, as general-information requests should not be ranked higher over more relevant, specific, serious medical needs.
- Self-answered questions: Samples 8 and 11 contain valid explanations or resolutions to the problem being inquired about. For example, the patient in sample 11 had a continuous dry cough and sore throat. However, they had already taken all requisite COVID-19 precautions (COVID-19 test, self-quarantine). These samples were labeled not severe and may prove challenging for future telemedical triage systems.

Textbox 2 highlights all the false negatives, as determined by the K-means clusters shown in Figure 2. Specifically, these are the severe (red) samples in Figure 2 that appear in the Not Severe cluster.

A qualitative analysis of Textbox 2, which contains all false positives, identified other potential triage challenges:

- Sparse symptom representation: Sample 1 contains patient symptom mentions (boop, HIV, anemia), which contain little representation in the training data. As similar data sets expand the number of diseases for which they are able to triage, learning good representations of large symptom sets may prove challenging.
- Implicit symptom mentions: Sample 2 states they have all COVID-19 symptoms except fever. The reader understands this as meaning there is likely cough, loss of taste/smell, etc, present in the patient. SBERT is unable to make this inference, making this type of sample challenging.

In summary, SBERT's errors are focused on samples with complex symptom presentation. Future work on telemedical triage may focus on explicit modeling of presented symptoms such that temporality, negation, intention, and other linguistic phenomena are accounted for.

Figure 1. SBERT embeddings projected to 2 dimensions using t-SNE. The left image depicts how the test samples are distributed in the embedding space prior to triplet loss–based fine-tuning. The right image displays how SBERT learns to separate query embeddings in the embedding space. Note: The comp-1 and comp-2 axes denote the names of the 2 dimensions onto which t-SNE projects the 768D embeddings, where "comp" is short for "component". SBERT: sentence bidirectional encoder representation from transformers; t-SNE: t-distributed stochastic neighbor embedding.

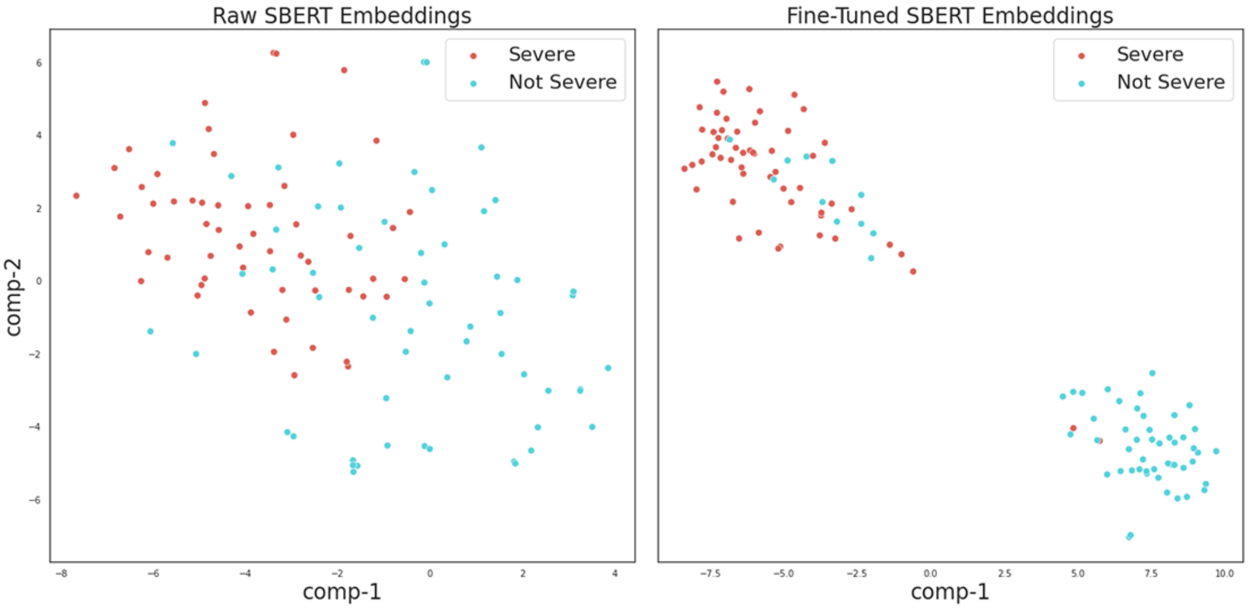
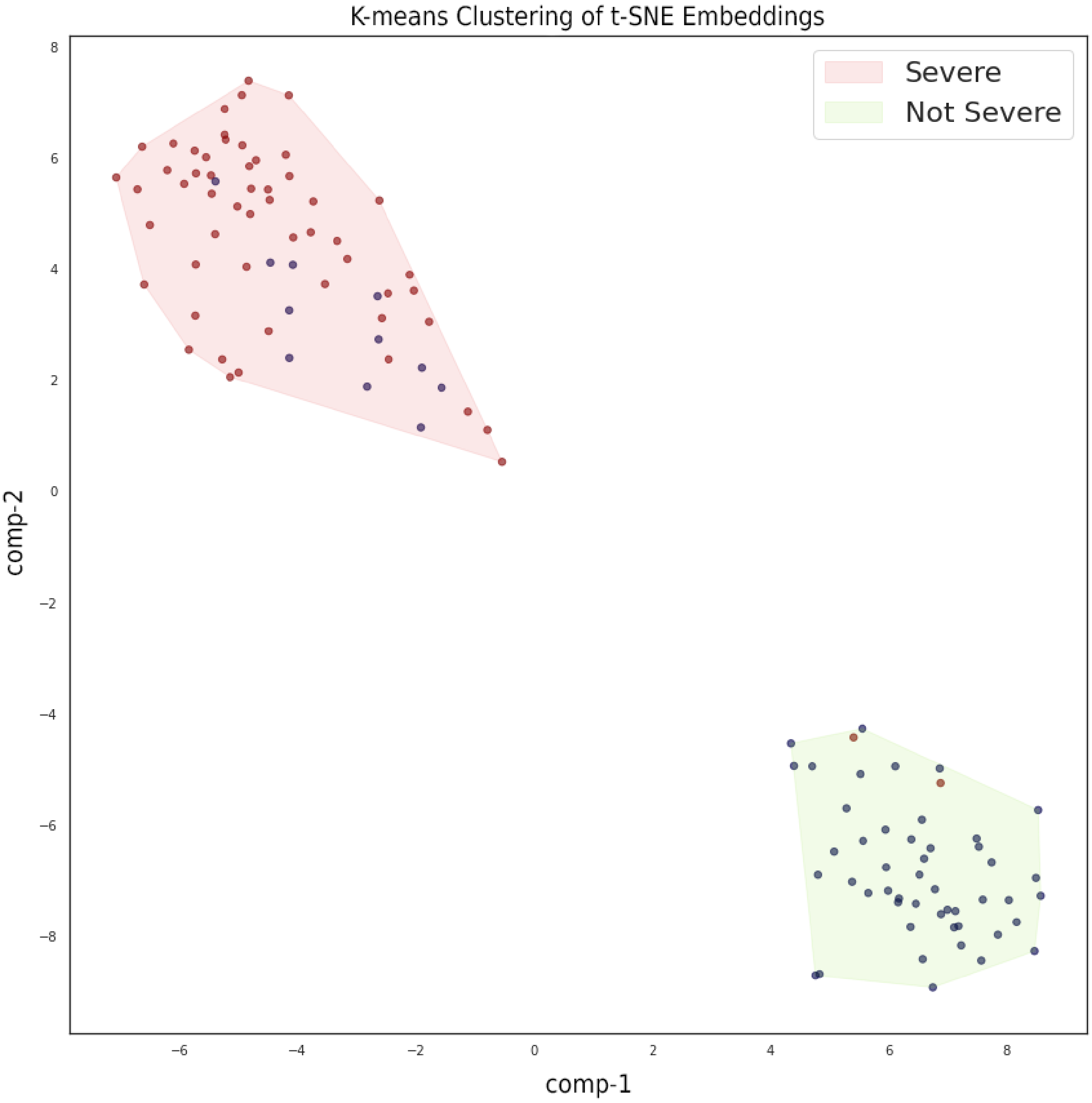


Figure 2. Visualizing the output of K-means clustering on test set t-SNEs. Note: The comp-1 and comp-2 axes denote the names of the 2 dimensions onto which t-SNE projects the 768D embeddings, where “comp” is short for “component”. t-SNE: t-distributed stochastic neighbor embedding.



Textbox 1. False-positive samples from the patient query test set.

1. "I came into close contact with someone who just flown back from Australia, I have been in self isolation since he landed, am I'm not showing any symptoms (dry cough, fever, sore throat) what is the next protocol? Do I go into work?"
2. "I have been recently diagnosed w the flu (nose swab test done). I am 34 years old- ex-smoker (quit completely for 7 years now) however each year I get 'walking pneumonia' at least once. I'm almost certain I've got it again now. Anything I can do to stop getting pneumonia??"
3. "Hi! So I'm a 20 year old female. I started working out about a year ago. I noticed some lower abdominal pain after partaking in abdominal workouts. But also notice it around the time of my period. It's right next to/under my hip bone on left side."
4. "I recently got over either the flu or pnemonia. I've noticed my feces are increasingly yellow or whitish. I had quit smoking thrity days ago and have been on the nicotine gum and now the lozenges. I don't feel bad, but this is unusual. Could the nicotine products be contributing to the discoloration? Thank you."
5. "Last year during flu season had a severe cough, difficulty breathing and xray show there was fluid in my chest/lungs. Any advice on what to do with this covid 19."
6. "Hi. What medication can I take for sinus and headaches during this time of the virus? Thank you."
7. "I was recently diagnosed with Pneumonia. I had pneumonia when I was a baby but never since. I was shocked when I heard the diagnosis. The symptoms became apparent on March 28. I began taking Doxycycline Monohydrate on Monday April 3 and must take them for seven days. I am mega healthy and have not been sick in years. I can t even remember the last time I was sick. This has really knocked me out. I have zero energy and very little appetite. How does pneumonia manifest in the lung and how does bacteria get in there? How long will it be until I recover? I am really having a hard time with this. Help!"
8. "Hi,my sinuses usually act up during seasonal change (like now). My worry is my symptoms resemble that of covid-19. Throat has been a bit irritated, lately nose has been runny. Wanted to know how I could get myself tested as I don't live alone, thanks?"
9. "Throat a bit sore and want to get a good imune booster, especially in light of the virus. Please advise. Have not been in contact with Nyone with the virus."
10. "Hello, I am a student and dealing with a microbiology assignment.I am given a paitent sample.My paitent is 4 years old -diagnosis Pneumonia - summary of peresent illness =Recurrant colds, ear infections,and bronchitis.She has been sick for past 3 weeks. Developed a fever yesterday.Also nausia and vomiting,muscle aches. Past Medical history= Cystic Fibrosis diagnosed at age 3. I did all the lab work and found out that the bacteria causes the disease is Psudomonas aeruginosa. What is the appropriate treatment? Please help."
11. "Hi. This COVID-19 outbreak is scary. I got screened this week and it was negative. But prior to screening I had a week of continuous dry coughs and also throat was sore. I've put myself in a quarantine. What next? Do I still need to screen again?"

Textbox 2. False-negative samples from the patient query test set.

1. "My dr. Did a routine CNBC last week. His nurse called and my blood showed signs of anemia. Ok today his nurse called and the blood they deeper searched on showed: the disease of chronic pneumonia? Ok I do not have hiv/AIDS. My question is she said no cure. My mom died from a chronic infection in her longs acronym Boop. I contacted cdc.gov they put me in touch with center for rare diseases. No cure for mom. Question: the told me Boop was not genetically transferred. This is exactly how my moms lung diseases started. is thimy lung disease genetic? Is it curable? Help please."
2. "Good morning I have all the symptoms for the coronavirus except a high fewer. I have been in contact with someone (who now also are displaying these symptoms) who are staying with a person who visited India in the last few weeks. Should I be worried?"

Discussion

Principal Results

In this study, we provided a novel extension of the COVID-Dialogue data set with telemedical query severity labels. Further, we thoroughly investigated the capability of several transfer learning approaches to predict severity in a resource-constrained setting. We concluded that transformer-based models are able to triage with high efficiency (all F1 scores>0.9). Further, we provided a thorough error analysis of SBERT, highlighting challenging samples that require a deep understanding of symptom presentation. Our error analysis highlights various avenues for future work that explicitly model various patient query types.

This is a new area of research and requires more investigation to define the requirements of a real deployment. It should be noted that such systems should not be used for diagnosis. Such a solution can benefit from online learning approaches, especially in the context of the pandemic (eg, temporal and spatial factors are important for detecting outbreaks of a new variant of infection).

Interpretability: Performance Trade-off

A commonly discussed limitation of deep neural networks (DNNs) is their lack of a natural way to explain the predictions they have made [39,40], so the use of DNNs makes it difficult to ask *why* a certain sample was predicted as severe or not severe. Models such as long short-term memory (LSTM) and transformer-based models make it challenging to identify when unfair biases or spurious correlations drive predictions. Thus, the use of transfer learning in telemedical triage must be done with care, as biases learned from other data sets may influence triage decisions.

Lexical models, such as TF-IDF, in combination with a linear classifier, provide straightforward access to a model's utilization of certain vocabulary terms.

Given that our BERT-based models only provided up to a 5-point increase in the F1 score, we compared the test set errors of SBERT and TF-IDF+SVM to highlight specific sample types that require complexity and knowledge transferability of the transformer architecture for accurate prediction (Table 3).

Across all test sets (ie, the test set from each fold in 5-fold cross-validation), we found that the lexical model using TF-IDF made 77 errors, while SBERT made 47 errors. Additionally, 81% of the errors made by SBERT were also made by TF-IDF. Of the 39 samples predicted correctly by SBERT but incorrectly by TF-IDF, we highlighted 7 queries in Table 3 that are representative of the TF-IDF errors. A qualitative analysis of Table 3 highlighted the following:

- General queries: These samples either inquire about general medical knowledge or request information about COVID-19 testing from a nonsymptomatic patient. Samples 1 and 2 highlight an example of each general query type. These samples were challenging for TF-IDF, given lexical models may struggle to understand query intent without contextual modeling, as our TF-IDF model only considers unigram features. From our 39-sample evaluation set, 22 predictions made by TF-IDF were false positives, with 59% of them being on general query samples.
- Ambiguous questions: These samples are queries that do not contain enough information for a valid response or do not pose a question that can benefit from a remote physician. Samples 3 and 4 are examples of ambiguous questions. We found that 15% of the TF-IDF errors that SBERT predicted correctly are predictions on ambiguous questions.

Many of the false negatives had no obvious content-based justification for TF-IDF's failure. In other words, TF-IDF's issue with the false negatives in Table 3 appear to be due to symptom sparsity and spelling errors. The ability of transformer-based models to transfer knowledge and analyze subwords makes it better suited for such samples, which are realistic issues to be faced by any telemedical triage system put into production.

Table 3. Subset of samples predicted incorrectly by TF-IDF[a]+SVM[b] but predicted correctly by SBERT[c].

Sample number	Patient query	Ground truth label
1	"About the ibuprofen and covid 19 should I quit taking it? It's got me paranoid. The way the media's been talking about it. I take it everyday for my neck pain and back pain. I can't take pain pills because they make me nauseas. Any insight please"	0
2	"Hi, I arrived from the Netherlands on Monday morning. No symptoms but have been around my helper. Should we get tested"	0
3	"I'm finding difficult to maintain precisely 6 ft in grocery stores. Today, as I was leaving, someone entering the store that was (possibly) 3 ft away was coughing lightly, and I took a shower when I got home. I'm a hypochondriac. Possible covid-19?"	0
4	"Hi, My uncle has been diagnosed with liver cancer and he is in the last stage. After the first chemotherapy he has been admitted to hospital due to pneumonia. Is he again diagnosed with lung cancer? And what are the chances of getting cure? What treatment you would like us to get it done."	0
5	"I believe I might have Covid 19 symtoms. It's possible to get testing done at home to confirm? Currently I have soar throat, started last night around 19:30."	1
6	"Hi my husband has been puking since this morning, has serious vertigo + is off balance. Im suspecting food poisoning but want to be sure. I gave him a pill for nausea, which is working. Do I still take him to a doctor to check that its nothing else?"	1
7	"I live in france.and now 7days for home quarantine.i have no fever.but I have parangities in my thoart. last few years it's comes and goes. now I am worried because of covid-19. Does only parangities is only symptoms of this???"	1

[a] TF-IDF: term frequency–inverse document frequency.

[b] SVM: support vector machine.

[c] SBERT: sentence bidirectional encoder representation from transformers.

Limitations and Future Work

We were unfortunately only able to look at telemedical triage through the lens of patients with COVID-19 or related pneumonia symptoms. Real-world systems will need to understand a diverse set of diseases and symptoms to handle the variance in queries doctors will receive. For example, HealthTap offers medical advice in over 147 specialties, necessitating a system with a deeper understanding of different medical conditions. In the future, we plan to extend our system so that it can classify patient queries that span a more diverse set of medical conditions.

Like any other automatic recommendation system, the performance of such an automated triage system might be affected based on the quality of user queries. For instance, an automatic triage system can assign lower severity to queries that have missing information (ie, a patient forgets to mention relevant symptoms or does not share enough details) or are not well written. This is similar to Google Search, where the quality of search results depends on the user query and where user satisfaction is correlated with the quality of the query. Such automatic triage can still be useful and significantly improve longitudinal user interaction at scale, as has been shown in other recommendation systems (eg, Google Search or Amazon recommendations). Another limitation is the binary classification system, as it ignores the spectrum of potential perceived severity. Future work could develop a score-based system where severity is scored on a continuous scale. Further research with multidisciplinary research teams is required to determine the impact of such automated solutions and identify potential techniques to address such limitations.

In future work, we will extend this system beyond the online medical Q&A forum and into doctor-patient messaging apps. We are actively in conversation with a local teaching hospital regarding the issue of doctors being overwhelmed with textual medical queries. Thus the NLP models explored in this paper may prove useful in future works modeling doctor-patient messaging data, including tasks such as the relative ranking of patient query importance of in-hospital private messaging systems. However, the triage problem becomes more challenging in the hospital messaging system context as patients will naturally assume their doctor is familiar with their medical history, providing narrow and incomplete information in their textual queries. To handle this problem, future systems must be able to make inferences using multiple modalities (eg, EHRs, medical images), as well as past conversations, which will likely require a shift in deep learning architecture as BERT-based models are restricted to processing of 512 tokens.

Conclusion

Telemedical triage is an important task in the world of telemedicine. Ranking medical queries according to severity both optimizes the doctor’s time and allows care to be administered to more of those with time-sensitive issues. We showed that even in the presence of a limited amount of data, transfer learning can be used to triage for COVID-19 and pneumonia patients with high accuracy. Specifically, we found a statistically significant difference in performance between transformer-based solutions and both lexical and GloVe embedding-based solutions. We additionally categorized all model errors into numerous interpretable categories, highlighting sample types that challenge our NLP-based triage systems. Queries with complex negation, temporality, and ambiguity (among other linguistic phenomena) were shown to be highly present in SBERT’s errors, giving specific direction for future work on telemedical triage.

Data Availability

Data and code from this study can be found on GitHub.

Conflicts of Interest

None declared.

References

1. Johnson C, Taff K, Lee BR, Montalbano A. The rapid increase in telemedicine visits during COVID-19. Patient Exp J 2020 Aug 04;7(2):72-79. [doi: 10.35680/2372-0247.1475]
2. Nanda M, Sharma R. A review of patient satisfaction and experience with telemedicine: a virtual solution during and beyond covid-19 pandemic. Telemed J E Health 2021 Dec 01;27(12):1325-1331. [doi: 10.1089/tmj.2020.0570] [Medline: 33719577]
3. Gmunder KN, Ruiz JW, Franceschi D, Suarez MM. Demographics associated with US healthcare disparities are exacerbated by the telemedicine surge during the COVID-19 pandemic. J Telemed Telecare 2021 Jun 23:1357633X2110259. [doi: 10.1177/1357633x211025939]
4. Healthtap Review. URL: https://www.telehealth.com/online-doctor/healthtap-review/ [accessed 2022-02-23]
5. Margolius D, Hennekes M, Yao J, Einstadter D, Gunzler D, Chehade N, et al. On the front (phone) lines: results of a covid-19 hotline. J Am Board Fam Med 2021 Feb 23;34(Supplement):S95-S102. [doi: 10.3122/jabfm.2021.s1.200237]
6. Dorn S. Backslide or forward progress? Virtual care at U.S. healthcare systems beyond the COVID-19 pandemic. NPJ Digit Med 2021 Jan 08;4(1):6 [FREE Full text] [doi: 10.1038/s41746-020-00379-z] [Medline: 33420420]
7. PersistLab/TelemedicalQueryClassification. URL: https://github.com/Persist-Lab/TelemedicalQueryClassification [accessed 2022-03-04]
8. Ju Z, Chakravorty S, He X, Chen S, Yang X, Xie P. UCSD-AI4H/COVID-Dialogue. URL: https://github.com/UCSD-AI4H/COVID-Dialogue [accessed 2022-08-23]
9. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. arXiv Preprint published on Aug 27, 2019. [FREE Full text] [doi: 10.18653/v1/d19-1410]
10. Vaswani A, Shazeer N, Parmar N. Attention is all you need. CoRR. arxiv Preprint published on Jun 12, 2017. [FREE Full text]
11. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR. arxiv Preprint published on Oct 11, 2018. [FREE Full text]
12. Romanov A, Shivade C. Lessons from natural language inference in the clinical domain. CoRR. arxiv Preprint published on Aug 21, 2018. [FREE Full text] [doi: 10.18653/v1/d18-1187]

13. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011 Sep 01;18(5):552-556 [FREE Full text] [doi: 10.1136/amiajnl-2011-000203] [Medline: 21685143]
14. Alsentzer E, Murphy J, Boag W. Publicly available clinical BERT embeddings. 2019 Presented at: 2nd Clinical Natural Language Processing Workshop; June 7, 2019; Minneapolis. [doi: 10.18653/v1/w19-1909]
15. Lai L, Wittbold KA, Dadabhoy FZ, Sato R, Landman AB, Schwamm LH, et al. Digital triage: novel strategies for population health management in response to the COVID-19 pandemic. Healthc (Amst) 2020 Dec;8(4):100493 [FREE Full text] [doi: 10.1016/j.hjdsi.2020.100493] [Medline: 33129176]
16. Yao L, Leung K, Tsai C, Huang C, Fu L. A novel deep learning-based system for triage in the emergency department using electronic medical records: retrospective cohort study. J Med Internet Res 2021 Dec 27;23(12):e27008 [FREE Full text] [doi: 10.2196/27008] [Medline: 34958305]
17. Gligorijevic D, Stojanovic J, Satz W. Deep attention model for triage of emergency department patients. CoRR. arxiv Preprint published on Mar 28, 2018. [FREE Full text] [doi: 10.1137/1.9781611975321.34]
18. Nicki GEA. Emergency Severity Index. Version 4: Implementation Handbook. 2012 Edition. Rockville, MD: Agency for Healthcare Research and Quality; 2012.
19. Si S, Wang R, Wosik J, Zhang H, Dov D, Wang G, et al. Students need more attention: BERT-based attention model for small data with application to automatic patient message triage. arXiv Preprint published on Jun 22, 2020. [FREE Full text]
20. Fu G, Song C, Li J, Ma Y, Chen P, Wang R, et al. Distant supervision for mental health management in social media: suicide risk classification system development study. J Med Internet Res 2021 Aug 26;23(8):e26119 [FREE Full text] [doi: 10.2196/26119] [Medline: 34435964]
21. Wang X, Chen S, Li T, Li W, Zhou Y, Zheng J, et al. Depression risk prediction for Chinese microblogs via deep-learning methods: content analysis. JMIR Med Inform 2020 Jul 29;8(7):e17958 [FREE Full text] [doi: 10.2196/17958] [Medline: 32723719]
22. Klein AZ, Magge A, O'Connor K, Flores Amaro JI, Weissenbacher D, Gonzalez Hernandez G. Toward using Twitter for tracking covid-19: a natural language processing pipeline and exploratory data set. J Med Internet Res 2021 Jan 22;23(1):e25314 [FREE Full text] [doi: 10.2196/25314] [Medline: 33449904]
23. Dahl D. How I Disrupted the Health Care Industry. URL: https://www.inc.com/magazine/201311/darren-dahl/how-healthtap-disrupted-the-health-care-industry.html [accessed 2022-02-23]
24. Tang L, Fujimoto K, Amith MT, Cunningham R, Costantini RA, York F, et al. "Down the rabbit hole" of vaccine misinformation on YouTube: network exposure study. J Med Internet Res 2021 Jan 05;23(1):e23262 [FREE Full text] [doi: 10.2196/23262] [Medline: 33399543]
25. Hansen ND, Mølbak K, Cox IJ, Lioma C. Relationship between media coverage and measles-mumps-rubella (MMR) vaccination uptake in Denmark: retrospective study. JMIR Public Health Surveill 2019 Jan 23;5(1):e9544 [FREE Full text] [doi: 10.2196/publichealth.9544] [Medline: 30672743]
26. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
27. Johnson A, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
28. Monselise M, Chang C, Ferreira G, Yang R, Yang CC. Topics and sentiments of public concerns regarding covid-19 vaccines: social media trend analysis. J Med Internet Res 2021 Oct 21;23(10):e30765 [FREE Full text] [doi: 10.2196/30765] [Medline: 34581682]
29. Hossain T, Logan RI, Ugarte A, Matsubara Y, Young S, Singh S. COVIDLies: detecting covid-19 misinformation on social media. 2020 Presented at: 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020; December 2020; Online. [doi: 10.18653/v1/2020.nlpcovid19-2.11]
30. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. JMLR 2011;12(85):2825-2830 [FREE Full text]
31. Manning C, Raghavan P, Schütze H. Introduction to information retrieval. Nat Lang Eng 2010;16(1):100-103. [doi: 10.1017/cbo9780511809071]
32. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297. [doi: 10.1007/bf00994018]
33. Pennington J, Socher R, Manning C. GloVe: global vectors for word representation. 2014 Presented at: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October 2014; Doha, Qatar. [doi: 10.3115/v1/d14-1162]
34. Mascio A, Kraljevic Z, Bean D. Comparative analysis of text classification approaches in electronic health records. CoRR. arxiv Preprint published on May 8, 2020. [FREE Full text] [doi: 10.18653/v1/2020.bionlp-1.9]
35. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. 2016 Presented at: 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 12-17, 2016; San Diego, CA. [doi: 10.18653/v1/n16-1174]

36. Liu Z, He H, Yan S, Wang Y, Yang T, Li G. End-to-end models to imitate traditional Chinese medicine syndrome differentiation in lung cancer diagnosis: model development and validation. JMIR Med Inform 2020 Jun 16;8(6):e17821 [FREE Full text] [doi: 10.2196/17821] [Medline: 32543445]
37. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947 Jun;12(2):153-157. [doi: 10.1007/BF02295996] [Medline: 20254758]
38. van der Maaten L, Hinton G. Visualizing data using t-SNE. JMLR 2008;9:2579-2605 [FREE Full text]
39. Zhang Y, Tino P, Leonardis A, Tang K. A survey on neural network interpretability. IEEE Trans Emerg Top Comput Intell 2021 Oct;5(5):726-742. [doi: 10.1109/tetci.2021.3100641]
40. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019 May 13;1(5):206-215 [FREE Full text] [doi: 10.1038/s42256-019-0048-x] [Medline: 35603010]

Abbreviations

AI: artificial intelligence
BERT: bidirectional encoder representation from transformers
bi-LSTM: bidirectional long short-term memory
DNN: deep neural network
ED: emergency department
EHR: electronic health record
GloVe: global vectors for text representation
HAN: hierarchical attention network
KNN: K-nearest neighbor
LSTM: long short-term memory
NLP: natural language processing
Q&A: question-and-answer
RQ: research question
SBERT: sentence bidirectional encoder representation from transformers
SVM: support vector machine
TF-IDF: term frequency–inverse document frequency
t-SNE: t-distributed stochastic neighbor embedding

Edited by T Hao; submitted 08.03.22; peer-reviewed by L Guo; comments to author 28.05.22; revised version received 17.06.22; accepted 11.08.22; published 02.09.22.

Please cite as:
Gatto J, Seegmiller P, Johnston G, Preum SM
Identifying the Perceived Severity of Patient-Generated Telemedical Queries Regarding COVID: Developing and Evaluating a Transfer Learning–Based Solution
JMIR Med Inform 2022;10(9):e37770
URL: https://medinform.jmir.org/2022/9/e37770
doi:10.2196/37770
PMID:35981230

Original Paper

Leveraging Representation Learning for the Construction and Application of a Knowledge Graph for Traditional Chinese Medicine: Framework Development Study

Heng Weng[1], MD; Jielong Chen[2], MD; Aihua Ou[1], BA; Yingrong Lao[1], PhD

[1]State Key Laboratory of Dampness Syndrome of Chinese Medicine, Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, China

[2]School of Information Science, Guangdong University of Finance & Economics, Guangzhou, China

Corresponding Author:
Yingrong Lao, PhD
State Key Laboratory of Dampness Syndrome of Chinese Medicine
Second Affiliated Hospital of Guangzhou University of Chinese Medicine
Dade road No. 111
Guangzhou, 510120
China
Phone: 86 81887233 ext 35933
Email: laoyingrong@qq.com

Abstract

Background: Knowledge discovery from treatment data records from Chinese physicians is a dramatic challenge in the application of artificial intelligence (AI) models to the research of traditional Chinese medicine (TCM).

Objective: This paper aims to construct a TCM knowledge graph (KG) from Chinese physicians and apply it to the decision-making related to diagnosis and treatment in TCM.

Methods: A new framework leveraging a representation learning method for TCM KG construction and application was designed. A transformer-based Contextualized Knowledge Graph Embedding (CoKE) model was applied to KG representation learning and knowledge distillation. Automatic identification and expansion of multihop relations were integrated with the CoKE model as a pipeline. Based on the framework, a TCM KG containing 59,882 entities (eg, diseases, symptoms, examinations, drugs), 17 relations, and 604,700 triples was constructed. The framework was validated through a link predication task.

Results: Experiments showed that the framework outperforms a set of baseline models in the link prediction task using the standard metrics mean reciprocal rank (MRR) and Hits@N. The knowledge graph embedding (KGE) multitagged TCM discriminative diagnosis metrics also indicated the improvement of our framework compared with the baseline models.

Conclusions: Experiments showed that the clinical KG representation learning and application framework is effective for knowledge discovery and decision-making assistance in diagnosis and treatment. Our framework shows superiority of application prospects in tasks such as KG-fused multimodal information diagnosis, KGE-based text classification, and knowledge inference–based medical question answering.

(JMIR Med Inform 2022;10(9):e38414) doi:10.2196/38414

KEYWORDS

knowledge graph; knowledge embedding; traditional Chinese medicine; knowledge discovery; medicine; clinical; framework

Introduction

Background

Having a long history of 5000 years, traditional Chinese medicine (TCM) is featured as the scientific thinking of holistic view and syndrome differentiation, as well as the long-time practice of technical methods of personalized treatment. TCM has the advantages of precise clinical efficacy, relatively safe medication, flexible treatment, and relatively low cost [1]. However, a large amount of empirical knowledge exists with Chinese physicians, which is difficult to be applied directly in assisting clinical decision-making systems. At the same time, the dismantling of medical guidelines alone cannot cope with all situations, and existing clinical assisted decision-making systems cannot explain the ins and outs of diagnostic decisions as senior experts do.

The combination of knowledge graphs (KGs) and artificial intelligence (AI) models has the bilateral advantages of "black box" and "logic." Using knowledge graph embedding (KGE) techniques, KGE models may partially simulate the cognitive process of the human brain by representing massive entities, relations, and attributes. By combining with the causal events extracted from the text of event descriptions by causality extraction techniques, event information can be presented in structured form. KGs and machine learning models are expected to be integrated to assist machine understanding and concept interpretation, allowing the decision-making process of machines to be interpretable. However, how to construct a TCM KG and apply it with KGE models is still a challengeable problem.

To that end, this paper proposes a new framework leveraging a representation learning method for TCM KG construction and application. TCM knowledge is extracted from Chinese physicians based on 1 of our previous works [2] by using an automatic procedure of information extraction concept normalization, *entity alignment*. The framework collects multimodal information about Chinese medicines to support the automatic construction of personalized KGs according to clinical disease treatments by Chinese physicians. Our framework has application potential in text classification, KG-based question answering, and recommendations of practitioners and specialties.

The main contributions of this paper are threefold: (1) A new framework for the construction and application of TCM KG by leveraging representation learning is proposed, (2) a transformer-based Contextualized Knowledge Graph Embedding (CoKE) model is applied to KG representation learning and knowledge distillation by integrating multihop relations, and (3) a TCM KG containing 59,882 entities, 17 relations, and 604,700 triples is constructed.

Related Work

Medical Knowledge Graph

The concept of KG was proposed by Google in 2012. Research applications evolved by previously improving the capabilities of search engines and enhancing the search quality and experience of users related to finance, healthcare, geography, e-commerce, and medical care. There exist many KGs, including on Google Knowledge Graph [3], DBpedia [4], Yet Another Great Ontology (YAGO; Max Planck Institute for Computer Science) [5], and FreeBase (Metaweb Technologies, Inc.) [6]. In China, there are Zhi Cube (Sogou), Zhi Xin (Baidu), zhishi.me (Shanghai Jiao Tong University) [7], and the GDM Lab Chinese KG project (Fudan University) [8]. In the medical field, the KG of medicine NKIMed [9] was developed by the Institute of Computer Technology of the Chinese Academy of Sciences, and the KG of Chinese medicine [10] was constructed by the Institute of Chinese Medicine Information of the Chinese Academy of Traditional Chinese Medicine. The Traditional Chinese Medicine Language System (TCMLS) is a relatively large semantic network for the KG of Chinese medicine [11], containing more than 100,000 concepts and 1 million semantic relations, which basically covers the conceptual system of TCM disciplines. The TCMLS was in the leading position of the TCM community in terms of its scale and completeness. Rotmensch et al [12] extracted positive mentions of diseases and symptoms (concepts) from structured and unstructured data in electronic medical records (EMRs) and used them to construct a health KG automatically.

Knowledge Graph Representation Learning

Graph neural networks (GNNs) are deep learning architectures for graph-structured data, which combine end-to-end learning with inductive reasoning. GNNs are promising research topics of AI, and they are expected to solve the problems of causal inference and interpretability that cannot be handled by traditional deep learning models. KG representation learning is a critical branch of the research on GNNs and plays a nontrivial role in knowledge acquisition and downstream application. KG representation learning consists of elements such as representation spaces (pointwise space, complex vector space, gaussian distribution, manifold, and group), scoring functions (distance-based and semantic-matching scoring functions), and encoding models (linear/bilinear, factorization models and neural networks).

Translational models leverage translational distances (eg, L1 or L2 norm) to model relations between head and tail entities. TransE is one of the representative translational models [13]. Dealing with 1-to-N, N-to-1, and N-to-N relations, TransE suffered from inefficiency problems in representing head or tail entities. To alleviate such problems, KGE models, including TransH [14], TransR [15], and TransD [16], were designed to impose translational distance constraints through different entity projection strategies. RotatE considers the embedding vectors of relations as rotations from source entities to target entities in a complex space [17].

The basic idea of factorization models is to decompose the matrix of each slice in a 3-way tensor into a product of entity vectors and relation matrices in the lower-dimensional space. The RESCAL model leveraged a relation-associated matrix to capture interactions between head and tail entities, which required a large number of parameters to model relations [18]. Therefore, vector forms of relations were introduced in DistMult [19] to decrease model parameters by restricting the interaction matrices to diagonal matrices. To increase the interactions between head and tail entities, a circular correlation operation was leveraged as the score function in the expressive HolE model [20]. Inspired by DistMult, the ComplEx model extended the representations of entities and relations by utilizing embedding vectors in a complex space [21]. An expressive KGE model named SimplE used 2 vectors for each entity to learn independent parameters through simplifying ComplEx by removing redundant computation [22].

In recent years, inspired by convolution operations, convolution-based KGE models, such as ConvE [23], ConvKB [24], and CapsE [25], were designed as different strategies to capture features between entities and relations for KG representation learning. A KGE model named knowledge base attention (KBAT) extended the graph attention (GAT) network by exploring the multihop representation of a given entity for representation aggregation via multihead attention and graph attention mechanisms [26]. The natural language pretraining model BERT [27] learned to integrate contextual information in the KG based on the representation of the transformer [28]. CoKE [29] used a transformer to encode edge and path sequences. These promising methods have attracted much attention due to the high efficiency of convolution in representation learning. CoKE aimed to learn the dynamic adaptive representations of entities and relations based on a rich graph structure context. Compared with static representations, the performance of contextual models is state of the art, since the representations combined with contextual semantic information are richer and more flexible. Despite the use of a transformer, CoKE was still parameter-efficient to obtain competitive performance with fewer parameters. The comparison of the KG representation learning models is shown in Table 1.

Table 1. Comparison of baseline KGE[a] models.

Model	Scoring function $f_r(h,t)$	Entity and relation embedding
Translational model		
TransE [13]	☒	☒
TransH [14]	☒	☒
TransR [15]	☒	☒
TransD [16]	☒	☒
Linear/bilinear model		
SimplE [22]	☒	☒
HolE [20]	☒	☒
Rotational model		
QuatE [30]	☒	☒
RotatE [17]	☒	☒
Convolutional neural network		
ConvE [23]	☒	☒
ConvKB [24]	☒	☒
GNN[b]		
KBAT[c] [26]	☒	☒
Neural network transformer		
CoKE[d] [29]	☒	☒

[a] KGE: knowledge graph embedding.

[b] GNN: graph neural network.

[c] KBAT: knowledge base attention.

[d] CoKE: Contextualized Knowledge Graph Embedding.

Application of Medical Knowledge Graphs

The hot topics related to the application of medical KGs are KG-fused multimodal information diagnosis, KGE-based text classification, and knowledge inference–based medical question answering and assisted diagnosis. Shen et al [31] reused the existing knowledge base to build a high-quality KG and designed a prediction model to explore pharmacology and KG features. The model allowed the user to gain a better understanding of the drug properties from a drug similarity perspective and insights that were not easily observed in individual drugs. Zheng et al [32] took advantage of 4 kinds of modality data (X-ray images, computed tomography [CT] images, ultrasound images, and text descriptions of diagnoses) to construct a KG. The model leveraged multimodal KG attention embedding for diagnosis of COVID-19. The experimental results demonstrated that it was essential to capture and join the importance of single- and multilevel modality information in a multimodal KG. Li et al [33] designed an AI-powered voice assistant by constructing a comprehensive knowledge base with ontologies of defined Alzheimer disease and related dementia (ADRD) diet care and user profiles. They extended the model with external KGs, such as FoodData Central and DrugBank, which personalized ADRD diet services provided through a semantics-based KG search and reasoning engine.

With the development of deep learning methods, diagnostic decisions have become interpretable. Theoretically, rule-based engines may infinitely approximate the performance of nonlinear classifiers by mining the expanded knowledge. In other words, through the integration of interpretable knowledge rules, rule-based engines may approximate the performance of deep learning models. Through deep mining of rules, the clinical assisted decision-making system may be able to perform multiple rounds of rule expansion under dynamic thresholds and further extend the capability of decision-making based on existing knowledge.

Methods

TCM Knowledge Graphs

To construct a TCM KG (Table 2) for ordinary usage, such as disease diagnosis and treatment assistance, we cleaned the EMR data set of diagnosis and treatment of TCM diseases and represented the relations of entities in triples. For instance, given a description text of insulin resistance as a mechanism of type 2 diabetes, the entities and relations in the sentence were extracted and organized into a disease mechanism triple of (insulin resistance, mechanism=>disease, diabetes). A KG was defined as G=(E,R,S), where entities, relations, and triples are , respectively, and |E| and |R| are the counts of entities and relations. The triples consisted of entities, relations, describing concepts, or attributes.

Traditional KGE models are designed to learn static representations of entities and relations. The features of graph contexts are obtained by representing neighbor entities and relations. Different meanings are expressed by entities and relations in diverse contexts, as words appear in different textual contexts. Multihop relations (ie, paths between entities) can provide rich contextual features for reasoning in KG [29]. Existing work [34] shows that multihop relation paths contain rich inference patterns between entities. Since not all relation paths are reliable, we designed a causal-constraint algorithm to filter the reliability of relation paths. Relation paths were represented via semantic composition of relation embeddings. The screened multihop relations were extended to triple alternative combinations.

The rules for screening potential multihop causal relations are shown in Figure 1. For example, there exist triples (*insulin resistance*, *treat*, *diabetes mellitus*) and (*metformin*, *mechanism*, *insulin resistance*) in a clinical KG describing the relations between clinical mechanism and disease (or drug) as a positive example in the figure. The relations can be inferred as the causal multihop relation between a drug and a disease by the rules drug=>mechanism and mechanism=>disease, indicating that metformin can treat insulin-resistant diabetes. The triples (*dyslipidemia*, *symptom*, *diabetes mellitus*) and (*dyslipidemia*, *symptom*, *CKD* [where CKD refers to chronic kidney disease]) co-occurred and thus could not reflect the causal relation between diabetes mellitus and CKD or dyslipidemia. Such negative triples were screened according to the rules.

An example of a casual multihop relation of TCM disease (abdominal mass)=>mechanism (*phlegm dampness*, *toxin*, *blood stasis*)–mechanism=>mechanism (*clearing heat-toxin*, *eliminating dampness*)–disease=>drug (*root of Chinese* Pulsatilla) can be inferred according to the rules (*abdominal mass*, *disease=>drug*, *phlegm dampness*, *toxin*, *blood stasis*), (*phlegm dampness*, *toxin*, *blood stasis*, *mechanism=>mechanism*, *clearing heat-toxin*, *eliminating dampness*), and (*abdominal mass*, *disease=>drug*, *root of Chinese* Pulsatilla). In other words, casual multihop relations of TCM can be inferred, which conform to the cognition of diseases–syndrome–principle–method–recipe–medicines of TCM, including the aforementioned path disease=>mechanism=>treatment=>drug.

The semantics of the entities *diabetes mellitus* and *metformin* were enriched by the embeddings of the 2-hop path inferred by triples (*metformin*, *mechanism*, *insulin resistance*) and (*insulin resistance*, *treat*, *diabetes mellitus*). To represent multihop relations, given the 2-hop path from the entity *metformin* to *diabetesmellitus*, triple forms (*metformin*, *mechanism-treat*, *diabetes mellitus*) were used for consistency. Since the multihop features were integrated, the representations of entities and relations tended to have strong inference capability, which facilitated entity link prediction. The KG was represented as textual triples that described multihop relations of entities.

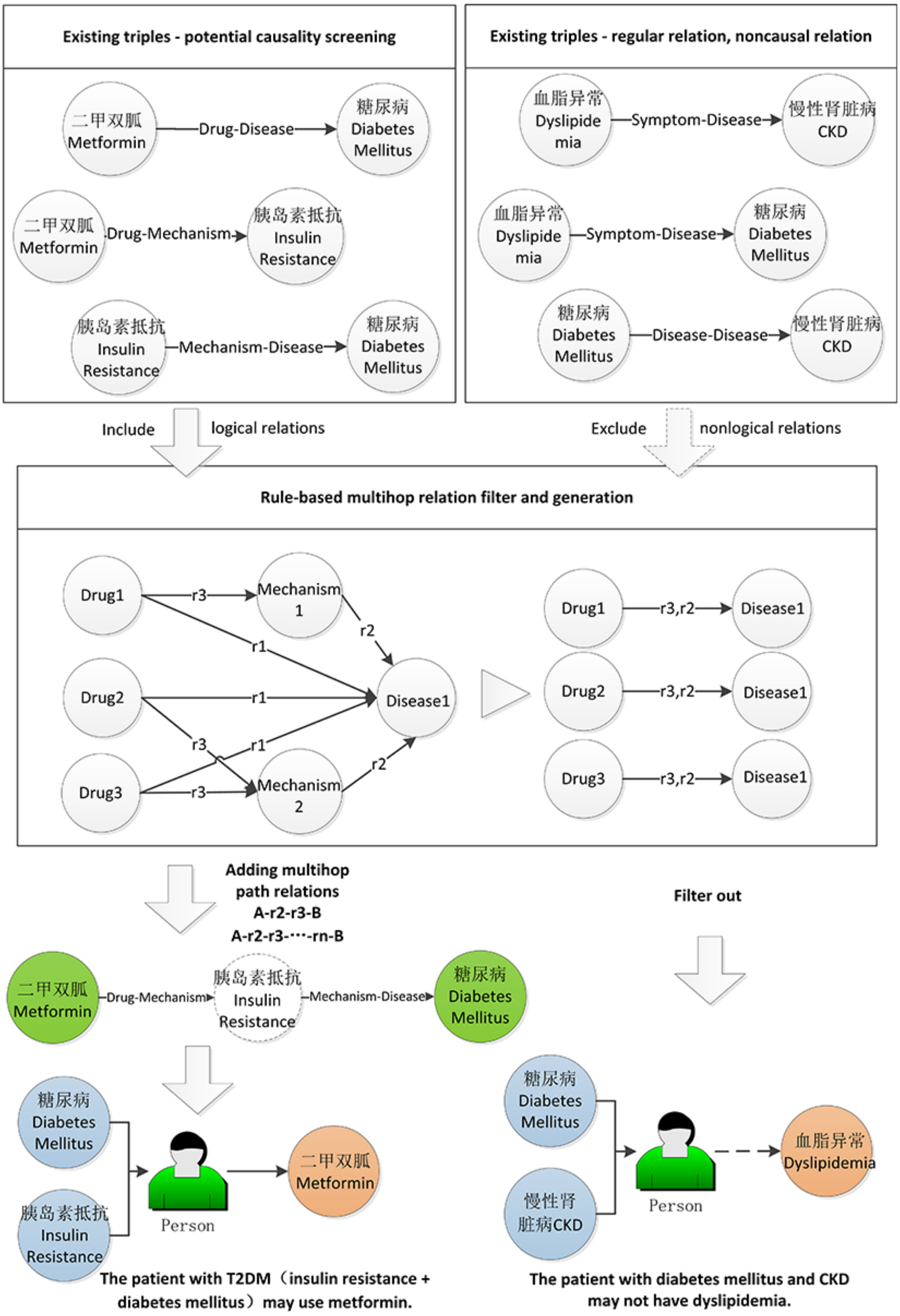
Table 2. Overview of the TCM[a] KG[b].

Relation name	Heads, n	Tails, n	Triples, n
symptom=>symptom	8101	8544	51,345
disease=>symptom	12,225	15,071	133,648
disease=>drug	12,650	11,526	84,524
mechanism=>mechanism	527	51	590
symptom=>drug	3941	6145	24,724
symptom=>mechanism	6544	1096	10,906
symptom=>disease	8101	10,391	87,651
mechanism=>department	1908	65	4408
symptom=>body parts	318	85	548
mechanism=>body parts	2217	72	3221
mechanism=>symptom	2147	4191	16,377
symptom=>department	10,157	178	24,870
disease=>mechanism	7774	5304	46,425
disease=>body parts	7607	110	13,505
disease=>department	14,484	284	40,762
disease=>disease	9728	10,545	40,575
mechanism =>disease	2228	5443	20,621

[a]TCM: traditional Chinese medicine.

[b]KG: knowledge graph.

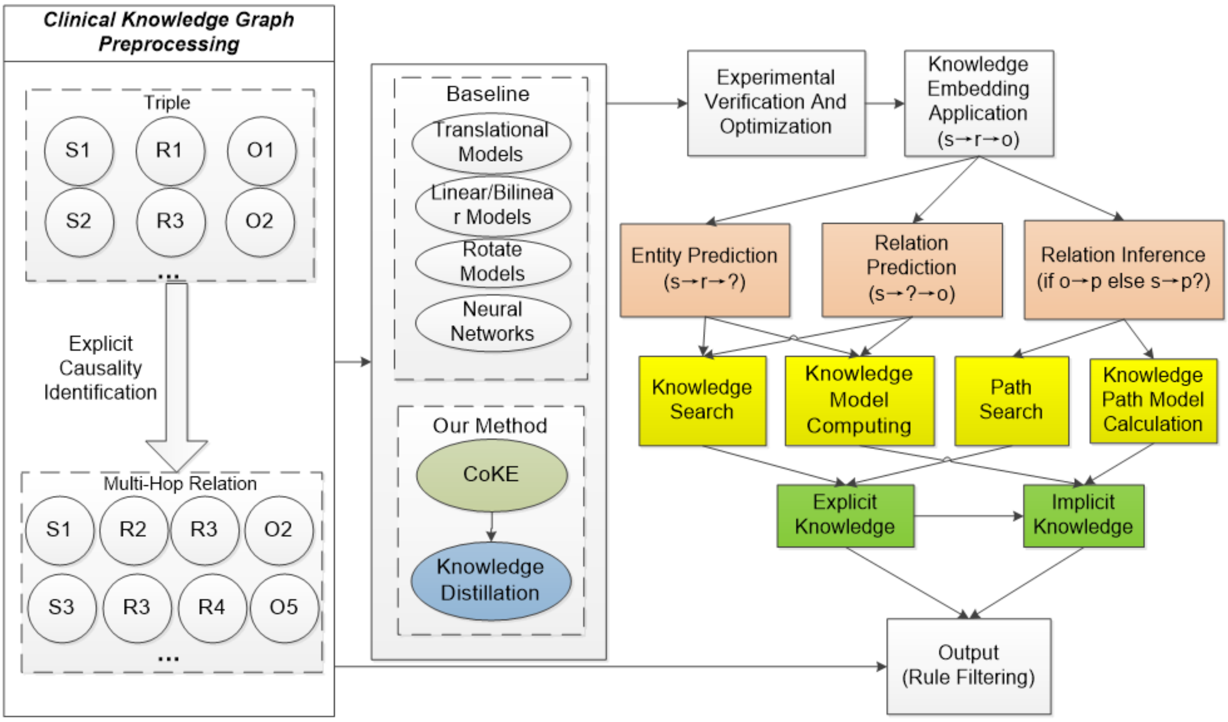
Figure 1. Positive and negative examples of multihop relation filtering and generation. CKD: chronic kidney disease; T2DM: type 2 diabetes mellitus.



Knowledge Graph Representation Framework

After preprocessing of the TCM KGs, we applied a CoKE-based KG representation learning model based on a diagnosis and treatment KG of Chinese and Western medicine and proposed a new KG representation framework. Compared with popular knowledge representation learning models, such as TransE and KBAT, our framework features the fusion of CoKE and multihop relations. The framework was verified with downstream applications, such as assisted decision-making and question answering, as shown in Figure 2.

Figure 2. Proposed framework of TCM KG representation learning. CoKE: Contextualized Knowledge Graph Embedding; KG: knowledge graph; TCM: traditional Chinese medicine.


Clinical Knowledge Graph Preprocessing
Triple
S1
R1
O1
S2
R3
O2
...
Explicit Causality Identification
Multi-Hop Relation
S1
R2
R3
O2
S3
R3
R4
O5
...
Baseline
Translational Models
Linear/Bilinear Models
Rotate Models
Neural Networks
Our Method
CoKE
Knowledge Distillation
Experimental Verification And Optimization
Knowledge Embedding Application (s→r→o)
Entity Prediction (s→r→?)
Relation Prediction (s→?→o)
Relation Inference (if o→p else s→p?)
Knowledge Search
Knowledge Model Computing
Path Search
Knowledge Path Model Calculation
Explicit Knowledge
Implicit Knowledge
Output (Rule Filtering)

Entity Link Prediction

The CoKE model was leveraged as the base model in this paper. The BERT model was leveraged to learn contextualized embeddings of entities and relations in CoKE. The input sequence X = (x_1, x_2, …, x_n) consisted of the embeddings of a head entity x_1 and a tail entity x_n, while the embeddings of relations were denoted as x_2 from x_{n-1}. Given $x_i$from the input sequence, the hidden representation h_i was expressed as Equation 1:

where is the embedding of an element and is the positional embedding of an element. The former was used to identify the current entities or relations in , and the latter presented the positional features of the element in the sequence. The constructed hidden representations were fed into transformer encoders of L layers as Equation 2:

where is the hidden representation of x_i at the l-th layer of the encoder. A multihead self-attention mechanism was leveraged by the transformer, which allowed each element to attend to other elements in the sequence effectively for contextual feature modeling. As the use of transformers has become ubiquitous recently, we omitted a detailed description of the transformer. The final hidden representations are representations for entities and relations within the sequence X. The learned representations were naturally contextualized and automatically adaptive to the input.

Multihop Relational Representation Learning

Given a triple (s,r,o) in a KG, the contexts between a head and a tail entity can be described as an edge and a path. An edge s→r→o is presented as a sequence that can be viewed as a triple. For instance, an edge *metformin→mechanism→insulin resistance* can form a triple (*metformin*, *mechanism*, *insulin resistance*) equivalently. As the basic unit of a KG, an edge (or a triple) is the simplest form of a graph context describing an entity. Another context is a path s→r_1→…→r_k→o as a sequence consisted of head and tail entities and a list of linked relations between them. For instance, the path describes multihop relations between the head entity *metformin* and the tail entity *diabetesmellitus*, where *insulin resistance* is the intermediate entity in the path, while *mechanism* and *treat* are the relations. The path can be expressed as a triple (*metformin*, *mechanism-treat*, *diabetes mellitus*). Consisting of contextual features of entities, the multihop path representation can be leveraged for reasoning in a KG.

To verify the effectiveness of the model, experiments of entity link prediction in knowledge graph completion (KGC) [35] and multihop relation representation learning were conducted. Entity link prediction refers to a task that predicts missing target entities of triples (h, r, ?) and (?, r, t) with a candidate entity set by semantic constraints of KGE models. PathQuery answering [36] was utilized in the experiments of multihop relation representation learning. Given a source entity s and a relation path p, a set of target entities that were inferred from the source entity s via the path p was predicted.

In entity link prediction, our model was trained to predict missing target entities, given a context in the KG, answering

1-hop or multihop queries. Different strategies were considered to train our model with respect to the cases of edges and paths. Each edge s→r→o is associated with 2 instances ?→r→o and s→r→?, which can be regarded as 1-hop query answering. For instance, *metformin→mechanism→?* is to answer the query, What is the mechanism of m*etformin*? Similarly, each path $s \to r_1 \to \ldots \to r_k \to o$ is also associated with 2 instances, one to predict s and the other to predict o, which can be viewed as multihop question answering. For instance, is to answer the query, What disease can be treated by the mechanism of *metformin*?

In the training procedure, edges or paths were unified as an input sequence $X = (x_1, x_2, \ldots, x_n)$. Two instances were created by replacing x_1 with a special token [MASK] for s prediction and by replacing x_n with [MASK] for o prediction. The masked sequence was fed into the transformer encoding blocks to obtain the final hidden representation for target entity prediction.

As in the BERT model, the representations of the masked entities were fed into a feedforward neural network and a standard Softmax layer was leveraged for classification (Equation 3):

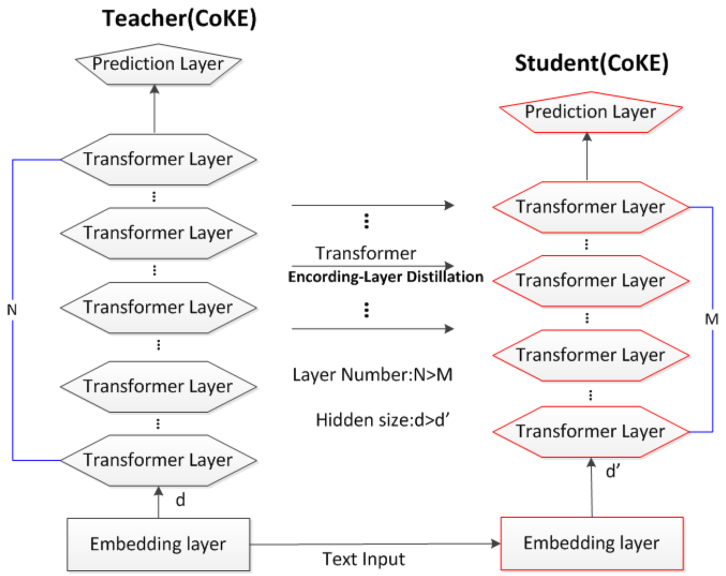
where z_1 and z_n are the representations of $h^L{}_1$ and $h^L{}_n$ produced by the feedforward layer, while is a matrix shared with the input element embedding matrix for classification. D is the hidden size, V is the size of the entity vocabulary, and p_1 and p_n are the predicted distributions of target entities s and o. Cross-entropy loss was leveraged as the loss function for classification (Equation 4):

where y_t and p_t are the t-th components of the 1-hot label vector y and the distribution vector p, respectively. A label-smoothing strategy was leveraged to lessen the restriction of 1-hot labels. In other words, the value of the target entity was set to ε, while $y_t = (1 – ε)/(V – 1)$ for incorrect entities in the candidate entity set.

Knowledge Distillation

Inspired by the idea of TinyBERT [37] for knowledge distillation, our model CoKE-distillation contains a teacher and a student model for knowledge distillation, as shown in Figure 3.

Figure 3. Architecture of CoKE-distillation. CoKE: Contextualized Knowledge Graph Embedding.



Our proposed CoKE-distillation model consists of 3 levels of distillation: embedding layer distillation, transformer -layer distillation, and prediction layer distillation. At the embedding layer distillation level, the embedding matrices of the student and teacher model are constrained by the mean-square error (MSE) loss (Equation 5):

where is a trainable linear transformation matrix to project the embedding of the student model into the semantic space of the teacher model. The embedding matrices of the student and teacher models are denoted by , where l is length of the sequence, d_0 is the size of the embeddings of the teacher model, and d is the size of the embeddings of the student model.

At the level of transformer layer distillation, the CoKE-distillation model distills knowledge in k-layer intervals. For instance, if the student model has 4 layers, a transformer loss is calculated every 3 layers, since the teacher model has 12 layers. The first layer of the student model corresponds to the third layer of the teacher model, while the second layer of the student model corresponds to the sixth layer of the teacher model and so on. The transformer loss of each layer is divided into 2 parts, attention-based knowledge distillation and implicit state–based knowledge distillation. The loss of each layer consists of an attention-based knowledge distillation loss and a hidden state-based knowledge distillation loss.

The attention-based knowledge distillation loss is expressed as Equation 6:

where h is the number of attention heads, refers to the attention matrix corresponding to the i-th head of the teacher or the student, and l is the length of the input text.

The hidden state-based knowledge distillation loss is expressed as Equation 7:

where the matrices refer to the hidden representations of student and teacher models, respectively. At the level of prediction layer distillation, prediction loss is shown as Equation 8:

where z^T and z^S are the logit vectors predicted by the student and the teacher respectively, CE means the cross-entropy loss, and t means the temperature value. In our experiment, t was set to .

Results

Data Set

To evaluate the proposed model, a widely used standard data set FB15k-237 [38] was used, which is a subset of the Freebase knowledge base [6] with 14,541 entities and 237 relations. Due to redundant relations existing in the FB15k data set, FB15K-237 removes the inverse relations, preventing models from directly inferring target entities by inverse relations. The FB15k-237 data set is randomly divided into 3 sets (training, validation, and test sets), with 272,115 triples in the training set, 17,535 triples in the validation set, and 20,466 triples in the test set.

We constructed a medical diagnosis and treatment data set of TCM, called TCMdt, consisting of entities and relations as triples. The data set contained 17 kinds of relations, 59,882 entities, and 604,700 triples without repetitive and inverse relations. There were 3811 kinds of N–1 relations, such as relation combinations *mechanism-body parts* and *mechanism-mechanism*. The rest of the relations were N–N relations, 600,868 in total. There were no 1–1 and 1–N relations in the data set. The data set was divided into a training, a validation and a test set, containing 59,882 entities and 17 relations in total. The details of the FB15k-237 and TCMdt data sets are shown as Table 3.

The hypertension data set (Table 4) in TCM for the multilabel modeling task was used in our experiment to evaluate the effectiveness of KGE learning. TCM has been used for the diagnosis of hypertension and has significant advantages. Symptom analysis and modeling of TCM provide a way for clinicians to accurately and efficiently diagnose hypertension. In this study, the initial data were collected from trained practitioners and clinical practitioners. Details of 928 cases of hypertension were collected from the clinical departments of the Guangdong Provincial Hospital, with both inpatient and outpatient medical records from the Liwan district [39]. All cases with incomplete information were removed from the data set, and the remaining 886 (95.47%) cases were used for analysis in this study.

Each case in the data set had 129 dimensions of TCM symptom features and syndrome diagnosis labels in 1-hot format. Each case had 2-5 labels of TCM syndrome diagnosis reidentified by trained clinicians. The KGE of the syndrome entities and the symptom vectors and matrix were constructed from the aforementioned TCMdt data set.

Table 3. Statistics of the FB15k-237 data set and the constructed TCMdt data set.

Data set	Entities, n	Relations, n	Triples in the training set, n	Triples in the validation set, n	Triples in the test set, n
FB15k-237	14,541	237	272,115	17,535	20,446
TCMdt	59,882	17	544,230	30,235	30,235

Table 4. Statistics of the hypertension data set in TCM[a].

Features, n	Classes, n	Total cases, N	Validation
121	8	886	10-fold cross-validation

[a]TCM: traditional Chinese medicine.

Baselines

Baseline methods were used for comparison in the experiments, including translational models, bilinear models, a rotational model, a GNN, and a transformer-based model. The details of the models and their types are shown in Table 5.

Table 5. Baseline methods for KG[a] representation learning.

Type of model	Models
Translational model	TransE [13], TranH [14], TransR [15], TransD [16]
Linear/bilinear model	ComplEx [21], DistMult [19], SimplE [22]
Rotational model	RotatE [17]
GNN[b]	KBAT[c] [26]
Transformer-based model	CoKE[d] [29]

[a] KGE: knowledge graph.
[b] GNN: graph neural network.
[c] KBAT: knowledge base attention.
[d] CoKE: Contextualized Knowledge Graph Embedding.

Evaluation Metrics

With respect to the evaluation metrics, Sun et al [40] found that some high performance can be attributed to the inappropriate evaluation protocols and proposed an evaluation protocol to address this problem. The proposed protocol was more robust to handle bias in the model, which could substantially affect the final results. Ruffinelli et al [41] conducted systematic experiments on the training methods used in various KGE models and found that some early models (eg, RESCAL) can outperform the state-of-the-art models, after adjusting the training strategies and exploring a larger search space of hyperparameters. This indicated that the performance improvement of the models might not reflect their advantage, since the training strategies might play a critical role. Therefore, we established a unified evaluation standard to mine the valuable ideas and superiority of the models.

We used the mean reciprocal rank (MRR) and Hits@N, which are frequently used evaluation metrics for link prediction task in KGs (Equations 9 and 10). Applying the filtered settings given by Wang et al [14], the rank of the head or tail entities in a test triple (e_i, r_k, e_j) was computed within a filtered entity set. The filtered entity set contained entities that could be used to generate valid triples without valid head or tail entities in the training set. A large value of the MRR indicates that the KGE model have the capability of precise entity representation, while Hits@N denotes a rate of head and tail entities that rank within N (1, 3, or 10) empirically.

In the equations, $|\Gamma_t|$ is the size of testing triple set Γ_t and $I(\cdot)$ is an indicator function, while denote values of ranks for a head and a tail entity e_i and e_j, respectively.

Model Performances

During the comparison, we evaluated the models with embedding vectors of 256, 512, 1024, and 2048 dimensions and sufficient iterations to ensure the obtained embeddings were qualified for the sake of the downstream task. The results are shown in Tables 6 and 7. Compared with the baseline models, the CoKE model showed a competitive performance on both the standard data set and the constructed TCMdt data set. The CoKE model had the highest MRR and CoKE-multihop model had the best Hits@10. The CoKE-multihop-distillation model still showed a competitive performance on the MRR and HIT@10 compared to the CoKE model.

To evaluate the effectiveness of the KGE learning, 10-fold cross-validation was used in the multilabel modeling task experiments. Compared with typical models multilabel k nearest neighbors (MLKNN), RandomForest-RAkEL (where RAkEL refers to random k-labelsets), LogisticRegression-RAkEL, and deep neural network (DNN) [42], the proposed model outperformed the baseline models on metrics precision, recall, and the F1 score, as shown in Table 8. In addition, multilabel models with KGE had better performance than those without KGE. The results demonstrate that learned KGE is capable of improving the performance of deep learning models.

As shown in Figure 4, the DNN+BILSTM-KGE (where BILSTM refers to bidirectional long short-term memory) outperformed the DNN on evaluation metrics (eg, precision and F1 score) in the training procedure. Compared with the DNN, the average precision and F1 score of DNN+BILSTM-KGE showed improvement, with the Hamming loss significantly decreasing for the first 50 iterations.

Table 6. Performance comparison of link prediction on the FB15k-237 data set.

Models	MRR[a]	Hits@N		
		@10	@3	@1
TransE	0.296	0.499	0.330	0.196
SimplE	0.306	0.496	0.341	0.212
RotatE	0.314	0.505	0.347	0.221
ComplEx	0.296	0.489	0.333	0.200
DistMult	0.309	0.506	0.346	0.211
KBAT[b]	0.103	0.337	0.248	0.103
ConvKB	0.407	0.527	0.333	0.200
CoKE[c]	0.362	0.550	0.400	0.269

[a] MSE: mean-square error.

[b] KBAT: knowledge base attention.

[c] CoKE: Contextualized Knowledge Graph Embedding.

Table 7. Performance comparison of link prediction on the TCMdt data set.

Models	MRR[a]	Hits@N		
		@10	@3	@1
TransE	0.243	0.428	0.279	0.150
SimplE	0.162	0.436	0.222	0.113
RotatE	0.146	0.424	0.193	0.090
ComplEx	0.137	0.411	0.177	0.080
DistMult	0.164	0.438	0.223	0.117
ConvKB	0.271	0.464	0.302	0.192
CoKE[b]	0.332	0.491	0.365	0.250
KBAT[c]	0.129	0.369	0.178	0.088
CoKE-multihop	0.251	0.515	0.278	0.261
CoKE-multihop-distillation	0.32	0.483	0.374	0.260

[a] MSE: mean-square error.

[b] KBAT: knowledge base attention.

[c] CoKE: Contextualized Knowledge Graph Embedding.

Table 8. Results of 10-fold cross-validation of deep learning multilabel models.

Index	Precision	Recall	F1 score
MLKNN[a] (Hamming loss=0.186; best parameter: K=26)			
Micro-avg	0.810	0.710	0.760
Macro-avg	0.800	0.610	0.660
RandomForest-RAkEL[b] (Hamming loss=0.186; best parameter: n_estimators=800)			
Micro-avg	0.790	0.740	0.760
Macro-avg	0.760	0.640	0.670
LogisticRegression-RAkEL (Hamming loss=0.173; best parameter: C=0.5)			
Micro-avg	0.810	0.750	0.780
Macro-avg	0.760	0.660	0.700
DNN[c] (Hamming loss=0.186; best parameters: hidden=500, layer=3)			
Micro-avg	0.790	0.740	0.760
Macro-avg	0.750	0.670	0.700
DNN+LSTM[d]-KGE[e] (Hamming loss=0.167; best parameters: hidden=500, layer=3, LSTM=128)			
Micro-avg	0.800	0.790	0.790
Macro-avg	0.740	0.740	0.740
DNN+BILSTM[f]-KGE (Hamming loss=0.127; best parameter: LSTM=128)			
Micro-avg	0.860	0.820	0.840
Macro-avg	0.810	0.770	0.790

[a]MLKNN: multilabel k nearest neighbors.

[b]RAkEL: random k-labelsets.

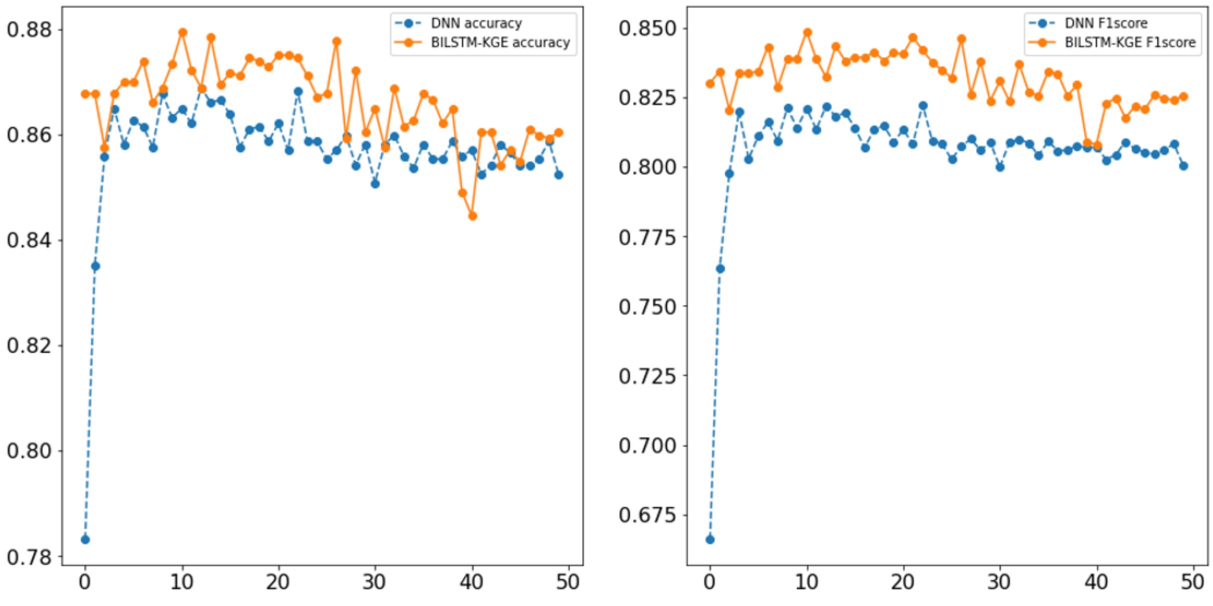
[c]DNN: deep neural network.

[d]LSTM: long short-term memory.

[e]KGE: knowledge graph embedding.

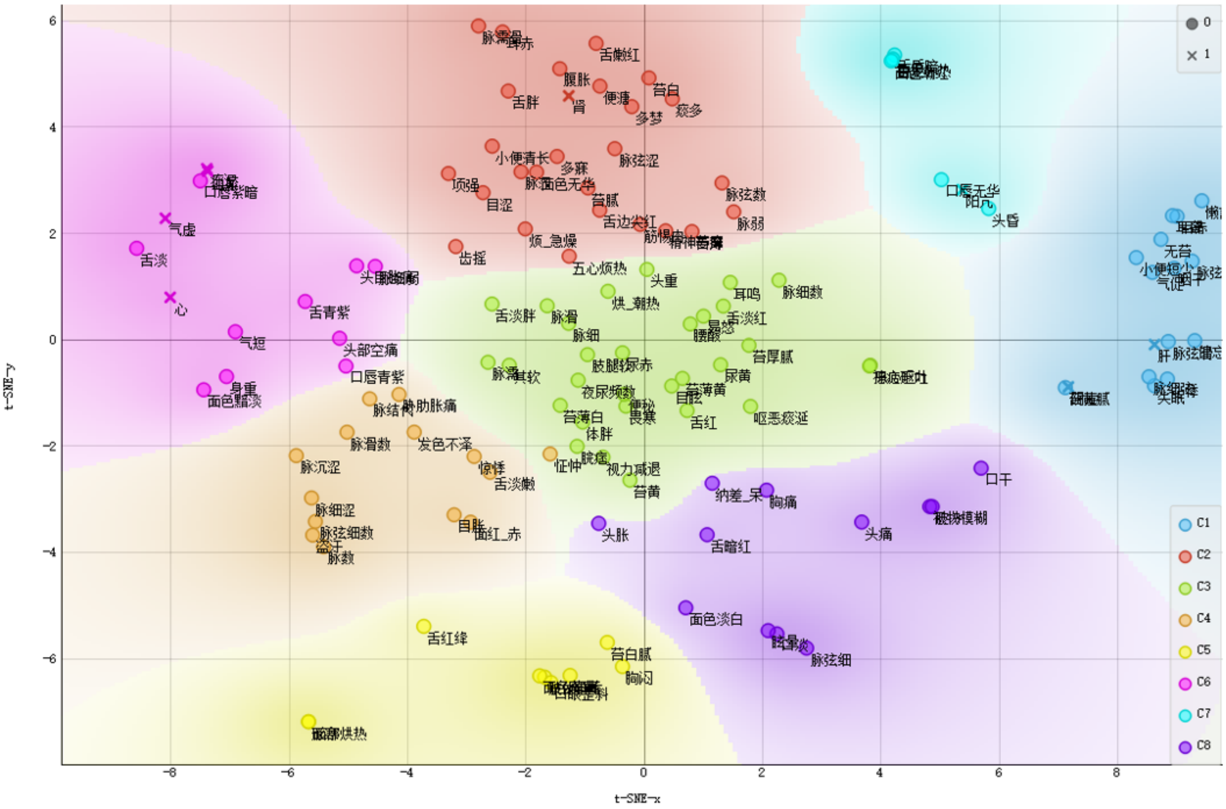
[f]BILSTM: bidirectional long short-term memory.

Figure 4. Performances of DNN and DNN+BILSTM-KGE. BILSTM: bidirectional long short-term memory; DNN: deep neural network; KGE: knowledge graph embedding.



Learned representations of entities were visualized by t-SNE, as shown in Figure 5. Symptoms and TCM syndrome elements are denoted by ○ and X, respectively. The representation distribution conformed to theoretical common sense in TCM with obvious boundaries (ie, silhouette score>0.44) between different classes of TCM syndromes. Intuitively, the learned representations preserved the semantic information about TCM syndromes by using the proposed KGE learning methods. In addition, the relation between entities *Yang hyperactivity* and *dizziness* was similar to the relation between entities *liver depression* and *stringy pulse*, indicating that the semantic constraint of translational distance is preserved after training. The results show that representations learned by the proposed KGE learning method are capable of providing semantic information in TCM.

Figure 5. Learned representations of entity visualization.



Discussion

Principal Findings

The experiments show that the CoKE model has a more stable performance and can be used for improving downstream tasks. We assume that downstream tasks may be improved by KGE learning, since semantic information provided by KGE is preserved in learned representations of missing entities and relations in a KGC task. KGE is suitable to be applied in scenarios that suffer from incompleteness issues, including knowledge discovery for diagnosis and treatment and assisted decision-making in TCM. Based on the clinical KGE model, we automatically extracted the information about dominant diseases treated by Chinese physicians, evidence, symptoms, theories, treatment methods, prescriptions, medicines, and concept mappings according to the definition of clinical knowledge ontology by the physicians. Inspired by Luo et al [43] and Jin et al [44], the triples in a clinical KG are used to learn a personalized KGE model of Chinese physicians.

The problem of incompleteness of a KG is alleviated by entity link prediction of the personalized KGE model. Through the visualization of the KG, our system assists experts in identifying and expanding the potential relations and neighbors of entities in order to obtain explicitness of the implicit knowledge. Through multiple iterations of embedded learning, the KGE model is suitable for treatment decision-making of Chinese physicians. The theories, treatment methods, prescriptions, capability of cause-effect reasoning, and interpretability are enhanced.

Consisting of theories, treatment methods, prescriptions, and medicines of endometriosis (EM) in TCM, the visualization of our KG is shown in Figure 6. A personalized KG for gynecology is constructed to assist experts in knowledge discovery and decision-making. The thickness of the arrows represents the strength of the potential causality, and the size of the nodes represents their importance in the KG of EM in gynecology. Our system clusters the nodes and represents them with different colors of the clusters. Different shapes of nodes represent different entity types.

We referred to a large amount of ancient and modern literature and the diagnosis and treatment data of Chinese and Western medicine, combined with the techniques of entity extraction and causality extraction in natural language processing. According to the definition of domain knowledge by Chinese physicians, valid entities and relations from real cases include the names of TCM diseases, Chinese medicines and prescriptions, tests and examinations, names of Western medicines and diseases, TCM symptoms, and hospital departments. In the training procedure, the weights of the CoKE model were updated until convergence in order to generate embedding vectors that captured semantic features for clinical interpretability. The proposed model can be applied for personalized recommendations of Chinese physicians, question answering, and optimization of diagnostic models.

Inspired by the heterogeneous network representation learning model [45], a framework for knowledge discovery and decision-making in TCM was proposed, as shown in Figure 7.

Figure 6. Visualization of a personalized KG that consists of theories, treatment methods, prescriptions, and medicines of EM in TCM. EM: endometriosis; KG: knowledge graph; TCM: traditional Chinese medicine.

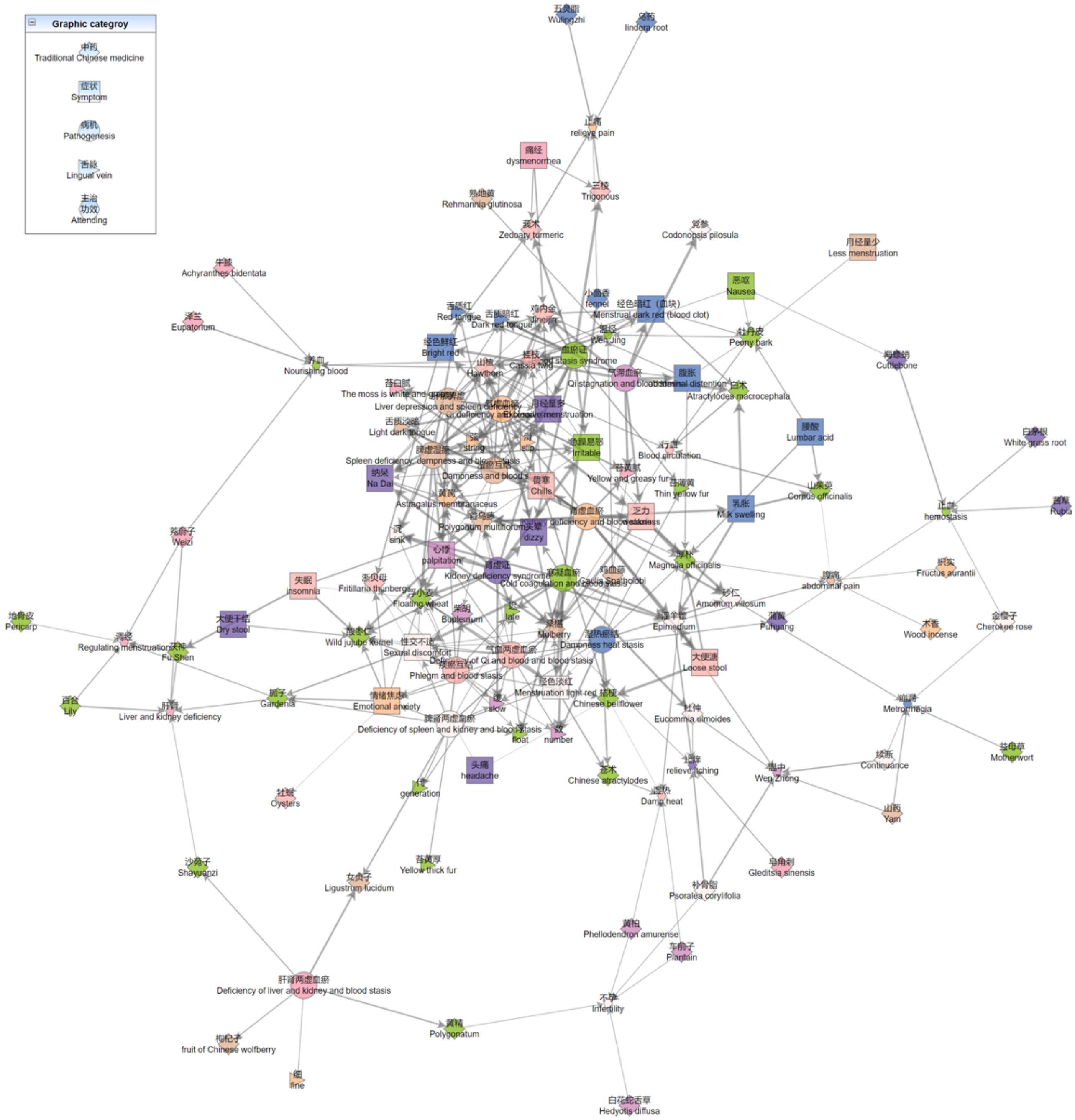
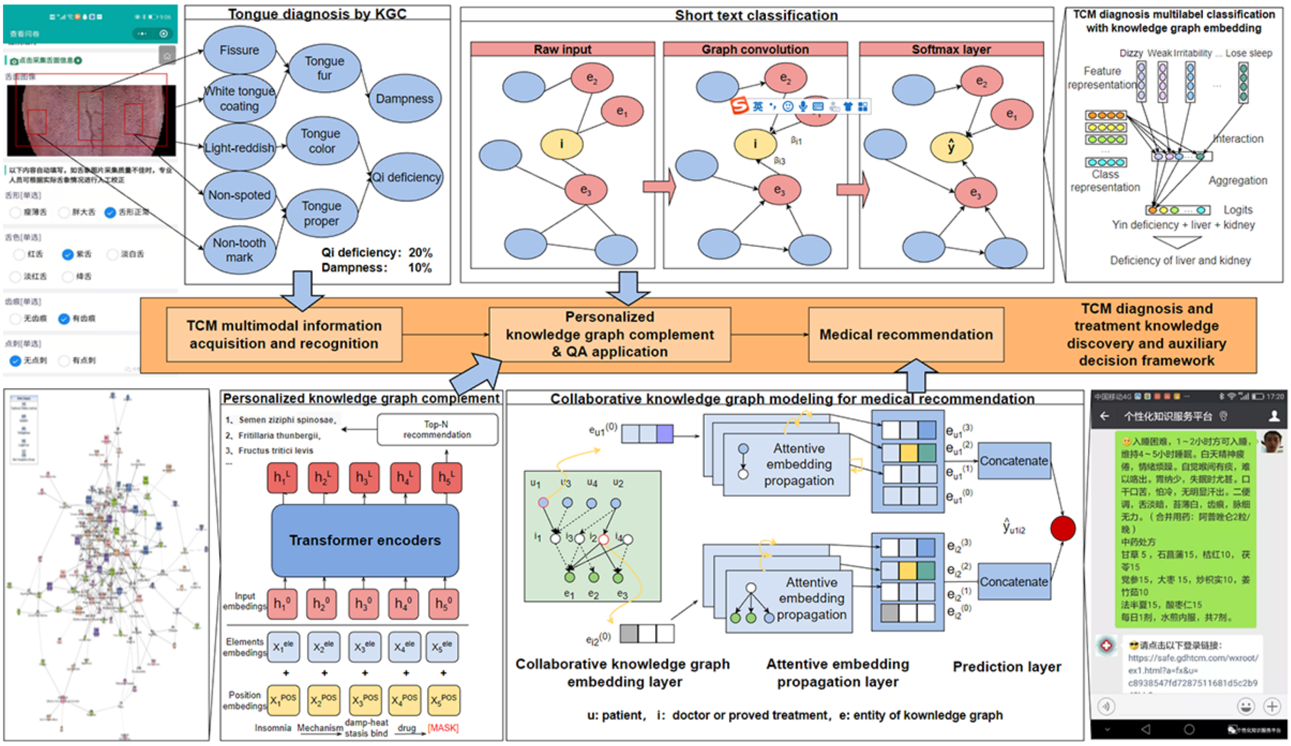


Figure 7. Application of the framework to knowledge discovery and decision-making in TCM. CKG: collaborative knowledge graph; TCM: traditional Chinese medicine; QA: question and answer.



查看问卷

点击采集舌图信息

舌面图像

以下内容自动填写，如舌面图片采集质量不佳时，专业人员可根据实际舌象情况进行人工校正

舌形[单选] ○ 瘦薄舌 ○ 胖大舌 ● 舌形正常

舌色[单选] ○ 红舌 ● 紫舌 ○ 淡白舌 ○ 淡红舌 ○ 绛舌

齿痕[单选] ○ 无齿痕 ● 有齿痕

点刺[单选] ● 无点刺 ○ 有点刺

个性化知识服务平台

入睡困难，1～2小时方可入睡，维持4～5小时睡眠。白天精神疲倦，情绪烦躁。自觉喉间有痰，难以咯出。胃纳少，失眠时尤甚。口干口苦，怕冷，无明显汗出。二便调，舌淡暗，苔薄白，齿痕，脉细无力。（合并用药：阿普唑仑2粒/晚）

中药处方

甘草 5，石菖蒲15，桔红10，茯苓15

党参15，大枣 15，炒枳实10，姜竹茹10

法半夏15，酸枣仁15

每日1剂，水煎内服，共7剂。

请点击以下登录链接：https://safe.gdhtcm.com/wxroot/ex1.html?a=fx&u=c8938547fd7287511681d5c2b9

Tongue diagnosis by KGC
Fissure
White tongue coating
Light-reddish
Non-spoted
Non-tooth mark
Tongue fur
Tongue color
Tongue proper
Dampness
Qi deficiency
Qi deficiency: 20%
Dampness: 10%
Short text classification
Raw input
Graph convolution
Softmax layer
e2
e1
i
e3
ŷ
TCM diagnosis multilabel classification with knowledge graph embedding
Dizzy
Weak
Irritability
...
Lose sleep
Feature representation
Class representation
Interaction
Aggregation
Logits
Yin deficiency + liver + kidney
Deficiency of liver and kidney
TCM multimodal information acquisition and recognition
Personalized knowledge graph complement & QA application
Medical recommendation
TCM diagnosis and treatment knowledge discovery and auxiliary decision framework
Personalized knowledge graph complement
1、Semen ziziphi spinosae,
2、Fritillaria thunbergii,
3、Fructus tritici levis
...
Top-N recommendation
Transformer encoders
Input embedings
Elements embedings
Position embedings
Insomnia
Mechanism
damp-heat stasis bind
drug
[MASK]
Collaborative knowledge graph modeling for medical recommendation
u1 u3 u4 u2
i1 i3 i2 i4
e1 e2 e3
Attentive embedding propagation
Concatenate
Collaborative knowledge graph embedding layer
Attentive embedding propagation layer
Prediction layer
u: patient, i: doctor or proved treatment, e: entity of kownledge graph

For medical recommendation and assisted decision-making, the first step is to collect objective information about the four diagnostic methods. The clinical KG incorporates multimodal information recognized from tongue and facial diagnosis equipment, which can be used to improve the performance of models, even in few-shot learning scenarios. KGs can be used to effectively solve the problems of sparsity and cold start in recommendation systems. Integrating KGs into recommendation systems as external information facilitates the systems with common-sense reasoning capability. Based on the powerful capability of information aggregation and the inference of GNNs, we designed a recommendation system to recommend symptoms, diseases, and Chinese physicians, which effectively improves the performance of recommendations. In addition, the information propagation and inference capability of GNNs also provide interpretability for the results of recommendations.

The model can be used for high-quality assisted decision-making in diagnosis and treatment based on multimodal information and specialty questionnaires. Our system helps practitioners and patients efficiently build online profiles, which enhances the research value of clinical cases. Constructed from natural language, KGs have a strong connection to text mining. KGE can be used to boost the performance of models for text classification and generation. For example, KGE can be leveraged for entity disambiguation when answering the question of what glucose-lowering drug is better for obese diabetics. Similar to link prediction, knowledge inference in question answering infers new relations between entities, given a KG, which is often a multihop relation inference process. For instance, the question can be viewed as a query which can be predicted by PathQuery answering of CoKE for medicine recommendation to obtain related medicines, including *metformin* [46-49].

Conclusion

In this paper, a KG-fused multihop relational adaptive CoKE framework was proposed for screening enhancement, knowledge complement, knowledge inference, and knowledge distillation. The superiority of the model in knowledge discovery and assisted decision-making in TCM was shown in experiments and clinical practice. TCM is a systematic discipline focusing on inheritance and practice. A large amount of knowledge is hidden in the ancient literature and experimental cases of Chinese physicians, which can be mined by researchers. In the future, we aim to improve the quality of the intelligent system of human-machine collaborative KGs in TCM. More in-depth research will be conducted on the knowledge fusion of heterogeneous GNNs, complex inference of KGs with GNNs, and interpretable learning of GNNs.

Acknowledgments

The work was supported by grants from the National Natural Science Foundation of China (#61871141) research and development plan projects in key areas of Guangdong Province (#2021A1111120008), the Collaborative Innovation Team of Guangzhou University of Traditional Chinese Medicine (#2021XK08), and the State Key Laboratory of Dampness Syndrome of Chinese Medicine (#SZ2021ZZ3004, SZ2021ZZ01).

Conflicts of Interest

None declared.

References

1. Du J, Shi D. The advantages of TCM in treating chronic diseases and the inspiration of TCM to modern medical treatment model. Beijing J Tradit Chin Med (in Chinese) 2010;29(4):3.
2. Weng H, Liu Z, Yan S. A framework for automated knowledge graph construction towards traditional chinese medicine. Health Info Sci 2017:170. [doi: 10.1007/978-3-319-69182-4_18]
3. Google Knowledge Graph Search API. URL: https://developers.google.com/knowledge-graph [accessed 2022-08-09]
4. Paulheim H. Data-driven joint debugging of the dbpedia mappings and ontology. 2017 Presented at: 14th International European Semantic Web Conference; May 28 to June 1, 2017; Portorož, Slovenia p. 404-418. [doi: 10.1007/978-3-319-58068-5_25]
5. Suchanek FM, Kasneci G, Weikum G. Yago: a core of semantic knowledge. 2007 Presented at: 16th International Conference on World Wide Web; May 8-12, 2007; Banff, Alberta, Canada p. 697-706. [doi: 10.1145/1242572.1242667]
6. Bollacker K, Evans C, Paritosh P. Freebase: a collaboratively created graph database for structuring human knowledge. 2018 Presented at: SIGMOD/PODS '08: International Conference on Management of Data; June 9-12, 2008; Vancouver Canada p. 1247-1250. [doi: 10.1145/1376616.1376746]
7. Liu Z, Cui A. Big Data Intelligence: Machine Learning and Natural Language Processing in the Internet Age. Beijing: Publishing House of Electronics Industry; 2016.
8. Cheng X, Jin X, Wang Y, Guo J, Zhang T, Li G. Survey on big data system and analytic technology. J Softw 2014(9):1889-1908.
9. Zhou X, Cao C. Medical Knowledge Acquisition: An Ontology - Based Approach. China: Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences; 2003.
10. Jia L, Liu J, Yu T, Zhu L, Gao B, Liu L. Construction of traditional chinese medicine knowledge graph. J Med Inform 2015:51-53.
11. Jia L, Zhu L, Dong Y. Study and establishment of appraisal system for traditional chinese medicine language system. Chin Digit Med 2012;07(010):13-16.
12. Rotmensch M, Halpern Y, Tlimat A, Horng S, Sontag D. Learning a health knowledge graph from electronic medical records. Sci Rep 2017 Jul 20;7(1):5994 [FREE Full text] [doi: 10.1038/s41598-017-05778-z] [Medline: 28729710]
13. Bordes A, Usunier N, Garcia-Duran A. Translating embeddings for modeling multi-relational data. Adv Neural Info Proc Syst 2013:26.
14. Wang Z, Zhang J, Feng J, Chen Z. Knowledge graph embedding by translating on hyperplanes. AAAI 2014 Jun 21;28(1):1112-1119. [doi: 10.1609/aaai.v28i1.8870]
15. Lin Y, Liu Z, Sun M, Liu Y, Zhu X. Learning entity and relation embeddings for knowledge graph completion. AAAI 2015 Feb 19;29(1):2181-2187. [doi: 10.1609/aaai.v29i1.9491]
16. Xiao H, Huang M, Zhu X. From one point to a manifold: knowledge graph embedding for precise link prediction. 2015 Presented at: 25th International Joint Conference on Artificial Intelligence; July 9-15, 2016; New York, NY p. 1315-1321.
17. Ji G, He S, Xu L, Liu K, Zhao J. Knowledge graph embedding via dynamic mapping matrix. 2015 Presented at: 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing; July 2015; Beijing, China p. 687-397. [doi: 10.3115/v1/p15-1067]
18. Nickel M, Tresp V, Kriegel HP. A three-way model for collective learning on multi-relational data. 2011 Presented at: 28th International Conference on Machine Learning; June 28 to July 2, 2011; Bellevue, WA p. 809-816.
19. Yang B, Yih W, He X, Gao J, Deng L. Embedding entities and relations for learning and inference in knowledge bases. ICLR 2015:13.
20. Nickel M, Rosasco L, Poggio T. Holographic embeddings of knowledge graphs. AAAI 2016 Mar 02;30(1):1955-1961. [doi: 10.1609/aaai.v30i1.10314]
21. Trouillon T, Welbl J, Riedel S, Gaussier E, Bouchard G. Complex embeddings for simple link prediction. ICML 2016:2071-2080.
22. Kazemi SM, Poole D. Simple embedding for link prediction in knowledge graphs. 2018 Presented at: NeurIPS 2018: Annual Conference on Neural Information Processing Systems; December 3-8, 2018; Montréal, Canada p. 4284-4295.
23. Dettmers T, Minervini P, Stenetorp P, Riedel S. Convolutional 2D knowledge graph embeddings. 2018 Apr 25 Presented at: AAAI-18: Thirty-Second AAAI Conference on Artificial Intelligence; February 2-7, 2018; New Orleans, LA p. 1811-1818. [doi: 10.1609/aaai.v32i1.11573]
24. Nguyen DQ, Nguyen T, NguyenPhung D. A novel embedding model for knowledge base completion based on convolutional neural network. 2018 Presented at: 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 1-6, 2018; New Orleans, LA p. 327-333. [doi: 10.18653/v1/n18-2053]

25. Vu T, Nguyen TD, Nguyen DQ. A capsule network-based embedding model for knowledge graph completion and search personalization. 2019 Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 1-8, 2019; Minneapolis, MN p. 2180-2189. [doi: 10.18653/v1/n19-1226]
26. Nathani D, Chauhan J, Sharma C. Learning attention-based embeddings for relation prediction in knowledge graphs. 2019 Presented at: 57th Annual Meeting of the Association for Computational Linguistics; July 28 to August 2, 2019; Florence, Italy p. 4710-4723. [doi: 10.18653/v1/p19-1466]
27. Devlin J, Toutanova LK. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 1-8, 2019; Minneapolis, MN p. 4171-4186.
28. Vaswani A, Shazeer N, Parmar N. Attention is all you need. 2017 Presented at: 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA p. 6000-6010.
29. Wang Q, Huang P, Wang H. Coke: Contextualized knowledge graph embedding. arXiv 2019:2168. [doi: 10.1090/mbk/121/79]
30. Qian W, Fu C, Zhu Y, Cai D, He X. Translating embeddings for knowledge graph completion with relation attention mechanism. 2018 Presented at: Twenty-Seventh International Joint Conference on Artificial Intelligence; July 13-19, 2018; Stockholm, Sweden p. 4286-4292. [doi: 10.24963/ijcai.2018/596]
31. Shen Y, Yuan K, Dai J, Tang B, Yang M, Lei K. KGDDS: a system for drug-drug similarity measure in therapeutic substitution based on knowledge graph curation. J Med Syst 2019 Mar 05;43(4):92. [doi: 10.1007/s10916-019-1182-z] [Medline: 30834481]
32. Zheng W, Yan L, Gou C, Zhang ZC, Jason Zhang J, Hu M, et al. Pay attention to doctor-patient dialogues: multi-modal knowledge graph attention image-text embedding for COVID-19 diagnosis. Inf Fusion 2021 Nov;75:168-185 [FREE Full text] [doi: 10.1016/j.inffus.2021.05.015] [Medline: 34093095]
33. Li J, Maharjan B, Xie B, Tao C. A personalized voice-based diet assistant for caregivers of Alzheimer disease and related dementias: system development and validation. J Med Internet Res 2020 Sep 21;22(9):e19897 [FREE Full text] [doi: 10.2196/19897] [Medline: 32955452]
34. Lin Y, Liu Z, Luan H. Modeling relation paths for representation learning of knowledge bases. 2015 Presented at: Conference on Empirical Methods in Natural Language Processing; September 17-21, 2015; Lisbon, Portugal p. 705-714. [doi: 10.18653/v1/d15-1082]
35. Kadlec R, Bajgar O, Kleindienst J. Knowledge base completion: baselines strike back. 2017 Presented at: 2nd Workshop on Representation Learning for NLP; August 2017; Vancouver, Canada p. 69-74. [doi: 10.18653/v1/w17-2609]
36. Guu K, Miller J, Liang P. Traversing knowledge graphs in vector space. 2015 Presented at: Conference on Empirical Methods in Natural Language Processing; September 2015; Lisbon, Portugal p. 318-327. [doi: 10.18653/v1/d15-1038]
37. Jiao X, Yin Y, Shang L, Jiang X, Li L, Wang F, et al. TinyBERT: distilling BERT for natural language understanding. In: Findings of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics; 2020:4163-4174.
38. Toutanova K. Observed versus latent features for knowledge base and text inference. 2015 Presented at: 3rd Workshop on Continuous Vector Space Models and their Compositionality; 2015; Beijing, China p. 57-66. [doi: 10.18653/v1/w15-4007]
39. Ou A, Lin X, Li G. LEVIS: a hypertension dataset in traditional Chinese medicine. 2013 Presented at: IEEE International Conference on Bioinformatics and Biomedicine; December 18-21, 2013; Shanghai, China p. 192-197. [doi: 10.1109/bibm.2013.6732672]
40. Sun Z, Vashishth S, Sanyal S. A re-evaluation of knowledge graph completion methods. 2020 Presented at: 58th Annual Meeting of the Association for Computational Linguistics; July 5-10, 2020; Seattle, WA p. 5516-5522. [doi: 10.18653/v1/2020.acl-main.489]
41. Ruffinelli D, Broscheit S, Gemulla R. You can teach an old dog new tricks! on training knowledge graph embeddings. 2019 Presented at: 7th International Conference on Learning Representations; May 6-9, 2019; New Orleans, LA. [doi: 10.4324/9781315657691-66]
42. Maxwell A, Li R, Yang B, Weng H, Ou A, Hong H, et al. Deep learning architectures for multi-label classification of intelligent health risk prediction. BMC Bioinform 2017 Dec 28;18(Suppl 14):523 [FREE Full text] [doi: 10.1186/s12859-017-1898-z] [Medline: 29297288]
43. Luo Y, Hou H, Lu J. Analysis of the law of Professor Yang Nizhi for diabetic kidney disease based on knowledge graph experimental mining. Modernizat Tradit Chin Med Materia Med-World Sci Technol 2020;22(5):1464-1471.
44. Jin L, Zhang T, He W. An analysis of clinical characteristics and prescription patterns of Professor Zhang Zhongde. Modernizat Tradit Chin Med Materia Med-World Sci Technol 2021:1-11.
45. Yang C, Xiao Y, Zhang Y, Sun Y, Han J. Heterogeneous Network Representation Learning: A Unified Framework with Survey and Benchmark. IEEE Transactions on Knowledge & Data Engineering 2020;01:1-1.
46. Hu L, Yang T, Shi C. Research progress of knowledge graph based on graph neural network. Commun CCF 2020;016(008):38.
47. Qiu X, Sun T, Xu Y, Shao Y, Dai N, Huang X. Pre-trained models for natural language processing: a survey. Sci Chin Technol Sci 2020 Sep 15;63(10):1872-1897. [doi: 10.1007/s11431-020-1647-3]

48. Du B, Wan G, Ji Y. A review of knowledge graph techniques from the view of geometric deep learning. Aero Weaponry (in Chinese) 2020;27(3):1-10.
49. Guan S, Jin X, Jia Y, Wang Y, Cheng X. Knowledge reasoning over knowledge graph: a survey. j Softw 2018;29(10):2966-2994.

Abbreviations

ADRD: Alzheimer disease and related dementia
AI: artificial intelligence
BILSTM: bidirectional long short-term memory
CKD: chronic kidney disease
CoKE: Contextualized Knowledge Graph Embedding
DNN: deep neural network
EM: endometriosis
EMR: electronic medical record
GNN: graph neural network
KBAT: knowledge base attention
KG: knowledge graph
KGC: knowledge graph completion
KGE: knowledge graph embedding
MLKNN: multilabel k nearest neighbors
MRR: mean reciprocal rank
MSE: mean-square error
RAkEL: random k-labelsets
TCM: traditional Chinese medicine
TCMLS: Traditional Chinese Medicine Language System

Edited by T Hao; submitted 31.03.22; peer-reviewed by C Zhang, J Li; comments to author 09.05.22; revised version received 04.07.22; accepted 27.07.22; published 02.09.22.

Please cite as:
Weng H, Chen J, Ou A, Lao Y
Leveraging Representation Learning for the Construction and Application of a Knowledge Graph for Traditional Chinese Medicine: Framework Development Study
JMIR Med Inform 2022;10(9):e38414
URL: https://medinform.jmir.org/2022/9/e38414
doi:10.2196/38414
PMID:36053574

Original Paper

Identification of Preterm Labor Evaluation Visits and Extraction of Cervical Length Measures from Electronic Health Records Within a Large Integrated Health Care System: Algorithm Development and Validation

Fagen Xie[1], PhD; Nehaa Khadka[1], MPH; Michael J Fassett[2,3], MD; Vicki Y Chiu[1], MS; Chantal C Avila[1], MA; Jiaxiao Shi[1], PhD; Meiyu Yeh[1], MS; Aniket Kawatkar[1], PhD; Nana A Mensah[1], PhD; David A Sacks[1], MD; Darios Getahun[1,4], MD, MPH, PhD

[1]Department of Research and Evaluation, Kaiser Permanente Southern California, Pasadena, CA, United States

[2]Department of Obstetrics & Gynecology, Kaiser Permanente West Los Angeles Medical Center, Los Angeles, CA, United States

[3]Department of Clinical Science, Kaiser Permanente Bernard J Tyson School of Medicine, Pasadena, CA, United States

[4]Department of Health Systems Science, Kaiser Permanente Bernard J. Tyson School of Medicine, Pasadena, CA, United States

Corresponding Author:
Fagen Xie, PhD
Department of Research and Evaluation
Kaiser Permanente Southern California
100 S. Los Robles Avenue
2nd Floor
Pasadena, CA, 91101
United States
Phone: 1 626 564 3294
Fax: 1 626 564 787
Email: fagen.xie@kp.org

Abstract

Background: Preterm birth (PTB) represents a significant public health problem in the United States and throughout the world. Accurate identification of preterm labor (PTL) evaluation visits is the first step in conducting PTB-related research.

Objective: We aimed to develop a validated computerized algorithm to identify PTL evaluation visits and extract cervical length (CL) measures from electronic health records (EHRs) within a large integrated health care system.

Methods: We used data extracted from the EHRs at Kaiser Permanente Southern California between 2009 and 2020. First, we identified triage and hospital encounters with fetal fibronectin (fFN) tests, transvaginal ultrasound (TVUS) procedures, PTL medications, or PTL diagnosis codes within $24^{0/7}$-$34^{6/7}$ gestational weeks. Second, clinical notes associated with triage and hospital encounters within $24^{0/7}$-$34^{6/7}$ gestational weeks were extracted from EHRs. A computerized algorithm and an automated process were developed and refined by multiple iterations of chart review and adjudication to search the following PTL indicators: fFN tests, TVUS procedures, abdominal pain, uterine contractions, PTL medications, and descriptions of PTL evaluations. An additional process was constructed to extract the CLs from the corresponding clinical notes of these identified PTL evaluation visits.

Results: A total of 441,673 live birth pregnancies were identified between 2009 and 2020. Of these, 103,139 pregnancies (23.35%) had documented PTL evaluation visits identified by the computerized algorithm. The trend of pregnancies with PTL evaluation visits slightly decreased from 24.41% (2009) to 17.42% (2020). Of the first 103,139 PTL visits, 19,439 (18.85%) and 44,423 (43.97%) had an fFN test and a TVUS, respectively. The percentage of first PTL visits with an fFN test decreased from 18.06% at $24^{0/7}$ gestational weeks to 2.32% at $34^{6/7}$ gestational weeks, and TVUS from 54.67% at $24^{0/7}$ gestational weeks to 12.05% in $34^{6/7}$ gestational weeks. The mean (SD) of the CL was 3.66 (0.99) cm with a mean range of 3.61-3.69 cm that remained stable across the study period. Of the pregnancies with PTL evaluation visits, the rate of PTB remained stable over time (20,399, 19.78%). Validation of the computerized algorithms against 100 randomly selected records from these potential PTL visits showed positive predictive values of 97%, 94.44%, 100%, and 96.43% for the PTL evaluation visits, fFN tests, TVUS, and CL, respectively, along with sensitivity values of 100%, 90%, and 90%, and specificity values of 98.8%, 100%, and 98.6% for the fFN test, TVUS, and CL, respectively.

Conclusions: The developed computerized algorithm effectively identified PTL evaluation visits and extracted the corresponding CL measures from the EHRs. Validation against this algorithm achieved a high level of accuracy. This computerized algorithm can be used for conducting PTL- or PTB-related pharmacoepidemiologic studies and patient care reviews.

(JMIR Med Inform 2022;10(9):e37896) doi:10.2196/37896

KEYWORDS

preterm labor; preterm birth; fetal fibronectin; transvaginal ultrasound; cervical length; natural language processing; computerized algorithm; data extraction; patient records; clinical notes; evaluation notes; patient care; patient notes; electronic health records

Introduction

Preterm birth (PTB, the birth of a child before $37^{0/7}$ weeks of gestation) occurs in nearly 10% of live births in the United States [1,2]. It is one of the leading causes of infant morbidity and mortality in the United States and throughout the world [3,4] and constitutes a significant public health burden [2]. The majority of PTBs are spontaneous or idiopathic, whereas the remaining are medically indicated due to fetal or maternal complications [5-7]. Surviving infants are at significantly increased risk for long-term sequelae, including respiratory, gastrointestinal, central nervous system, hearing, and vision problems, as well as long-term cognitive, motor, and behavioral delays with long-lasting effects [2].

The identification of pregnant women at high risk for imminent spontaneous PTB (sPTB) is critical for appropriate and timely management of preterm labor (PTL), including timely administration of antenatal corticosteroids and magnesium sulfate for accelerating fetal lung maturation and neuroprotection [8-11]. On the other hand, accurate assessment of the risk of sPTB including cervical examination and observation of clinical signs and symptoms can allow for better timing of antenatal corticosteroid administration, avoid unnecessary interventions, and reduce costs. Fetal fibronectin (fFN) testing [12] and transvaginal ultrasound (TVUS) measurement of the cervical length (CL) prior to 24 weeks [13] have been used as indicators of potential sPTB risk. For instance, a CL measuring over 3 cm [14] or a negative fFN test [15] obtained from a pregnant woman with presumed PTL may rule against PTL and therefore avoid overdiagnosis and unnecessary treatment. Standardized clinical procedures for the assessment and management of pregnant women with suspected signs and symptoms of PTL have been established [16,17], and although not widely implemented, they have shown significant health care cost reduction by avoiding unnecessary hospitalization of pregnant women who may have signs and symptoms of PTL but are not likely to deliver prematurely [18].

One historical challenge in the evaluation of retrospective patient data has been with respect to the ability to incorporate some of these free-text elements in the electronic health record (EHR); despite being rich sources of data, they have been challenging to incorporate into studies without reliable, consistent, and efficient ways to identify these elements and classify them in data analyses. Natural language processing (NLP) is a field of computer-based methods aimed at standardizing and analyzing free text, for allowing inclusion of these free-text data elements even in large data sets [19-23]. It converts medical information residing in natural language into a more structured format for various medical research and patient care management purposes [24-27]. Although there have been fruitful attempts to predict the risk of sPTB [12-15,28,29] with structured EHRs or machine learning approaches, to our knowledge, there is no available automated algorithm to identify PTL evaluation visits among patients presenting at triage or hospitals from the EHR. The ability to examine all cases of threatened PTL in a large data set, their associated methods of evaluation, and their outcomes and costs will ultimately help inform the discussion surrounding the standardization of threatened PTL assessment and the associated use of TVUS and fFN. The purpose of the present study was to develop and validate a computerized NLP algorithm and process to effectively identify PTL evaluation visits and extract corresponding CL data from the EHRs, including free-text clinical notes, within a large integrated health care system.

Methods

Study Setting and Population

Kaiser Permanente Southern California (KPSC) is a large integrated health care system providing comprehensive medical services to over 4.7 million members across 15 large medical center areas. The demographic characteristics of KPSC members are diverse and largely representative of the residents in Southern California [30] with health insurance through group plans, individual plans, Medicare, and Medicaid programs, representing >260 ethnicities and >150 spoken languages. KPSC's extensive EHR data contain individual-level structured data (including diagnosis codes, procedure codes, medications, immunization records, laboratory results, and pregnancy episodes and outcomes) and unstructured data (including free-text clinical notes, radiology reports, pathology reports, imaging, and videos) covering all medical visits across all health care settings (ie, outpatient, inpatient, emergency department, virtual, etc). Clinical care of KPSC members provided by external contracted providers is captured in the EHR through reimbursement claim requests.

Ethics Approval

The study protocol was reviewed and approved by the KPSC Institutional Review Board with a waiver of the requirement for informed consent (approval number: 12670). Only authorized persons were given access permission to perform all analyses.

Identification of PTL Evaluation Visits

The details of PTL assessments are documented in the EHR system in both structured (eg, fFN results, TVUS, and medication) and unstructured (eg, contraction frequency and CL) formats. We conducted a retrospective cohort study including all pregnancies and live births delivered at KPSC hospitals (N=441,673) between 2009 and 2020. The encounters between $24^{0/7}$ and $34^{6/7}$ weeks of gestation for each pregnancy episode and the corresponding medical information including clinical notes were extracted from the KPSC EHR system. The extracted information was then used to develop the computerized algorithm and process for identifying PTL evaluation visits through a refined iterative chart review process by the following steps. The encounters between $20^{0/7}$ and $23^{6/7}$ weeks as well as those between $35^{0/7}$ and $36^{6/7}$ weeks of gestation were excluded because fFN testing was not indicated in these gestational age groups.

Step 1: Based on the codes described in Table A1 of Multimedia Appendix 1, any of the following potential PTL-related encounters for each pregnancy episode were identified and assembled: encounter involving fFN testing, encounter involving TVUS, encounter with PTL diagnosis codes, and encounter with PTL medication.

If any of the above encounters was detected, it was passed to Step 3 for further processing.

Step 2: The evidence or indicator of PTL evaluation was identified from the clinical notes through the following process:

1. Clinical notes associated with triage or hospitalization encounters between $24^{0/7}$and $34^{6/7}$ weeks of gestation for each pregnancy episode were extracted, but these were limited to the notes of interest to the study, as shown in Table A2 of Multimedia Appendix 1. Experienced obstetric gynecologists determined these note types.
2. Extracted clinical notes were preprocessed through letter lowercase conversion and sentence separation and tokenization (ie, segmenting text into linguistic units such as words and punctuation) [20]. The separated sentences were further cleaned up by removing the nondigital or nonletter characters except for spaces, periods, commas, and colons, while correcting misspelled words and standardizing abbreviated words or terms detected from the process of algorithm development. The complete corrected and standardized word lists are summarized in Table A3 of Multimedia Appendix 1.
3. Sentences extracted with at least one of the following predefined keywords are listed in Table A4 of Multimedia Appendix 1: preterm labor, fetal fibronectin, transvaginal ultrasound, abdominal pain, and uterine contraction. These keywords of interest to the study were compiled through consultations with experienced obstetric gynecologists. Sentences without any predefined keywords were not passed for further processing.

The following indicators of PTL evaluation were extracted from the above extracted sentences: performed fetal fibronectin test, performed transvaginal ultrasound, abdominal pain, uterine contraction, and explicit descriptions regarding preterm labor evaluation, such as “in preterm labor,” “ruled out PTL,” and “assessment: preterm labor.” Any negated, general, history-related, and uncertain descriptions were excluded.

If any of the above indicators was detected, the corresponding encounter was defined as a PTL evaluation encounter.

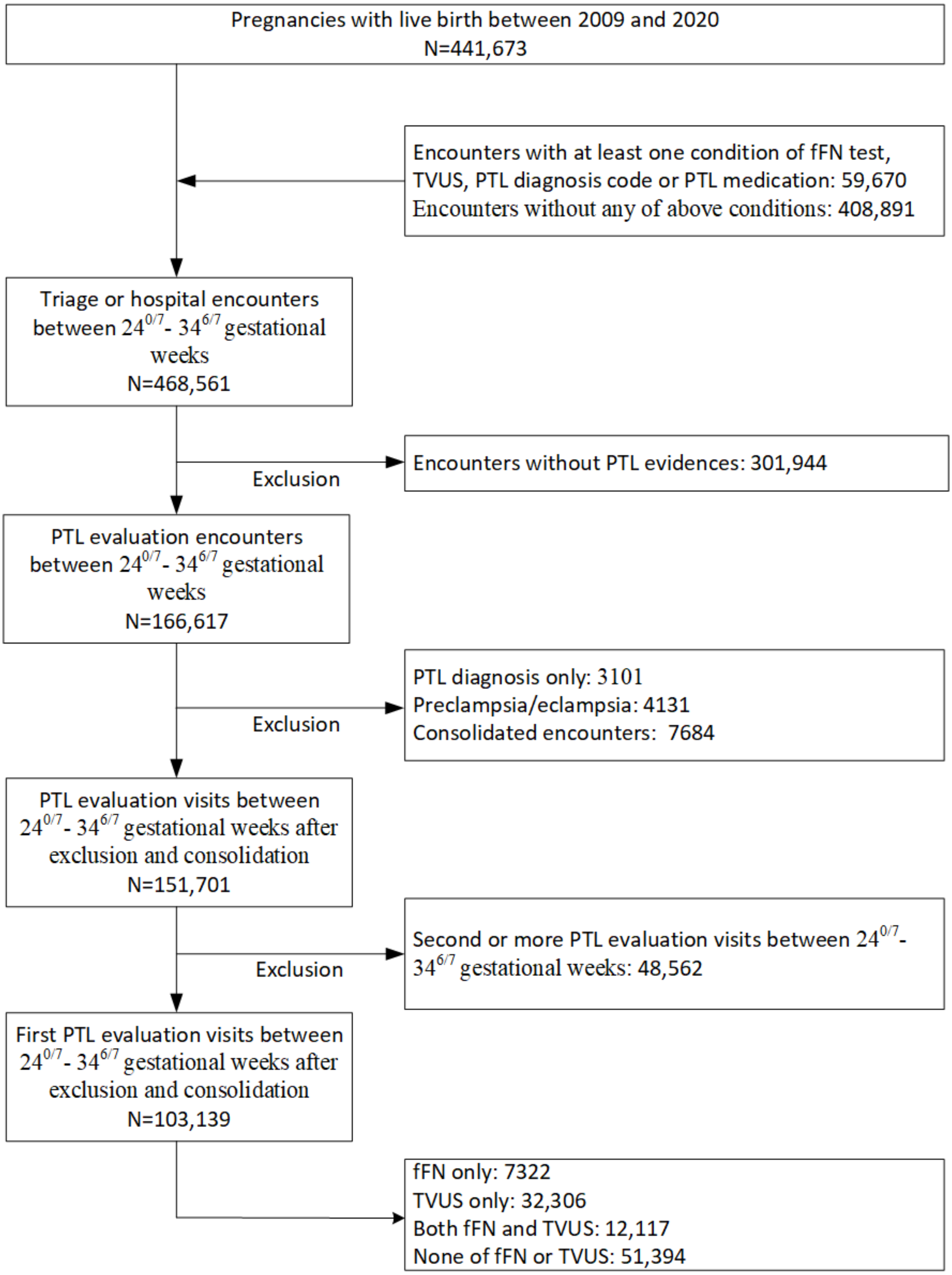
Step 3: The PTL evaluation encounters identified in Step 1 and Step 2 were combined, and deduplication was performed if the same encounter was found multiple times. However, encounters with the following conditions were excluded:

1. The encounter was a delivery encounter in patients with the pre-eclampsia/eclampsia diagnosis code. These were excluded due to potential confounding results related to a medically indicated PTB.
2. The encounter had a PTL diagnosis code but without any other evidence of evaluation for PTL in the same encounter (eg, TVUS, uterine contraction, and fFN test). The percentage of this group was relatively small (1.9%). We decided to exclude these potential cases due to the low confirmed rate from the chart review of a randomly selected sample (see the chart review process below).

Step 4: If the identified PTL encounters had an overlapping time window, these encounters were consolidated as a combined PTL encounter, in which the admitted time was the earlier admitted time, whereas the discharge time was the later discharge time.

Figure 1 presents the number of encounters derived from the process between $24^{0/7}$ and $34^{6/7}$ weeks.

Figure 1. Flowchart showing preterm labor evaluation visits. fFN: fetal fibronectin; PTL: preterm labor; TVUS: transvaginal ultrasound.


Pregnancies with live birth between 2009 and 2020
N=441,673
Encounters with at least one condition of fFN test, TVUS, PTL diagnosis code or PTL medication: 59,670
Encounters without any of above conditions: 408,891
Triage or hospital encounters between 24$^{0/7}$- 34$^{6/7}$ gestational weeks
N=468,561
Exclusion
Encounters without PTL evidences: 301,944
PTL evaluation encounters between 24$^{0/7}$- 34$^{6/7}$ gestational weeks
N=166,617
Exclusion
PTL diagnosis only: 3101
Preclampsia/eclampsia: 4131
Consolidated encounters: 7684
PTL evaluation visits between 24$^{0/7}$- 34$^{6/7}$ gestational weeks after exclusion and consolidation
N=151,701
Exclusion
Second or more PTL evaluation visits between 24$^{0/7}$- 34$^{6/7}$ gestational weeks: 48,562
First PTL evaluation visits between 24$^{0/7}$- 34$^{6/7}$ gestational weeks after exclusion and consolidation
N=103,139
fFN only: 7322
TVUS only: 32,306
Both fFN and TVUS: 12,117
None of fFN or TVUS: 51,394

CL Measurement Extraction

Cervical assessment may be performed via transvaginal or transabdominal ultrasound to determine CL during PTL evaluation visits; it can be used as a guide for either admission to the hospital or discharge home, as well as for making management decisions when interpreted in the context of clinical assessment and fFN where possible [16,17]. The measured CL was usually documented in the clinical notes or radiology reports by the examining health care provider. However, retrieving and formatting this measure presented a challenge due to the wide variety of free-text formats used. Therefore, a computerized process was developed to extract CL measures from clinical notes associated with a particular PTL evaluation encounter as in the following steps.

Step 1: Lists of keywords or phrases used to describe CL were compiled based on the knowledge of conventional usage by experienced obstetric gynecologists and enriched by iterative refinement. The complete lists are summarized in Table A5 of Multimedia Appendix 1 and separated into 3 priority groups.

Step 2: The sentences in each clinical note were searched for the preidentified keywords or phrases. If one of the predefined keywords was identified in a sentence, then Step 3 was performed. If no keyword was detected, the search was stopped, and the algorithm moved to the next note.

Step 3: The numeric values associated with the keyword "forward" within 10 tokens in the same sentence were searched starting from the position where the predefined keyword was found. If no values were found during forward searching, then the potentially associated values were searched "backward" within 5 tokens before the keyword position because some values were described before the keyword. However, the extracted value was ignored or excluded if it described other measures rather than the CL, such as cervical dilation. The retrieved measures could be 1 or multiple values or a range of values. In addition, each value could contain a unit (cm or mm) or not have any unit. Examples include "cervical length measures 1.6 cm," "tvus cl 2.6-2.7 cm no funneling," "cervical length 3.3 to 4.4," and "transvaginal ultrasound at bedside 41 mm long cervical length."

Step 4: The final CL measure was determined for each clinical note based on the keyword or phrase priority. If multiple keywords with different priorities were found in the note, the measured values associated with the keywords with the highest priority were retained. If the retained highest priority group still contained multiple different values, the shortest one was retained.

Step 5: The CL measures were determined for each PTL evaluation visit. A PTL evaluation visit could contain multiple CL values measured at different times. If the encounter was a delivery encounter, the first measure was used as the final CL. Otherwise, the measure closest to discharge was used as the final CL.

Step 6: The CL was standardized and finalized for each PTL evaluation visit. If the measure did not have an associated unit, it was considered cm by default. When the unit was mm/millimeter, the values were divided by 10. Finally, if ranges or multiple values were extracted, then the average value of the extracted values was considered the CL.

Chart Reviews and Validation Process

To validate the computerized algorithm for identifying true PTL evaluation visits in the EHRs, an iterative chart review process was completed by trained research chart abstractors and adjudicated by experienced obstetrician-gynecologists via multiple iterations. The trained research chart abstractors were provided a spreadsheet with the patients' unique medical record numbers and visit encounters with the encounter start and end dates. An encounter was considered a true PTL evaluation visit if any of the following criteria were met based on the review of free text in the medical notes: fFN test performed, TVUS performed, clinician description or mention of PTL in the encounter note, clinician description of contractions or abdominal pain in the encounter note, CL obtained, and administration of a PTL-related medication (eg, tocolysis, magnesium sulfate, and corticosteroids).

If any of the evaluation criteria were marked as "yes," then the encounter was categorized as a PTL evaluation visit. Otherwise, it was not categorized as a PTL evaluation visit. The corresponding supporting information for the decision was documented in detail as well.

First, a sample of 20 encounters was randomly selected from the group with PTL diagnosis codes only but without any other evidence of evaluation for PTL, and the trained research chart abstractors reviewed the chart. Of the 20 encounters, 7 (35%) PTL diagnosis codes were confirmed as PTL evaluation encounters. Due to the low confirmed rate, the encounters with PTL diagnosis only were excluded from further processing. Second, another sample of 20 potential PTL evaluation visits identified by the computerized process was randomly selected for chart review. Among these, 17 (85%) were confirmed as true PTL evaluation visits, and the chart review results were then used for refining and finalizing the process. Finally, 100 potential PTL evaluation visits were randomly selected for full chart review, and the chart review results were used as the reference standard to assess the algorithm's performance to accurately identify true cases of threatened PTL evaluation.

Data Analysis

Results of PTL evaluation visits, fFN tests, TVUS procedures, and CL measurements generated from the computerized algorithm and process were first evaluated against the chart-reviewed and adjudicated reference standard, including their sensitivity, specificity, and positive predictive value (PPV). Descriptive analyses were then conducted to report the distribution of the first identified PTL evaluation visit of each pregnancy episode by birth year, PTB status, and gestational age in detail. Gestational age at birth was based on the clinical estimate and captured as a structured format in the EHRs.

Results

A total of 441,673 live birth pregnancy episodes were extracted from the KPSC EHR system from January 1, 2009, to December 31, 2020. Of them, 103,139 (23.35%) were identified by the computerized algorithm and process with at least 1 PTL evaluation visit between $24^{0/7}$ and $34^{6/7}$ gestation weeks. The percentage of pregnancies with PTL evaluation visits was stable at approximately 24% between 2009 and 2015 and decreased starting in 2016 (Table 1). The annual trend of PTB associated with PTL among these pregnancies with PTL evaluation visits is shown in Table 2. The overall rate of PTB among pregnant women triaged for PTL evaluation was 19.78% and stable at a range of 18%-20% across the study period.

Table 3 presents the distribution of the identified first PTL evaluation visit of each pregnancy with fFN tests, TVUS procedures, and CL measures by birth year. The rate of the performed fFN tests decreased from 28.33% in 2009 to 9.01% in 2020, whereas the percentage of TVUS procedures increased from 36.72% in 2009 to 45.22% in 2020 and the rate of CL reporting increased from 35.32% in 2009 to 42.36% in 2020. In addition, the rate of PTL with both the fFN test and TVUS procedure decreased from 14.64% in 2009 to 6.85% in 2020. The mean CL was 3.66 cm (SD=0.99 cm) and remained relatively stable over the study period.

Table 4 summarizes the distribution of the identified first PTL evaluation visit of each pregnancy with PTB, fFN tests, TVUS procedures, and CL measurements by the corresponding gestation age at the PTL evaluation visit. For the percentage of patients who ultimately had an sPTB varying by gestational age at the time of assessment, the sPTB decreased from 20.75% in patients presenting at $24^{0/7}$-$24^{6/7}$ gestational weeks to 16.7% at $27^{0/7}$-$27^{6/7}$ gestational weeks; it stayed in the range of 16%-19% between $27^{0/7}$ and $30^{6/7}$ gestational weeks and then increased from 19.38% at $31^{0/7}$-$31^{6/7}$ gestational weeks to 24.52% at $34^{0/7}$-$34^{6/7}$ gestational weeks.

Table 1. Trend showing pregnancies resulting in live births with preterm labor evaluation visits within $24^{0/7}$-$34^{6/7}$ gestational weeks by birth year.

Birth year	Live birth pregnancy, N	Live birth pregnancy with preterm labor evaluation visit, n (%)
2009	31,476	7682 (24.41)
2010	31,388	7798 (24.84)
2011	32,896	8084 (24.57)
2012	34,765	8514 (24.49)
2013	34,968	8477 (24.24)
2014	36,148	8993 (24.88)
2015	37,782	9109 (24.11)
2016	39,605	9486 (23.95)
2017	40,030	9412 (23.51)
2018	41,026	9511 (23.18)
2019	41,326	9061 (21.93)
2020	40,263	7012 (17.42)
Overall	441,673	103,139 (23.35)

Table 2. Live birth pregnancies with preterm labor evaluation visits between $24^{0/7}$ and $34^{6/7}$ weeks of gestation by birth year and preterm birth status.

Birth year	Preterm birth status		
	Yes[a], n (%)	No, n (%)	Total (N)
2009	1556 (20.26)	6126 (79.74)	7682
2010	1602 (20.54)	6196 (79.46)	7798
2011	1638 (20.26)	6446 (79.74)	8084
2012	1698 (19.94)	6816 (80.06)	8514
2013	1644 (19.39)	6833 (80.61)	8477
2014	1645 (18.29)	7348 (81.71)	8993
2015	1755 (19.27)	7354 (80.73)	9109
2016	1859 (19.6)	7627 (80.4)	9486
2017	1870 (19.87)	7542 (80.13)	9412
2018	1814 (19.07)	7697 (80.93)	9511
2019	1809 (19.96)	7252 (80.04)	9061
2020	1509 (21.52)	5503 (78.48)	7012
Overall	20,399 (19.78)	82,740 (80.22)	103,139

[a] Yes: preterm births among those pregnancies with preterm labor evaluations.

Table 3. First preterm labor evaluation visit of each pregnancy identified by the computerized algorithm between $24^{0/7}$ and $34^{6/7}$ weeks of gestation by birth year.

Birth year	Total PTL[a], N	Yes for fFN[b], n (%)	Yes[c] for TVUS[d], n (%)	Yes for both fFN and TVUS, n (%)	Cervical length	
					n (%)	Mean (SD), cm
2009	7682	2176 (28.33)	2821 (36.72)	1125 (14.64)	2713 (35.32)	3.62 (1.01)
2010	7798	2145 (27.51)	2958 (37.93)	1129 (14.47)	2847 (36.51)	3.63 (1.01)
2011	8084	2223 (27.5)	3221 (39.84)	1233 (15.25)	3131 (38.73)	3.63 (0.99)
2012	8514	2155 (25.31)	3579 (42.04)	1276 (15)	3482 (40.9)	3.64 (0.99)
2013	8477	2106 (24.84)	3846 (45.37)	1349 (15.91)	3685 (43.47)	3.61 (0.99)
2014	8993	1848 (20.55)	4134 (45.97)	1264 (14.05)	3949 (43.91)	3.64 (1.00)
2015	9109	1653 (18.15)	4278 (46.96)	1113 (12.22)	4103 (45.04)	3.69 (1.00)
2016	9486	1470 (15.5)	4269 (45)	991 (10.44)	4097 (43.19)	3.68 (0.99)
2017	9412	1172 (12.45)	4045 (42.98)	803 (8.53)	3881 (40.23)	3.69 (0.96)
2018	9511	1009 (10.61)	4025 (42.32)	714 (7.51)	3805 (40.01)	3.68 (0.98)
2019	9061	850 (9.38)	3976 (43.88)	640 (7.06)	3762 (41.52)	3.70 (0.98)
2020	7012	632 (9.01)	3171 (45.33)	480 (6.85)	2970 (43.36)	3.65 (1.00)
Overall	103,139	19,439 (18.85)	44,423 (43.97)	12,117 (11.75)	42,425 (41.13)	3.66 (0.99)

[a] PTL: preterm labor.

[b] fFN: fetal fibronectin.

[c] Yes: It means that the column contains patient records with documented transvaginal ultrasound assessment or cervical length values.

[d] TVUS: transvaginal ultrasound.

Table 4. First preterm labor evaluation visit of each pregnancy identified by the computerized algorithm between $24^{0/7}$ and $34^{6/7}$ weeks of gestation by gestational age.

Gestation age of PTL[a] (weeks)	Total PTL cases, N	PTB[b] -Yes[c], n (%)	fFN[d]-Yes, n (%)	TVUS[e]-Yes[f], n (%)	Both fFN and TVUS -Yes, n (%)	Cervical length	
						n (%)	Mean (SD)
$24^{0/7}$-$24^{6/7}$	7691	1596 (20.75)	1397 (18.16)	4205 (54.67)	1013 (13.17)	4009 (52.13)	3.70 (1.06)
$25^{0/7}$-$25^{6/7}$	7496	2468 (19.58)	1403 (18.72)	3983 (53.14)	971 (12.95)	3813 (50.87)	3.73 (1.03)
$26^{0/7}$-$26^{6/7}$	7923	1392 (17.57)	1524 (19.24)	4060 (51.24)	1037 (13.09)	3894 (49.15)	3.76 (1.01)
$27^{0/7}$-$27^{6/7}$	8122	1356 (16.7)	1733 (21.34)	4186 (51.54)	1143 (14.07)	3995 (49.19)	3.75 (0.97)
$28^{0/7}$-$28^{6/7}$	8417	1562 (18.56)	1771 (21.04)	4220 (50.14)	1166 (13.85)	4060 (48.24)	3.71 (0.98)
$29^{0/7}$-$29^{6/7}$	8823	1535 (17.4)	2032 (23.03)	4229 (50.2)	1290 (14.62)	4262 (48.31)	3,68 (0.97)
$30^{0/7}$-$30^{6/7}$	9224	1709 (18.53)	2114 (22.92)	4436 (48.09)	1279 (13.87)	4274 (46.34)	3.67 (0.94)
$31^{0/7}$-$31^{6/7}$	9932	1925 (19.38)	2446 (24.63)	4638 (46.7)	1475 (14.85)	4492 (45.23)	3.59 (0.97)
$32^{0/7}$-$32^{6/7}$	11,158	2234 (20.02)	2639 (23.65)	4752 (42.59)	1520 (13.62)	4567 (40.93)	3.58 (0.95)
$33^{0/7}$-$33^{6/7}$	11,770	2537 (21.55)	2088 (17.74)	3898 (33.12)	1113 (9.46)	3722 (31.62)	3.50 (0.97)
$34^{0/7}$-$34^{6/7}$	12,583	3085 (24.52)	292 (2.32)	1516 (12.05)	100 (0.8)	1337 (10.63)	3.42 (1.08)
Overall	103,139	20,399 (19.78)	19,439 (18.85)	44,423 (43.97)	12117 (11.75)	42,425 (41.13)	3.66 (0.99)

[a] PTL: preterm labor.

[b] PTB: preterm birth.

[c] Yes: It implies preterm births among those pregnancies with preterm labor evaluations.

[d] fFN: fetal fibronectin.

[e] TVUS: transvaginal ultrasound.

[f] Yes: It means that the column contains patient records with documented transvaginal ultrasound assessment or cervical length values.

The percentage of PTL evaluation visits with fFN tests, TVUS procedures, and CL measurements also varied over the gestational age at presentation. fFN testing increased from 18.16% at $24^{0/7}$-$24^{6/7}$ gestational weeks to 24.63% at $31^{0/7}$-$31^{6/7}$ gestational weeks and then dropped significantly to 2.32% at $34^{0/7}$-$34^{6/7}$ gestational weeks. In contrast, the percentage decreased from 54.67% at $24^{0/7}$-$24^{6/7}$ gestational weeks to 12.05% at $34^{0/7}$-$34^{6/7}$ gestational weeks for TVUS procedures, and 52.13% at $24^{0/7}$-$24^{6/7}$ gestational weeks to 10.63% at $34^{0/7}$-$34^{6/7}$ gestational weeks for CL measurements. The mean CL also slightly decreased from 3.7 cm (SD=1.06 cm) at $24^{0/7}$-$24^{6/7}$ gestational weeks to 3.43 cm (SD=1.08 cm) at $34^{0/7}$-$34^{6/7}$ gestational weeks. The trend of PTL evaluation with both fFN tests and TVUS procedures by gestational age had a pattern similar to PTL visits with fFN tests.

The validation of 100 randomly selected PTL evaluation visits identified by the computerized algorithm against the manual chart review (which served as the gold standard) is presented in Table 5. Of the 100 PTL evaluation visits identified by the NLP algorithm, 18 PTL evaluations involved fFN tests, 27 involved TVUS procedures, and 28 involved CL measures. Further, 97 of the 100 were confirmed PTL evaluation visits, 17 of 18 had confirmed fFN tests, all 27 had confirmed TVUS procedures, and 27 of 28 had confirmed CL measurements recorded. The computerized algorithm missed 3 PTL evaluation visits with TVUS performed and 3 CL measurements. The algorithm yielded PPVs of 97%, 94.44%, 100%, and 96.43% for PTL evaluation visits, fFN tests, TVUS procedures, and CL measurements, respectively, and sensitivity values of 100%, 90%, and 90%, along with specificity values of 98.8%, 100%, and 98.6% for fFN tests, TVUS procedures, and CL measurements, respectively, as observed in Table 6.

Table 5. Validation results of the preterm labor evaluation and cervical length measures extraction algorithm.

Computerized results	Total (N)	Status after chart review	
		Yes, n	No, n
Preterm labor evaluation visits	100	97	3
Fetal fibronectin test			
Yes	18	17	1
No	82	0	82
Transvaginal ultrasound			
Yes	27	27	0
No	73	3	70
Cervical length			
Yes-same value	27	27	0
Yes-different value	1	1	0
No	72	3	69

Table 6. Performance metrics of the algorithm.

Performance	PPV[a] (%)	Sensitivity (%)	Specificity (%)
Preterm labor evaluation visit	97	NE[b]	NE
Fetal fibronectin test	94.44	100	98.8
Transvaginal ultrasound	100	90	100
Cervical length	96.43	90	98.6

[a] PPV: positive predictive value.

[b] NE: not estimated.

Discussion

When pregnant women presented in triage with signs and symptoms of PTL, a PTL assessment was performed, and the details of the assessment were documented and stored in the EHR system in both structured and unstructured formats. In this study, we developed a computerized algorithm and process to identify PTL evaluation visits and extract associated methods of evaluation for threatened PTL, including fFN, TVUS, and CL. This algorithm identified the population of patients who presented with threatened PTL and underwent these associated assessments with high sensitivity and specificity. With this algorithm, 23.35% of pregnancies in the study were identified with PTL evaluation visits within $24^{0/7}$-$34^{6/7}$ gestational weeks and 19.78% of these pregnancies ultimately led to sPTB. This result is consistent with findings reported in previous studies [18,31,32].

It is worth exploring the details of misclassifications against the manual chart review, although the disagreement between manual chart review and NLP outputs was small. Of the 3 false positive PTL evaluation visits, 1 presented for a scheduled cesarean section at 36 gestational weeks; the visit mentioned uterine contractions, which was one of the conditions used to define PTL. The second case with uterine contractions presented for elective induction of labor at 39 gestational weeks and was not excluded due to the inaccurate estimation of the pregnancy start date. The third case was not excluded because the algorithm detected documented discussion of untreated infection potentially increasing the risk of PTL rather than the true PTL assessment. The algorithm produced only 1 false positive finding for fFN because it wrongly identified the phrase “fFN uninterpretable given a recent sex activity” as a positive fFN result. The algorithm missed 3 TVUS procedures, among which 2 were missed because the terms “vaginal ultrasound” and “formal ultrasound” were used to describe ultrasound for CL measurements, and these were not present in the compiled term list. The other missed case was due to the location of the imaging; TVUS was performed during the regular obstetrician office visit rather than in the hospital triage unit. Additionally, the algorithm incorrectly extracted 1 CL measure but missed 3. The CL measures of the missed cases were falsely excluded because the measures were inaccurately associated with other terms by the algorithm, such as “cervix opening/dilation” or “deepest vertical amniotic fluid pocket.” The incorrect one resulted from the false selection of a measurement performed during the obstetrician office visit rather than as part of the hospital triage service because both were mentioned in the same triage clinical notes.

Clinicians routinely conduct PTL assessments when pregnant women present with signs and symptoms of PTL. Such assessments may help distinguish true PTL cases from false ones, for which the subsequent application of appropriate interventions may improve neonatal outcomes [33]. Conversely, discharges to home for false PTL cases prevent unnecessary hospitalization, as well as unnecessary, costly, and potentially harmful interventions [34]. The current use of CL measures and fFN tests during pregnancy is limited to situations where a negative result can avoid unnecessary interventions. Our study algorithm tried to identify all PTL evaluation visits as long as the performed assessment was detected from clinical notes regardless of whether the encounter resulted in an sPTB or the continuation of the pregnancy. The identified PTL evaluation visits will provide a unique opportunity to explore the association of PTL assessment with fetal outcomes. This approach will also provide us the opportunity to accurately ascertain sPTB outcomes and its impact on successive pregnancies as well as differentiate PTBs by subtypes (sPTB from indicated PTB) in future studies.

In recent years, NLP applications have either embraced machine learning techniques alone or in combination with rule-based NLP [27,35]. Machine learning techniques proved advantageous because they improved accuracy when used in situations where the performance obtained with the existing rule-based algorithms was not satisfactory [36]. This technique has been applied in the prediction of PTBs using structured EHR data [27]. To our knowledge, this is the first NLP approach in the medical field to be used for identifying PTL evaluation visits based on either structured or unstructured data. Future work warrants further research in this area via machine learning approaches to improve the performance in terms of identifying PTL evaluation visits.

Our study has several potential limitations. First, our algorithm relied on the available (structured and unstructured) information and the accuracy of the variable in our EHR system. Although clinical notes are not available for individuals receiving outside care, it is less likely that pregnant women would receive their care elsewhere as long as a pregnancy episode was established in our care system. Second, although PTL visits with concomitant medical indications for preterm delivery were not the focus of the study (no PTL assessment performed, directly admitted for delivery), our algorithm only excluded the PTL visits with pre-eclampsia/eclampsia. Other medical conditions, such as scheduled cesarean sections and medically indicated induction of labor, were not integrated into the exclusion criteria of the algorithm due to relatively small sample sizes. Third, when applying to other health care systems and settings, this specific computerized algorithm may require some modifications due to variations in the format and presentation of clinical notes in different health care settings. Finally, this computerized algorithm was limited by the precompiled search terms and lexicons of interest in screening for potentially relevant clinical notes. It may be enhanced by more extensive and representative chart review samples in future work.

In conclusion, the developed NLP algorithm effectively identified PTL evaluation visits and extracted corresponding methods of evaluation for PTL, including fFN, TVUS, and CL measurements from the EHRs. Validation of this algorithm indicated a high level of accuracy. This NLP algorithm can be used to conduct PTL- or PTB-related pharmacoepidemiologic studies and patient care reviews.

Acknowledgments

This study was supported by Hologic, Marlborough, Massachusetts, United States. The opinions expressed are solely the responsibility of the authors and do not necessarily reflect the official views of the funding agency. The authors thank the patients of Kaiser Permanente Southern California for helping improve care using information collected through our electronic health record systems.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary tables containing the diagnosis codes, procedure, medication list, and key phrases and terms for the preterm labor evaluation visit algorithm.

[DOCX File , 37 KB - medinform_v10i9e37896_app1.docx]

References

1. Preterm birth. Centers for Disease Control and Prevention. 2020. URL: https://www.cdc.gov/reproductivehealth/maternalinfanthealth/pretermbirth.htm [accessed 2022-03-10]
2. Public Health Service, Health Resources and Services Administration. Child health USA '96-'97. DHHS publication HRSA-M-DSEA-97-48. 1997. URL: https://www.hrsa.gov/ [accessed 2022-08-30]
3. Behrman RE, Butler AS, editors. Preterm Birth: Causes, Consequences, and Prevention. Washington, DC: National Academies Press; 2007.
4. Goldenberg RL, Jobe AH. Prospects for research in reproductive health and birth outcomes. JAMA 2001 Feb;285(5):633-639. [doi: 10.1001/jama.285.5.633] [Medline: 11176872]
5. Menon R. Spontaneous preterm birth, a clinical dilemma: etiologic, pathophysiologic and genetic heterogeneities and racial disparity. Acta Obstet Gynecol Scand 2008 Jan;87(6):590-600 [FREE Full text] [doi: 10.1080/00016340802005126] [Medline: 18568457]

6. International classification of diseases. Centers for Disease Control and Prevention. 2005. URL: https://www.cdc.gov/nchs/data/dvs/Volume-1-2005.pdf [accessed 2022-03-10]
7. Ruma MS, Banker WM. Availability and use of fetal fibronectin testing and transvaginal ultrasound for preterm labor evaluation in the United States. J Matern Fetal Neonatal Med 2021 Oct:1-8. [doi: 10.1080/14767058.2021.1989403] [Medline: 34648390]
8. Roberts D, Dalziel S. Antenatal corticosteroids for accelerating fetal lung maturation for women at risk of preterm birth. Cochrane Database Syst Rev 2006 Jul;3(3):CD004454. [doi: 10.1002/14651858.CD004454.pub2] [Medline: 16856047]
9. Doyle L, Crowther C, Middleton P, Marret S. Antenatal magnesium sulfate and neurologic outcome in preterm infants: a systematic review. Obstet Gynecol 2009 Jun;113(6):1327-1333. [doi: 10.1097/AOG.0b013e3181a60495] [Medline: 19461430]
10. DeFranco EA, Lewis DF, Odibo AO. Improving the screening accuracy for preterm labor: is the combination of fetal fibronectin and cervical length in symptomatic patients a useful predictor of preterm birth? A systematic review. Am J Obstet Gynecol 2013 Mar;208(3):233.e1-233.e6. [doi: 10.1016/j.ajog.2012.12.015] [Medline: 23246314]
11. Society for Maternal-Fetal Medicine (SMFM), Hamm R, Combs C, Aghajanian P, Friedman A, Patient Safety and Quality Committee. Society for maternal-fetal medicine special statement: quality metrics for optimal timing of antenatal corticosteroid administration. Am J Obstet Gynecol 2022 Jun;226(6):B2-B10 [FREE Full text] [doi: 10.1016/j.ajog.2022.02.021] [Medline: 35189094]
12. Hezelgrave NL, Abbott DS, Radford SK, Seed PT, Girling JC, Filmer J, et al. Quantitative fetal fibronectin at 18 weeks of gestation to predict preterm birth in asymptomatic high-risk women. Obstet Gynecol 2016 Feb;127(2):255-263. [doi: 10.1097/AOG.0000000000001240] [Medline: 26942351]
13. Romero JA, Downes K, Pappas H, Elovitz MA, Levine LD. Cervical length change as a predictor of preterm birth in symptomatic patients. Am J Obstet Gynecol MFM 2021 Jan;3(1):100175. [doi: 10.1016/j.ajogmf.2020.100175] [Medline: 33451622]
14. Owen J, Yost N, Berghella V, MacPherson C, Swain M, Dildy GA, Maternal-Fetal Medicine Units Network. Can shortened midtrimester cervical length predict very early spontaneous preterm birth? Am J Obstet Gynecol 2004 Jul;191(1):298-303. [doi: 10.1016/j.ajog.2003.11.025] [Medline: 15295382]
15. Berghella V, Saccone G. Fetal fibronectin testing for reducing the risk of preterm birth. Cochrane Database Syst Rev 2019 Jul;7(7):CD006843 [FREE Full text] [doi: 10.1002/14651858.CD006843.pub3] [Medline: 31356681]
16. Preterm labor assessment toolkit. March of Dimes. 2016. URL: https://ohiohospitals.org/OHA/media/OHA-Media/Documents/Patient%20Safety%20and%20Quality/Infant%20Mortality/EED%20Webpage%20Resources/March-of-Dimes-Preterm-Toolkit.pdf [accessed 2022-03-10]
17. A practical guide for preterm labor protocol implementation. Hologic. 2019. URL: https://hologiced.com/wp-content/uploads/2019/04/MED-00342-fFN_Standard_Handbook_Digital-Version_Final_022119.pdf [accessed 2022-03-09]
18. Rose CH, McWeeney DT, Brost BC, Davies NP, Watson WJ. Cost-effective standardization of preterm labor evaluation. Am J Obstet Gynecol 2010 Sep;203(3):250.e1-250.e5. [doi: 10.1016/j.ajog.2010.06.037] [Medline: 20816147]
19. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994 Mar;1(2):161-174 [FREE Full text] [doi: 10.1136/jamia.1994.95236146] [Medline: 7719797]
20. Loper E, Bird S. NLTK: the natural language toolkit. In: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. 2002 Presented at: ETMTNLP 02; July 7, 2002; Philadelphia, PA p. 63-70. [doi: 10.3115/1118108.1118117]
21. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford CoreNLP natural language processing toolkit. 2014 Presented at: 52nd Annual Meeting of the Association for Computational Linguistics. System Demonstrations; June 23-24, 2014; Baltimore, MD p. 55-60. [doi: 10.3115/v1/p14-5010]
22. Chapman B, Chapman WW, Dayton G. Python implementation of the ConText Algorithm. chapmanbe / pyConTextNLP. URL: https://github.com/chapmanbe/pyConTextNLP/ [accessed 2022-03-10]
23. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010 Sep;17(5):507-513 [FREE Full text] [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
24. Crowley RS, Castine M, Mitchell K, Chavan G, McSherry T, Feldman M. caTIES: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research. J Am Med Inform Assoc 2010 May;17(3):253-264 [FREE Full text] [doi: 10.1136/jamia.2009.002295] [Medline: 20442142]
25. Zheng C, Yu W, Xie F, Chen W, Mercado C, Sy LS, et al. The use of natural language processing to identify Tdap-related local reactions at five health care systems in the Vaccine Safety Datalink. Int J Med Inform 2019 Jul;127:27-34 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.04.009] [Medline: 31128829]
26. Yu W, Zheng C, Xie F, Chen W, Mercado C, Sy LS, et al. The use of natural language processing to identify vaccine-related anaphylaxis at five health care systems in the Vaccine Safety Datalink. Pharmacoepidemiol Drug Saf 2020 Feb;29(2):182-188 [FREE Full text] [doi: 10.1002/pds.4919] [Medline: 31797475]

27. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform 2019 Apr;7(2):e12239 [FREE Full text] [doi: 10.2196/12239] [Medline: 31066697]
28. Włodarczyk T, Płotka S, Szczepański T, Rokita P, Sochacki-Wójcicka N, Wójcicki J, et al. Machine learning methods for preterm birth prediction: a review. Electronics 2021 Mar;10(5):586. [doi: 10.3390/electronics10050586]
29. Kiefer DG, Vintzileos AM. The utility of fetal fibronectin in the prediction and prevention of spontaneous preterm birth. Rev Obstet Gynecol 2008;1(3):106-112. [Medline: 19015761]
30. Koebnick C, Langer-Gould AM, Gould MK, Chao CR, Iyer RL, Smith N, et al. Sociodemographic characteristics of members of a large, integrated health care system: comparison with US Census Bureau data. Perm J 2012 Sep;16(3):37-41 [FREE Full text] [doi: 10.7812/TPP/12-031] [Medline: 23012597]
31. Blackwell SC, Sullivan EM, Petrilla AA, Shen X, Troeger KA, Byrne JD. Utilization of fetal fibronectin testing and pregnancy outcomes among women with symptoms of preterm labor. Clinicoecon Outcomes Res 2017 Oct;9:585-594 [FREE Full text] [doi: 10.2147/CEOR.S141061] [Medline: 29042802]
32. McPheeters ML, Miller WC, Hartmann KE, Savitz DA, Kaufman JS, Garrett JM, et al. The epidemiology of threatened preterm labor: a prospective cohort study. Am J Obstet Gynecol 2005 Apr;192(4):1325-1329-discussion 1329-1330. [doi: 10.1016/j.ajog.2004.12.055] [Medline: 15846230]
33. Sayres Jr WJ. Preterm labor. Am Fam Physician 2010 Feb;81(4):477-484 [FREE Full text] [Medline: 20148502]
34. Haas DM, Imperiale TF, Kirkpatrick PR, Klein RW, Zollinger TW, Golichowski AM. Tocolytic therapy: a meta-analysis and decision analysis. Obstet Gynecol 2009 Mar;113(3):585-594. [doi: 10.1097/AOG.0b013e318199924a] [Medline: 19300321]
35. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014 Mar;21(2):221-230 [FREE Full text] [doi: 10.1136/amiajnl-2013-001935] [Medline: 24201027]
36. Castro SM, Tseytlin E, Medvedeva O, Mitchell K, Visweswaran S, Bekhuis T, et al. Automated annotation and classification of BI-RADS assessment from radiology reports. J Biomed Inform 2017 May;69:177-187 [FREE Full text] [doi: 10.1016/j.jbi.2017.04.011] [Medline: 28428140]

Abbreviations

CL: cervical length
EHR: electronic health record
fFN: fetal fibronectin
KPSC: Kaiser Permanente Southern California
NLP: natural language processing
PPV: positive predictive value
PTB: preterm birth
PTL: preterm labor
sPTB: spontaneous preterm birth
TVUS: transvaginal ultrasound

Edited by T Hao; submitted 10.03.22; peer-reviewed by Z Reis; comments to author 06.07.22; revised version received 15.07.22; accepted 12.08.22; published 06.09.22.

Please cite as:
Xie F, Khadka N, Fassett MJ, Chiu VY, Avila CC, Shi J, Yeh M, Kawatkar A, Mensah NA, Sacks DA, Getahun D
Identification of Preterm Labor Evaluation Visits and Extraction of Cervical Length Measures from Electronic Health Records Within a Large Integrated Health Care System: Algorithm Development and Validation
JMIR Med Inform 2022;10(9):e37896
URL: https://medinform.jmir.org/2022/9/e37896
doi:10.2196/37896
PMID:36066930

Original Paper

Mining Severe Drug Hypersensitivity Reaction Cases in Pediatric Electronic Health Records: Methodology Development and Applications

Yuncui Yu[1*], MA; Qiuye Zhao[2*], PhD; Wang Cao[1], MA; Xiaochuan Wang[1], MA; Yanming Li[1], MA; Yuefeng Xie[1], MA; Xiaoling Wang[1], MA

[1]National Center for Children's Health, Beijing Children’s Hospital, Capital Medical University, Beijing, China

[2]Bohui Yishu (Beijing) Co, Ltd, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Xiaoling Wang, MA
National Center for Children's Health
Beijing Children’s Hospital
Capital Medical University
56 Nanlishi Road
Xicheng District
Beijing, 100045
China
Phone: 86 59617173
Fax: 86 59616083
Email: wangxiaoling@bch.com.cn

Abstract

Background: Severe drug hypersensitivity reactions (DHRs) refer to allergic reactions caused by drugs and usually present with severe skin rashes and internal damage as the main symptoms. Reporting of severe DHRs in hospitals now solely occurs through spontaneous reporting systems (SRSs), which clinicians in charge operate. An automatic identification system scrutinizes clinical notes and reports potential severe DHR cases.

Objective: The goal of the research was to develop an automatic identification system for mining severe DHR cases and discover more DHR cases for further study. The proposed method was applied to 9 years of data in pediatrics electronic health records (EHRs) of Beijing Children’s Hospital.

Methods: The phenotyping task was approached as a document classification problem. A DHR dataset containing tagged documents for training was prepared. Each document contains all the clinical notes generated during 1 inpatient visit in this data set. Document-level tags correspond to DHR types and a negative category. Strategies were evaluated for long document classification on the openly available National NLP Clinical Challenges 2016 smoking task. Four strategies were evaluated in this work: document truncation, hierarchy representation, efficient self-attention, and key sentence selection. In-domain and open-domain pretrained embeddings were evaluated on the DHR dataset. An automatic grid search was performed to tune statistical classifiers for the best performance over the transformed data. Inference efficiency and memory requirements of the best performing models were analyzed. The most efficient model for mining DHR cases from millions of documents in the EHR system was run.

Results: For long document classification, key sentence selection with guideline keywords achieved the best performance and was 9 times faster than hierarchy representation models for inference. The best model discovered 1155 DHR cases in Beijing Children’s Hospital EHR system. After double-checking by clinician experts, 357 cases of severe DHRs were finally identified. For the smoking challenge, our model reached the record of state-of-the-art performance (94.1% vs 94.2%).

Conclusions: The proposed method discovered 357 positive DHR cases from a large archive of EHR records, about 90% of which were missed by SRSs. SRSs reported only 36 cases during the same period. The case analysis also found more suspected drugs associated with severe DHRs in pediatrics.

(JMIR Med Inform 2022;10(9):e37812) doi:10.2196/37812

KEYWORDS

drug hypersensitivity reactions; electronic health records; clinical notes; phenotyping; natural language processing; medical language processing; bidirectional encoder representation from transformers

Introduction

Drug hypersensitivity reactions (DHRs) are one of the adverse drug reactions resembling allergy occurs. DHRs affect more than 7% of the population and are a significant cause of the postmarketing withdrawal of drugs [1]. Severe DHRs, such as anaphylactic shock, drug-induced hypersensitivity syndrome, Stevens-Johnson syndrome, and epidermolysis bullosa, have been observed worldwide with an annual incidence of 0.05 to 3 persons per million population. With mortality rates varying between 5% to 30%, severe DHRs in pediatric populations, including children, infants, and even newborns, comprise 10% to 20% of reported cases [2,3].

Reporting of severe DHRs in hospitals now solely occurs through spontaneous reporting systems (SRSs), which clinicians in charge operate. Previous studies showed that only 10% to 30% of severe adverse drug reactions were reported in SRSs [4]. Even though the missed cases were properly handled and simply not logged into the SRS system, a more thorough report would have helped improve drug guidelines. Recently, routinely collected medical data such as electronic health records (EHRs) are increasingly being used to complement the SRS and enable active pharmacovigilance. EHR systems contain detailed data with timestamps for admissions, discharges, diagnoses, medications, and laboratory tests. However, severe DHR rely on symptoms and signs for detection, which in turn often reside in the free-text areas of EHRs and require the use of natural language processing to extract information.

One of the most well-studied medical language processing applications is phenotyping (eg, the automatic evaluation of phenomics traits such as smoking status) [5]. Automatic identification of severe DHRs in patients can also be explored as a phenotyping task. When no structural data are available, the phenotyping of clinical notes can be formulated as a document classification task, which has been well studied in the natural language processing field.

Recent work [6-8] has reported that clinical documents are too long for contextualized language models to process. Our research group has integrated the medical data from a hospital and established a vertical data warehouse in its early stage. Unlike previous works that only process discharge summaries [5-7], this DHR task deals with documents consisting of all clinical notes associated with 1 inpatient visit. The average word length of discharge summaries is typically hundreds of words. However, in this DHR data set, the average word length is up to several thousand Chinese characters, and some documents contain tens of thousands of Chinese characters. Therefore, picking the best strategy for long document classification is crucial for achieving our objective.

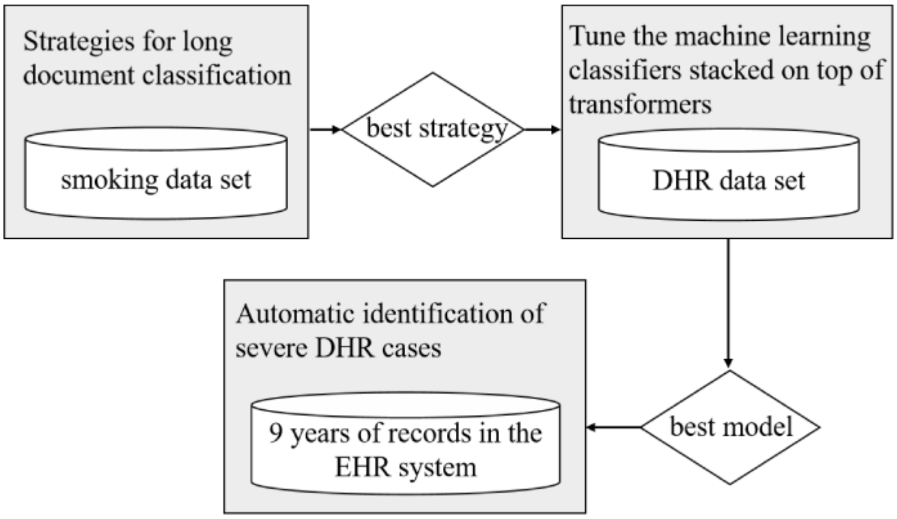
Methods

Pipeline Design

This work approaches the automatic identification of DHR cases as a long document classification problem. For training purposes, domain experts prepared a corpus containing document-level tags.

Figure 1 demonstrates the proposed system pipeline. First, 4 strategies for long document classification on the openly available smoking task were compared and evaluated. Second, the best strategy for the DHR task was applied. The pretrained embedding models of Chinese medical text on our own DHR task were compared and evaluated. A grid search to tune machine learning classifiers for the best document classification performance on the DHR data set was performed. Finally, the best pipeline to 9 years of data in a paramedic EHR was applied.

Figure 1. Proposed system pipeline in this study. DHR: drug hypersensitivity reaction; EHR: electronic health record.


Strategies for long document classification
smoking data set
best strategy
Tune the machine learning classifiers stacked on top of transformers
DHR data set
Automatic identification of severe DHR cases
9 years of records in the EHR system
best model

Ethics Approval

The study was reviewed and approved (2019-k-5) by the Institutional Ethics Committee of Beijing Children's Hospital in China, with a waiver of informed consent.

Data Set and Metrics

Smoking Task

The smoking challenge [5] automatically determines patients' smoking status from their discharge summaries. The 502 discharge summaries present 5 statuses: past smoker, current smoker, smoker, nonsmoker, and unknown. Following previous work, the class smoker was ignored. Table 1 shows the training and test data distribution.

Table 1. The training and test data distribution of the smoking task.

	Past smoker	Current smoker	Nonsmoker	Unknown	Total
Train data set	36	35	66	252	389
Test data set	11	11	16	63	101

Severe DHR Task

Data Source

Beijing Children's Hospital's information system allows for a patient's history and physician notes to be digitally recorded and instantaneously available via the network to all patient departments. A vertical data warehouse was built based on the integration of medical data in the early stage. It contains 431,972 hospitalization records of 315,608 patients from January 1, 2012, to December 31, 2020, including detailed diagnostic information, medication information, laboratory tests, disease course data, etc. Among them, a hospitalization record represents a hospitalization process. If a patient is hospitalized multiple times, the same patient will have multiple hospitalization records.

Corpus Construction

Positive cases that present severe DHRs were collected from 2 pools: the 31 positive cases logged to National Medical Products Administration reporting system and the 183 positive cases discovered by chart review. After deduplication, 200 positive cases were collected. Each positive case was assigned 1 of 4 subcategories. Furthermore, 400 negative cases were randomly sampled from Beijing Children's Hospital's EHR system. These cases were assigned a negative (NEG) tag and hand-checked by physicians to ensure they did not present severe DHRs.

The definitions of the 4 subtypes of severe DHR are shown in Multimedia Appendix 1 as found in the Guidelines for Medical Nomenclature Use of Adverse Drug Reactions issued by the Center for Drug Reevaluation of the China National Medical Products Administration in 2016 [9].

Training and Test Data Set

These 5 categories of documents were randomly sampled into the training and test data sets. The training and test data distribution is shown in Table 2. The positive and negative ratio is close to the corresponding ratio in the smoking task.

Table 2. The training and test data distribution of the severe drug hypersensitivity reaction data set.

	SJS[a]	DIHS[b]	AS[c]	EB[d]	NEG[e]	Total
Training data set	56	44	18	32	323	473
Test data set	18	3	5	7	77	110

[a] SJS: Stevens-Johnson syndrome.

[b] DIHS: drug-induced hypersensitivity syndrome.

[c] AS: anaphylactic shock.

[d] EB: epidermolysis bullosa.

[e] NEG: negative.

Evaluation Metrics

The micro-averaged F1 score was used to evaluate the performance of different models following previous study [6]. This metric is used for multiclass classification problems, measuring a balance between precision and recall and giving equal weights to each category.

Strategies for Long Document Classification

Four strategies were evaluated and compared: document truncation [10], hierarchy representation [6,11], more efficient self-attention [12], and key sentence selection [7,8,13,14]. The best strategy for long document classification was based on the openly available National NLP Clinical Challenges 2016 smoking task results [5]. The results of this task can be more fairly compared to other related works.

Document Truncation

The most straightforward way to apply a transformer model with a length limit is to truncate the input and pick the first block of tokens. These models typically require a length limit of 512 words.

More Efficient Self-Attention

Self-attention models, such as bidirectional encoder representation from transformer (BERT), require quadratic computational time and space with respect to the input sequence length. The Longformer model uses sparse self-attention instead of full self-attention to process longer documents (up to 4096 tokens).

Hierarchy Representation

In a hierarchy approach, sentence representations are built first and then aggregated into a document-level representation. In previous work on the phenotyping task of clinical notes, document representation is built by a sampling layer on top of the BERT blocks of each sentence [6].

Key Sentence Selection

A few key sentences could be enough for the document classification task. In previous works, unsupervised methods were explored to generate key sentences, which did not always perform well [13,15]. In this work, the keywords extracted from the task-specific guidelines were explored. The sentences containing keywords were selected as key sentences.

For the smoking task, unigrams and bigrams from previous work were taken as the keyword list: cigarette, smoke, smoked, smoker, smokes, smoking, tobacco [16].

For the DHR task, 2 sets of keywords were evaluated and compared. As an unsupervised method, the term frequency-inverse document frequency (TF-IDF) algorithm computed top feature words. Those containing numbers, foreign alphabets, and special characters were removed from these 2000 words. A total of 163 feature words with a score higher than zero were added to the keyword list.

The parts of the clinical notes that make references to the corresponding guidelines are most relevant for differential classification. Each positive category in the DHR data set is well defined in the corresponding guideline [17-20]. Medical terms were hand-picked from the guidelines. No domain knowledge was required to distinguish medical terms from general text. These keywords are shown in Textbox 1 in Chinese and Textbox 2 in English.

Textbox 1. The guideline keywords for the severe drug hypersensitivity reaction task in Chinese. AS: anaphylactic shock; DIHS: drug-induced hypersensitivity syndrome; EB: epidermolysis bullosa; IVIG: intravenous immunoglobulin; SJS: Stevens-Johnson syndrome; TEN: toxic epidermal necrolysis.

1. Stevens-Johnson综合征, 过敏性休克, 药物超敏反应综合征, 大疱表皮松解症, AS, EB, TEN, SJS, DIHS
2. 过敏，超敏，黏膜，红斑，松解，喘鸣，支气管痉挛，发绀，呼气流量峰值下降，肌张力减退，荨麻疹，血管性水肿，紫绀，低血容量性低血压，斑疹，斑丘疹，无菌性脓疱，紫癜，剥脱性皮炎，融合成片，松弛性水疱，表皮松解，大疱，表皮剥脱，叶状鳞屑，表皮剥离，猩红热样，麻疹样，弥漫性，黏膜侵蚀，大疱
3. 糖皮质激素，肾上腺素，甲基泼尼松龙，泼尼松，地塞米松， IVIG，甲泼尼龙

Textbox 2. The guideline keywords for the severe drug hypersensitivity reaction task in English. AS: anaphylactic shock; DIHS: drug-induced hypersensitivity syndrome; EB: epidermolysis bullosa; IVIG: intravenous immunoglobulin; SJS: Stevens-Johnson syndrome; TEN: toxic epidermal necrolysis.

1. Stevens-Johnson syndrome, anaphylactic shock, drug-induced hypersensitivity syndrome, epidermolysis bullosa, AS, EB, TEN, SJS, DIHS
2. Allergy, hypersensitivity, mucous membrane, erythema, epidermolysis, wheezing, bronchospasm, cyanosis, decreased peak expiratory flow, dystonia, urticaria, angioedema, hypovolemic hypotension, macula, maculopapular, sterile pustules, purpura, confluent, flaccid blister, bulla, exfoliative, scales, Scarlet fever–like, measles, diffuse, mucosal erosion, IVIG
3. glucocorticoid, adrenaline, prednisolone, prednisone, dexamethasone, methylpred

Data Set With Selected Text

An oracle test was conducted to evaluate whether the strategy of key sentence selection affects performance. This oracle test was performed as follows: (1) for each document that contains any keyword, assign its gold tag, and (2) for all the documents that contain no keywords, assign the UNKNOWN tag (for the smoking task) or the NEG tag (for the DHR task).

As shown in Table 3, key sentence selection reduced the maximum word count and the average word count for both data sets of the smoking task. The oracle micro-F1 was 1.0 for both the training and test set, which meant that the key sentence selection strategy did not affect the overall performance.

Two lists of keywords were evaluated for the DHR task: TF-IDF keywords and guideline keywords. As shown in Table 4, key sentence selection reduced the maximum word count and the average word count for both training and test data sets of the DHR task. The oracle test showed that with TF-IDF keywords, the oracle micro-F1 score was almost 1.0. With guideline keywords, about 2% to 3% of errors in the whole pipeline were introduced by this strategy.

Table 3. Statistics on the original and selected text in the smoking task[a].

	Maximum word count	Average word count	Oracle micro-F1
Train			
Original	3025	766	—[b]
Selected	194	18	1
Test			
Original	2529	851	—
Selected	117	18	1

[a] For word counting, all terms split by space delimiters were considered words.

[b] Not applicable.

Table 4. Statistics on the original and selected text in the severe drug hypersensitivity reaction task[a].

Keywords	Maximum average count	Average character count	Oracle micro-F1
Train			
Original	27198	4615	—[b]
Selected			
TF-IDF[c]	4681	770	0.99
Guideline	1926	199	0.98
Test			
Original	15454	3963	—
Selected			
TF-IDF	3210	687	1
Guideline	636	177	0.97

[a] For the drug hypersensitivity reaction data set, Chinese characters were counted.

[b] Not applicable.

[c] TF-IDF: term frequency-inverse document frequency.

Transformers

In-domain and open-domain pretrained embeddings by contextualized language models were evaluated in this work. For implementation, the SBERT library [10] computes document embedding with pretrained open-domain or domain-specific language models. There was no fine-tuning conducted for these pretrained models.

This work evaluated the open-domain model bert-base-uncased [21] and domain-specific models ClinicalBERT and DischargeBERT [20] for English clinical notes.

This work evaluated the open-domain model bert-base-chinese [21] and domain-specific model Medbert-kd-chinese [22] for Chinese clinical notes.

Machine Learning Classifiers

Machine learning classifiers were stacked on top of deep learning transformers. Each machine learning classifier was tuned by 10-fold cross-validation on the training data set. An automatic grid search framework [10] searched for optimal hyperparameters. This work evaluated linear models with stochastic gradient descent (SGD) learning and libsvm for support vector classification (SVC).

Results

Smoking Task: Strategies for Long Document Classification

Document Truncation

The library SBERT implemented this strategy with pretrained models BERT, ClinicalBERT, and DischargeBERT. As shown in Table 5, these models performed poorly. When long documents were straightforwardly fed into the transformers, only the first 512-word pieces were reserved.

Table 5. Phenotyping results (micro-averaged F1) of the smoking task.

Transformer	Classifier	Micro-averaged F1 (%)	
		Original text	Selected text
Longformer	SGD[a]	63.37	78.22
Bert-base-uncased	SGD	67.33	90.01
DischargeBERT	SGD	63.37[b]	91.09
ClinicalBERT	SGD	60.40	94.06

[a] SGD: stochastic gradient descent.

[b] Given the size of the data set, some models may have the same results.

More Efficient Self-Attention

The Longformer model uses sparse self-attention instead of full self-attention to process longer documents (up to 4096 tokens). However, as shown in Table 5, it did not outperform BERT baselines.

Key Sentence Selection

This work used unigrams and bigrams from Pedersen [16] to select key sentences. As shown in Table 5, each model performs better on the selected text. The domain-specific pretrained language model, ClinicalBERT (91.09%), and DischargeBERT (93.07%) outperformed the open-domain model, bert-base-uncased (90.01%).

Hierarchy Representation

In a hierarchy approach, sentence representations are built first and then aggregated into a document-level representation. For a fair comparison, we evaluated and reported the results of previous work [6] with our own evaluation script. As shown in Table 6, the f_{mean} architecture in [6] (94.2%) achieved state-of-the-art performance.

As shown in Table 6, our method (94.1%) achieved comparable performance with the top-performing method. Other earlier work for the smoking task (F1 ranged from 77.0% to 90.0%) did not achieve the same level of performance.

The strategies of key sentence selection and hierarchy representation achieve comparable performance. Furthermore, their efficiency and memory requirements were compared. As summarized in Table 7, GPU was not required for training machine learning classifiers in the proposed pipeline. The hierarchy representation model required a Tesla M40 GPU (Nvidia Corp) to train for 1 day. Our method was about 9 times faster than the hierarchy representation model for inference. With the strategies of both documentation truncation and key sentence selection, only 1 block was processed by the transformer models for each document, so the inference time was not reduced by key sentence selection.

Table 6. Phenotyping results (micro-averaged F1) of our methods and previous work[a] of the smoking task.

Transformer	Micro-averaged F1 (%)
ClinicalBERT (ours)	94.1
f_{mean} [6]	94.2
Shared task 1st place [23]	90.0
Majority label baseline [6]	81.0
CNN[b] [24]	77.0

[a] Our method and f_{mean} were evaluated by the same script over the test data set. Other results were found directly from their published reports. For comparison, the precision of the results is 0.1%.

[b] CNN: convolutional neural networks.

Table 7. Runtime and memory requirements of each model. The training time and GPU requirement of f_{mean} are taken from previous work [6]. The inference time on the test data set was evaluated on a GPU server with NVIDIA T4 and 4*cpu (Nvidia Corp).

Model	Documents	Inference time on test data set (seconds)	Training time (hours)	GPU memory
f_{mean} [6]	text	35.52	24	16
ClinicalBert	text	0.46	—[a]	—
+MLClassifier	selected text	0.437	1	—

[a] Not applicable.

Severe DHR Task: Stacked Transformers and Classifiers

The smoking task showed that key sentence selection improved self-attention transformers with length limits. In the DHR task, this strategy was evaluated with various transformers and classifiers. As discussed in Methods, 2 kinds of keywords were evaluated and compared. As an unsupervised method, top TF-IDF [8] feature words were used for key sentence selection. Considering that clinical notes comply with guidelines, keywords were drawn from the DHR guidelines.

As shown in Table 8, the guideline keywords always improved the performance, regardless of the stacked transformers and classifiers. The TF-IDF keywords only help with the SVC classifier.

Table 8. Phenotyping results (micro-averaged F1) of different transformers for the severe drug hypersensitivity reaction task.

Transformers and classifiers	Micro-averaged F1(%)		
	Original text	Selected text	
		TF-IDF[a]	guidelines
Bert-base-chinese			
SVC[b]	80.91	82.73	87.27
SGD[c]	80.00	77.27	86.36
Medbert-kd-chinese			
SVC	81.82	83.64	89.09
SGD	82.73	73.64	87.27

[a] TF-IDF: term frequency-inverse document frequency.

[b] SVC: support vector classification.

[c] SGD: stochastic gradient descent.

Applications in a 9-Year EHR

Finally, the best configuration was applied to the 9 years of data in Beijing Children's Hospital's EHRs. A total of 1155 cases were alerted. After double-checking by 2 clinicians and 2 pharmacists in pediatrics based on the criterion of severe DHRs, 357 cases of severe DHRs in children were found (Table 9): anaphylactic shock (n=39), drug-induced hypersensitivity syndrome (n=178), Stevens-Johnson syndrome (n=86), and epidermolysis bullosa (n=54). Only 36 of 356 severe DHRs had been reported to SRS before. About 89.89% of cases were underreported, resulting in insufficient attention from drug regulators and clinicians. This suggests that our method could actively identify severe DHRs providing additional evidence for pharmacovigilance in children.

The case analysis indicated many suspected drugs that may cause severe DHRs in pediatrics. The suspected drugs leading to anaphylactic shock mainly included pegaspargase injection, L-asparaginase, cefoperazone sulbactam, etc. Phenobarbital, nimesulide, and cephalosporin antibiotics were the key suspected drugs leading to drug-induced hypersensitivity syndrome and Stevens-Johnson syndrome. In addition, lamotrigine, lysine acetylsalicylate, and meropenem were closely related to the occurrence of epidermolysis bullosa.

Table 9. Distribution of the severe drug hypersensitivity reactions cases in 9 years of electronic health records found by the proposed pipeline.

Severe DHR[a]	Reported in SRS[b] of BCH[c], n	DHR cases confirmed by experts, (n)		
		Diagnosed in BCH	Diagnosed in other hospitals	Total
AS[d]	4	26	13	39
DIHS[e]	16	29	149	178
SJS[f]	7	9	77	86
EB[g]	9	8	46	54
Total	36	72	285	357

[a] DHR: drug hypersensitivity reaction.

[b] SRS: spontaneous reporting system.

[c] BCH: Beijing Children's Hospital.

[d] AS: anaphylactic shock.

[e] DIHS: drug-induced hypersensitivity syndrome.

[f] SJS: Stevens-Johnson syndrome.

[g] EB: epidermolysis bullosa.

Discussion

Principal Findings

The results showed that clinical documents were too long to perform document classification baselines. Among the 4 strategies of long document classification, hierarchy representation and key sentence selection were best performed on the smoking task. Moreover, key sentence selection was 9 times faster than hierarchy representation models for inference. The keywords extracted from task-specific guidelines performed better than the unsupervised method. Domain-specific language models always performed better than general embeddings.

A total of 1155 cases were alerted, among which clinicians and pharmacists identified 357 cases of severe DHRs in children. Only 36 of these cases have been reported by SRS. This result suggested that the reporting rate of SRS may be as low as 10.08%. The automatic pipeline that scrutinized clinical notes and reported potential severe DHR cases can help decrease the number of missed positive DHR cases and reduce the cost of labor at the same time.

The case analysis also found more suspected drugs associated with severe DHRs in pediatrics. The analysis could help promote postmarketing drug risk assessment conducive to rational drug use and improve drug guidelines.

Comparison With Prior Work

Our method achieved comparable performance for the smoking task with the top-performing method (94.1% vs 94.2%). For the DHR task, our method discovered 357 positive cases, about 90% of which were missed by SRS.

Recent work has studied that clinical documents are too long for contextualized language models to process [6-8]. Unlike previous works that only process discharge summaries [5-7], this DHR task deals with documents consisting of all clinical notes associated with 1 inpatient visit. The average word length of discharge summaries is typically hundreds of words. However, in the DHR data set, the average word length is up to several thousand Chinese characters, and some documents contain tens of thousands of Chinese characters.

This work has 4 strategies evaluated and compared: document truncation [10], hierarchy representation [6,11], more efficient self-attention [12], and key sentence selection [7,8,13,14]. None of these works considered the use of guidelines.

Limitations

The proposed method required the annotation of about 200 positive cases for supervised training. When applying to the large archive of EHRs in hospital databases, certain preprocessing steps are still required to prevent malfunctions from badly formatted documents. Such preprocessing steps may vary for each hospital's system.

Conclusions

Automatic identification of severe DHRs can be approached as a document classification problem. The best strategy for long document classification of clinical notes is key sentence selection with task-specific guidelines. The reporting of DHR cases cannot only rely on clinicians in charge. In the same period of data, the SRS system reported 36 cases, whereas the automatic process discovered 357 cases. The case analysis also found more suspected drugs associated with severe DHRs in pediatrics.

Acknowledgments

This work was supported by grant CST2020CT108 from the Clinical Toxicology Program of Chinese Society of Toxicology, grant DSM2021004 from the Post-marketing Drug Risk Assessment Program of China Society for Drug Regulation, and grant CNHDRC-KJ-W-2021-58 from the Clinical Technology Training Program for Comprehensive Evaluation of Pediatric Medication of China National Health and Development Research Center. The funder had no role in conducting the study; collection, management, analysis, and interpretation of data; preparation, review, and approval of the manuscript; or decision to submit the manuscript for publication.

Authors' Contributions

XLW undertook work of framework design and overall guidance of whole research. YCY, XCW, WC, YML, and YFX took responsibility for the data collection. YCY and QYZ performed the data processing and article writing. QYZ and XLW provided data interpretation and methodological advice.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Types of drug hypersensitivity reactions and criteria.
[DOCX File , 15 KB - medinform_v10i9e37812_app1.docx]

References

1. Naisbitt DJ. Drug hypersensitivity reactions in skin: understanding mechanisms and the development of diagnostic and predictive tests. Toxicology 2004 Jan 15;194(3):179-196. [doi: 10.1016/j.tox.2003.09.004] [Medline: 14687965]

2. Gomes ER, Brockow K, Kuyucu S, Saretta F, Mori F, Blanca-Lopez N, ENDA/EAACI Drug Allergy Interest Group. Drug hypersensitivity in children: report from the pediatric task force of the EAACI Drug Allergy Interest Group. Allergy 2016 Feb;71(2):149-161. [doi: 10.1111/all.12774] [Medline: 26416157]
3. Rukasin CRF, Norton AE, Broyles AD. Pediatric Drug Hypersensitivity. Curr Allergy Asthma Rep 2019 Feb 22;19(2):11. [doi: 10.1007/s11882-019-0841-y] [Medline: 30793223]
4. Lopez-Gonzalez E, Herdeiro MT, Figueiras A. Determinants of under-reporting of adverse drug reactions: a systematic review. Drug Saf 2009;32(1):19-31. [doi: 10.2165/00002018-200932010-00002] [Medline: 19132802]
5. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc 2008;15(1):14-24 [FREE Full text] [doi: 10.1197/jamia.M2408] [Medline: 17947624]
6. Andriy M, Elliot S, Masoud R, Mark D. Phenotyping of clinical notes with improved document classification models using contextualized neural language models. ArXiv. Preprint posted online on October 30, 2019 2019:1 [FREE Full text]
7. Huang K, Garapati S, Rich A. An interpretable end-to-end fine-tuning approach for long clinical text. ArXiv. Preprint posted online on November 12, 2020 2020:1 [FREE Full text]
8. Valmianski I, Goodwin C, Finn I. Evaluating robustness of language models for chief complaint extraction from patient-generated text. ArXiv. Preprint posted online on November 15, 2019 2019:1 [FREE Full text]
9. Guidelines for Medical Nomenclature Use of Adverse Drug Reactions. Beijing: National Medical Products Administration; 2016.
10. Reimers NG. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. ArXiv. Preprint posted online on August 27, 2019 2019:1 [FREE Full text] [doi: 10.18653/v1/d19-1410]
11. Pappagari RZ, Villalba J, Carmiel Y, Dehak N. Hierarchical transformers for long document classification. ArXiv. Preprint posted online on October 23, 2019 2019:1 [FREE Full text] [doi: 10.1109/asru46091.2019.9003958]
12. Beltagy IP, Cohan A. Longformer: the long-document transformer. ArXiv. Preprint posted online on April 10, 2020 2020:1 [FREE Full text]
13. Ding MZ, Yang H, Tang J. Cogltx: applying bert to long texts. Adv Neural Inf Process Syst 33. URL: https://proceedings.neurips.cc/paper/2020/file/96671501524948bc3937b4b30d0e57b9-Paper.pdf [accessed 2022-08-18]
14. Fiok K, Karwowski W, Gutierrez-Franco E, Davahli MR, Wilamowski M, Ahram T, et al. Text guide: improving the quality of long text classification by a text selection method based on feature importance. IEEE Access 2021;9:105439-105450. [doi: 10.1109/access.2021.3099758]
15. Park H, Vyas Y, Shah K. Efficient classification of long documents using transformers. ArXiv. Preprint posted online on March 21, 2022 2021:1 [FREE Full text]
16. Pedersen T. Determining smoker status using supervised and unsupervised learning with lexical features. URL: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.116.1948&rep=rep1&type=pdf [accessed 2022-08-18]
17. Li X, Zhai S, Wang Q, Wang Y, Yin J, Chen Y. Recommendations in guideline for emergency management of anaphylaxis. Adverse Drug React J 2019;21(2):85-91. [doi: 10.1201/9780429083129-12]
18. Allergic Diseases Committee. Expert consensus on diagnosis and treatment of drug hypersensitivity syndrome. Chin J Dermatol 2018;51(11):787-790. [doi: 10.3760/cma.j.issn.0412-4030.2018.11.002]
19. Adverse Drug Reaction Research Center of Chinese Society of Dermatology. Expert consensus on the diagnosis and treatment of Stevens-Johnson syndrome/toxic epidermal necrolysis. Chin J Dermatol 2021 May 15;54(5):376-381. [doi: 10.35541/cjd.20201177]
20. Alsentzer E, Murphy J, Boag W, Weng W, Jin D, Naumann T. Publicly Available Clinical BERT Embeddings. ArXiv. Preprint posted online on April 6, 2019 2019:1 [FREE Full text] [doi: 10.18653/v1/w19-1909]
21. Turc I, Chang M, Lee K, Kristina T. Well-read students learn better: on the importance of pre-training compact models. ArXiv. Preprint posted online on August 23, 2019 2019:1 [FREE Full text]
22. trueto: research and application of BERT model in Chinese clinical Natural language processing. 2021. URL: https://github.com/trueto/medbert [accessed 2021-03-01]
23. Clark C, Good K, Jezierny L, Macpherson M, Wilson B, Chajewska U. Identifying smokers with a medical extraction system. J Am Med Inform Assoc 2008;15(1):36-39 [FREE Full text] [doi: 10.1197/jamia.M2442] [Medline: 17947619]
24. Wang Y, Sohn S, Liu S, Shen F, Wang L, Atkinson EJ, et al. A clinical text classification paradigm using weak supervision and deep representation. BMC Med Inform Decis Mak 2019 Jan 07;19(1):1 [FREE Full text] [doi: 10.1186/s12911-018-0723-6] [Medline: 30616584]

Abbreviations

BERT: bidirectional encoder representation from transformer
DHR: drug hypersensitivity reaction
EHR: electronic health record
NEG: negative
SGD: stochastic gradient descent
SRS: spontaneous reporting system

SVC: support vector classification
TF-IDF: term frequency-inverse document frequency

Edited by T Hao; submitted 08.03.22; peer-reviewed by J Luo, L Chen; comments to author 04.05.22; revised version received 06.07.22; accepted 12.08.22; published 13.09.22.

Please cite as:
Yu Y, Zhao Q, Cao W, Wang X, Li Y, Xie Y, Wang X
Mining Severe Drug Hypersensitivity Reaction Cases in Pediatric Electronic Health Records: Methodology Development and Applications
JMIR Med Inform 2022;10(9):e37812
URL: https://medinform.jmir.org/2022/9/e37812
doi:10.2196/37812
PMID:36099001

Original Paper

One Clinician Is All You Need–Cardiac Magnetic Resonance Imaging Measurement Extraction: Deep Learning Algorithm Development

Pulkit Singh[1], BA; Julian Haimovich[2,3,4], MD; Christopher Reeder[1], PhD; Shaan Khurshid[2,3,5], MPH, MD; Emily S Lau[3,4], MD; Jonathan W Cunningham[4,6], MD; Anthony Philippakis[1,7], MD, PhD; Christopher D Anderson[8,9,10], MMSc, MD; Jennifer E Ho[4,11], MD; Steven A Lubitz[2,3,4,5], MPH, MD; Puneet Batra[1], PhD

[1]Data Sciences Platform, The Broad Institute of Harvard and MIT, Cambridge, MA, United States
[2]Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States
[3]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, United States
[4]Cardiovascular Disease Initiative, The Broad Institute of Harvard and MIT, Cambridge, MA, United States
[5]Demoulas Center for Cardiac Arrhythmias, Massachusetts General Hospital, Boston, MA, United States
[6]Division of Cardiology, Brigham and Women’s Hospital, Boston, MA, United States
[7]Eric and Wendy Schmidt Center, The Broad Institute of Harvard and MIT, Cambridge, MA, United States
[8]Department of Neurology, Brigham and Women’s Hospital, Boston, MA, United States
[9]Henry and Allison McCance Center for Brain Health, Massachusetts General Hospital, Boston, MA, United States
[10]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, United States
[11]CardioVascular Institute and Division of Cardiology, Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, United States

Corresponding Author:
Puneet Batra, PhD
Data Sciences Platform
The Broad Institute of Harvard and MIT
415 Main Street
Cambridge, MA, 02142
United States
Phone: 1 617 714 7000
Email: gpbatra@gmail.com

Abstract

Background: Cardiac magnetic resonance imaging (CMR) is a powerful diagnostic modality that provides detailed quantitative assessment of cardiac anatomy and function. Automated extraction of CMR measurements from clinical reports that are typically stored as unstructured text in electronic health record systems would facilitate their use in research. Existing machine learning approaches either rely on large quantities of expert annotation or require the development of engineered rules that are time-consuming and are specific to the setting in which they were developed.

Objective: We hypothesize that the use of pretrained transformer-based language models may enable label-efficient numerical extraction from clinical text without the need for heuristics or large quantities of expert annotations. Here, we fine-tuned pretrained transformer-based language models on a small quantity of CMR annotations to extract 21 CMR measurements. We assessed the effect of clinical pretraining to reduce labeling needs and explored alternative representations of numerical inputs to improve performance.

Methods: Our study sample comprised 99,252 patients that received longitudinal cardiology care in a multi-institutional health care system. There were 12,720 available CMR reports from 9280 patients. We adapted PRAnCER (Platform Enabling Rapid Annotation for Clinical Entity Recognition), an annotation tool for clinical text, to collect annotations from a study clinician on 370 reports. We experimented with 5 different representations of numerical quantities and several model weight initializations. We evaluated extraction performance using macroaveraged F_1-scores across the measurements of interest. We applied the best-performing model to extract measurements from the remaining CMR reports in the study sample and evaluated established associations between selected extracted measures with clinical outcomes to demonstrate validity.

Results: All combinations of weight initializations and numerical representations obtained excellent performance on the gold-standard test set, suggesting that transformer models fine-tuned on a small set of annotations can effectively extract numerical quantities. Our results further indicate that custom numerical representations did not appear to have a significant impact on extraction performance. The best-performing model achieved a macroaveraged F_1-score of 0.957 across the evaluated CMR measurements (range 0.92 for the lowest-performing measure of left atrial anterior-posterior dimension to 1.0 for the highest-performing measures of left ventricular end systolic volume index and left ventricular end systolic diameter). Application of the best-performing model to the study cohort yielded 136,407 measurements from all available reports in the study sample. We observed expected associations between extracted left ventricular mass index, left ventricular ejection fraction, and right ventricular ejection fraction with clinical outcomes like atrial fibrillation, heart failure, and mortality.

Conclusions: This study demonstrated that a domain-agnostic pretrained transformer model is able to effectively extract quantitative clinical measurements from diagnostic reports with a relatively small number of gold-standard annotations. The proposed workflow may serve as a roadmap for other quantitative entity extraction.

(JMIR Med Inform 2022;10(9):e38178) doi:10.2196/38178

KEYWORDS

natural language processing; transformers; machine learning; cardiac MRI; clinical outcomes; deep learning

Introduction

Cardiac magnetic resonance imaging (CMR) facilitates the characterization of many important cardiac diseases including left and right ventricular failure, left ventricular hypertrophy, and aortic root aneurysms. Quantification of left ventricular ejection fraction (LVEF) and classification of patients with heart failure into those with reduced, moderately reduced, or preserved ejection fraction is the cornerstone of selecting appropriate therapies for a given patient [1]. CMR also quantifies right ventricular function and is notably the only noninvasive diagnostic modality able to fully evaluate the right ventricle [2]. Anatomic information from CMR is also diagnostic of other important cardiac diseases, including left ventricular hypertrophy, which is an important marker for overall cardiac health, and thoracic aortic root aneurysms [3]. CMR measurements, in addition to other diagnostic information, are embedded in narrative clinical text. In many electronic health record (EHR) systems, these measurements are unavailable in easily accessible harmonized structured formats. The development of tools to automatically extract quantitative measurements from unstructured CMR reports would facilitate their use in research, including as inputs to machine learning models.

Existing approaches for extracting measurements from clinical text are often based on manually developed heuristics or machine learning methods that learn from labeled data but do not leverage pretrained language representations. Rule-based approaches [4], while computationally efficient, require substantial manual effort to construct and can suffer performance degradation with shifts in linguistic structure of reports [5]. Other work has used machine learning approaches such as support vector machines and long short-term memory models to extract measurements from clinical notes, but these approaches have required large quantities of expert annotations due to absence of pretraining [6]. In addition, prior methods for clinical measurement extraction rely on considerable data-specific preprocessing, which may not translate well to EHRs outside of where the heuristics were developed [7].

Transformer-based neural networks like Bidirectional Encoder Representations from Transformers (BERT) [8,9] have achieved state-of-the-art results across a wide variety of natural language processing (NLP) tasks [10]. These models are pretrained on large amounts of text to learn general linguistic structure and produce contextualized representations of language. The advantage of this pretraining paradigm is that these networks can be fine-tuned using minimal problem-specific labels to achieve state-of-the-art performance on many natural language tasks. BERT was originally pretrained on general domain text such as Wikipedia but has since been adapted for use in clinical applications by pretraining on domain-specific text [11-14]. Although transformer-based models have shown efficacy in extracting nonnumerical entities such as anatomical terms and disease states from clinical text [14], their application to extracting numerical quantities from clinical text has been limited [15,16].

In this study, we hypothesized that pretrained transformers fine-tuned on a small set of annotations can efficiently extract numerical quantities from diagnostic text. We fine-tuned a range of pretrained transformers, including clinically oriented ones, to develop an NLP workflow that simultaneously extracts 21 specific measurements of cardiac structure and function from CMR reports in a cardiology-based EHR cohort. This set represents all clinically meaningful quantitative imaging findings available in the CMR reports. We also explored whether alternative numerical representations impact extraction quality compared to the default representations that appear in reports. After selecting the best-performing model, we applied our workflow to extract measurements from all available CMR reports in the study cohort. To demonstrate the accuracy of these extractions, we assessed the expected associations between extracted cardiac anatomy and function indices and incident clinical outcomes.

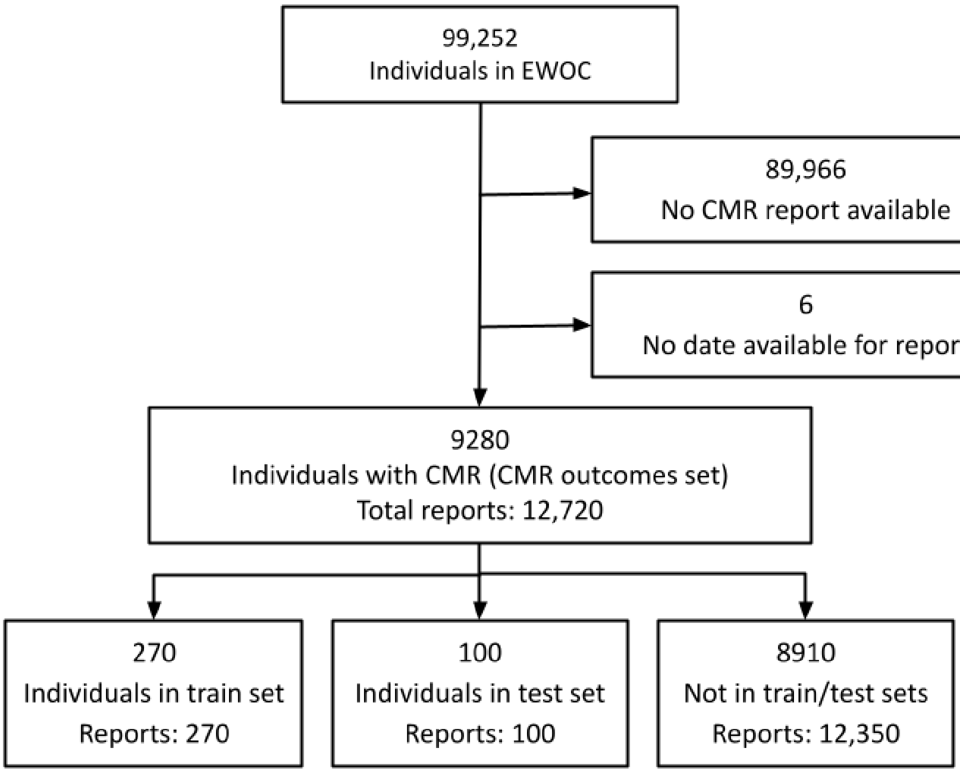
Methods

Study Sample

Individuals were selected from a retrospective community-based ambulatory cardiology sample (Enterprise Warehouse of

Cardiology [EWOC]) in a multi-institutional academic health care system (Mass General Brigham). EWOC comprises 99,252 adults aged 18 years or older with ≥2 cardiology clinic visits within 1 to 3 years between 2000 and 2019. A broad range of EHR data are available for each individual in the cohort, including demographics, anthropometrics, vital signs, narrative notes, laboratory results, medication lists, radiology and cardiology diagnostic test results, pathology reports, and procedural and diagnostic administrative billing codes [16]. These data were processed using the JEDI Extractive Data Infrastructure [17]. After excluding 6 individuals and reports that had no CMR date available, 12,720 CMR reports were available for 9280 individuals in EWOC (Figure 1).

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram for study sample. CMR: cardiac magnetic resonance imaging; EWOC: Enterprise Warehouse of Cardiology.



Ethics Approval

This research was approved by the Massachusetts General Brigham Institutional Review Board (2017P001650).

Clinical Feature Ascertainment

Baseline characteristics were defined using previously published groupings of International Classification of Diseases, 9th and 10th revision diagnosis codes [16]. Definitions for clinical features used in the analysis are provided in Table S1 in Multimedia Appendix 1. Baseline characteristics of individuals in the modeling sample were ascertained prior to the date of the CMR (Table 1).

Table 1. Baseline characteristics of training the set, test set, and CMR outcomes set.

	Training set (N=278)	Test set (N=100)	CMR[a] outcomes set[b] (N=9280)
Age (years), median (Q1, Q3)	54 (46, 64)	58 (45, 66)	57 (46, 67)
Female sex, n (%)	95 (34.2)	33 (33)	3666 (39.5)
Diabetes mellitus, n (%)	23 (8.3)	10 (10)	1216 (13.1)
Coronary artery disease, n (%)	69 (24.8)	31 (31)	3406 (36.7)
Myocardial infarction, n (%)	42 (15.1)	15 (15)	1791 (19.3)
Atrial fibrillation, n (%)	104 (37.4)	24 (24)	3164 (34.1)
Obesity, n (%)	12 (4.3)	7 (7)	631 (6.8)
Chronic kidney disease, n (%)	26 (9.4)	7 (7)	1123 (12.1)
Hypertension, n (%)	130 (46.8)	55 (55)	5563 (59.9)
Ethnicity, n (%)			
White	237 (85.3)	93 (93)	7814 (84.2)
Asian	14 (5.0)	1 (1)	251 (2.7)
Black	13 (4.7)	2 (2)	520 (5.6)
Other	7 (2.5)	1 (1)	195 (2.1)
Hispanic	4 (1.4)	0 (0)	111 (1.2)
Unknown	3 (1.1)	3 (3)	390 (4.2)

[a] CMR: cardiac magnetic resonance imaging.

[b] Includes all individuals in Enterprise Warehouse of Cardiology with a CMR report.

CMR Labeling

Similar to other EHRs, quantitative CMR measurements are contained in free-text diagnostic reports in the Mass General Brigham EHR [14,18]. We leveraged PRAnCER (Platform Enabling Rapid Annotation for Clinical Entity Recognition) [19], an open-source software application for intuitive labeling, to annotate 21 clinically important measurements from EWOC CMR reports (Textbox 1). We adapted PRAnCER to work with a custom schema containing CMR features rather than the Unified Medical Language System vocabulary [20] for which it was designed. There is significant variability in the format and context of measurement instances. This includes the ordering of measurements in the report, the language used to reference a particular measurement, the presence or absence of units, and the positional relationship between a measurement name and the value itself (Figure 2).

Of all available reports, 370 were randomly selected from unique individuals for annotation by a study clinician (JSH). From these reports, 270 were randomly partitioned into a training set while the remaining 100 were reserved for model testing (Figure 1). No individuals appeared in both the training and test sets. As CMR protocols may vary based on the clinical indication for the study, the total number of measurements per report ranged from 1 to 21. The counts of each unique feature across the training and test sets are available in Table S2 in Multimedia Appendix 1. Total clinician labeling time for all 370 reports was estimated at 15 hours.

Finally, to address the quality of clinical annotations, we employed a secondary annotator (PB) to label only the 100 reports reserved for model testing. We computed interannotator agreement as the proportion of matched extractions between annotators, in line with clinical entity extraction literature [15]. Overall agreement was excellent at 91.6%, and measurementwise agreement values are available in Table S3 in Multimedia Appendix 1**.** Given the nature of the annotation task, there was perfect precision when both annotators picked out a measurement from a report, and any disagreement represents values missed due to fatigue or difference in guidelines. Given the high agreement, we performed model derivation and validation on annotations from the study clinician (JSH) only.

Textbox 1. Clinical measurements extracted from cardiac magnetic resonance imaging reports.

Left ventricle anatomy and function

- Left ventricular end diastolic volume
- Left ventricular end diastolic volume index
- Left ventricular end diastolic diameter
- Left ventricular end systolic volume
- Left ventricular end systolic volume index
- Left ventricular end systolic diameter
- Left ventricular ejection fraction
- Left ventricular stroke volume
- Left ventricular mass
- Left ventricular mass index
- Cardiac output
- Cardiac index

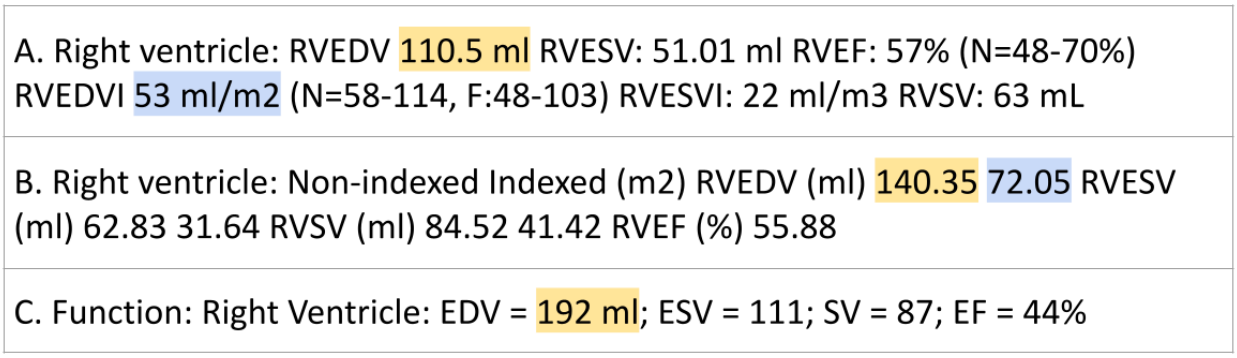
Right ventricle anatomy and function

- Right ventricular end diastolic volume
- Right ventricular end diastolic volume index
- Right ventricular end systolic volume
- Right ventricular end systolic volume index
- Right ventricular stroke volume
- Right ventricular stroke volume

Other cardiac structural anatomy

- Left atrial anterior-posterior dimension
- Pulmonary artery dimension
- Aortic root dimension

Figure 2. Example text from 3 cardiac magnetic resonance imaging reports (A,B,C) quantifying right ventricular function. The lack of consistency in how equivalent measurements are presented makes accurately extracting measurements challenging. Yellow highlighted features indicate right ventricular end diastolic volume (RVEDV), whereas blue highlighted features indicate right ventricular end diastolic volume index (RVEDVI). Example C does not contain the RVEDVI feature. EDV: end diastolic volume; EF: ejection fraction; ESV; end systolic volume; RVEF: right ventricular ejection fraction; RVESV: right ventricular end systolic volume; RVESVI: right ventricular end systolic volume index; RVSV: right ventricular stroke volume.



A. Right ventricle: RVEDV 110.5 ml RVESV: 51.01 ml RVEF: 57% (N=48-70%) RVEDVI 53 ml/m2 (N=58-114, F:48-103) RVESVI: 22 ml/m3 RVSV: 63 mL

B. Right ventricle: Non-indexed Indexed (m2) RVEDV (ml) 140.35 72.05 RVESV (ml) 62.83 31.64 RVSV (ml) 84.52 41.42 RVEF (%) 55.88

C. Function: Right Ventricle: EDV = 192 ml; ESV = 111; SV = 87; EF = 44%

Numerical Representations

Previous work has shown that the use of alternative representations in place of default surface representations of numbers has a significant impact on a transformer model's ability to perform quantitative manipulations within text, such as simple arithmetic [21]. The vocabularies of most transformer-based models include a limited number of numerical values and generally no decimal numbers since they are constructed from the most frequently occurring words in the corpus used for pretraining. The tokenization procedure employed by most transformer models separates "words" based on punctuation and does not distinguish between periods and decimal places, which results in decimal numbers being broken up into multiple tokens. Given the potential limitations of default numerical representations, we investigated whether implementing alternative numerical representations impacts the extraction quality of quantitative clinical measures. We designed 4 different types of numerical transformations for quantitative tokens in the CMR reports, which were applied to both the training and test samples for model derivation. These included replacing decimal points with a special token to ensure that decimal numbers stay intact during tokenization, a consistent number of digits for all values, scientific notation, and converting quantities to words. Table 2 demonstrates these transformations for 1 snippet of text, and Multimedia Appendix 1 contains more information about their implementations.

Table 2. Numerical transformations for an example snippet of text.

Transformation name	Transformed snippet	Notes
Original	RVESV[a]: 51.01 ml	No transformation; for reference
Replaced decimal	RVESV: 51|01 ml	Decimal points replaced with special separator character; enables parsing as a single token rather than being broken up
Consistent digits	RVESV: 051010 ml	All numbers converted to be 6 digits in length
Scientific notation	RVESV: 5.10100e+01	All numbers converted to scientific notation, with 5 significant digits
Words	RVESV: fifty one point zero one ml	Number converted to corresponding word representation

[a]RVESV: right ventricular end systolic volume.

Model Derivation and Validation

Our modeling approach involved fine-tuning transformer-based models using the HuggingFace transformers library [22] to predict a label for each token in a given CMR report. To do so, we attached a linear classification head on top of the last layer of a BERT architecture. The classification head produces a distribution over 22 possible labels—the 21 cardiac measurements of interest plus a “0” label for all other tokens (Figure 3). We preprocessed report text into sections containing 128 tokens, accounting for subword tokenization, in accordance with input size limitations of the transformer-based models. We used cross-entropy loss with a learning rate of 5e^{-5} and a batch size of 32 across all experiments. To evaluate the impact of clinical pretraining on numerical clinical value extraction, we experimented with initializing the weights of the BERT architecture with the weights provided by BERT$_{\text{LARGE}}$ [8,9] cased (~340 million parameters) as well as the clinically oriented weights of PubMedBERT [11], SapBERT [12], and Bio+DischargeSummaryBERT [13] (each with ~110 million parameters). Pretrained weights were downloaded from the HuggingFace model hub [23]. Each pretrained architecture was paired with the 5 numerical representations.

Each model was fine-tuned on the Center for Clinical Data Science computational cluster hosted by Mass General Brigham. On a graphic processing unit–equipped machine, each model trained at a rate of approximately 2 minutes per epoch. Each combination of weight initialization and numerical representation strategy was fine-tuned for 20 epochs, requiring an average of 40 minutes. For the purpose of model evaluation, we assigned a label to a token if the predicted score for that label was greater than 0.5. Performance was evaluated using the macroaveraged F_1-score over all 21 measurements of interest, as this metric captures featurewise performance regardless of the frequency of occurrence in the reports. For each model, we selected the number of epochs that maximized the macroaveraged F_1-score.

Minimal postprocessing was applied based on the results of the labels assigned by our modeling experiments. This included merging with additional significant digits that should obviously be included as part of a measurement and the consolidation of model-predicted tokens into a structured format (Multimedia Appendix 1**)**. Finally, we applied upper and lower bounds on extracted values using reference ranges derived from the CMR literature [24-26] (Table S4, Multimedia Appendix 1). An overview of the workflow, including collecting clinical annotations, modeling, and postprocessing to extract final measurements is provided in Figure 4.

Figure 3. Architecture for fine-tuning pretrained transformer architecture with gold-standard cardiac resonance imaging annotations and predicting labels for each token. BERT: Bidirectional Encoder Representations from Transformers; ESV: end systolic volume.

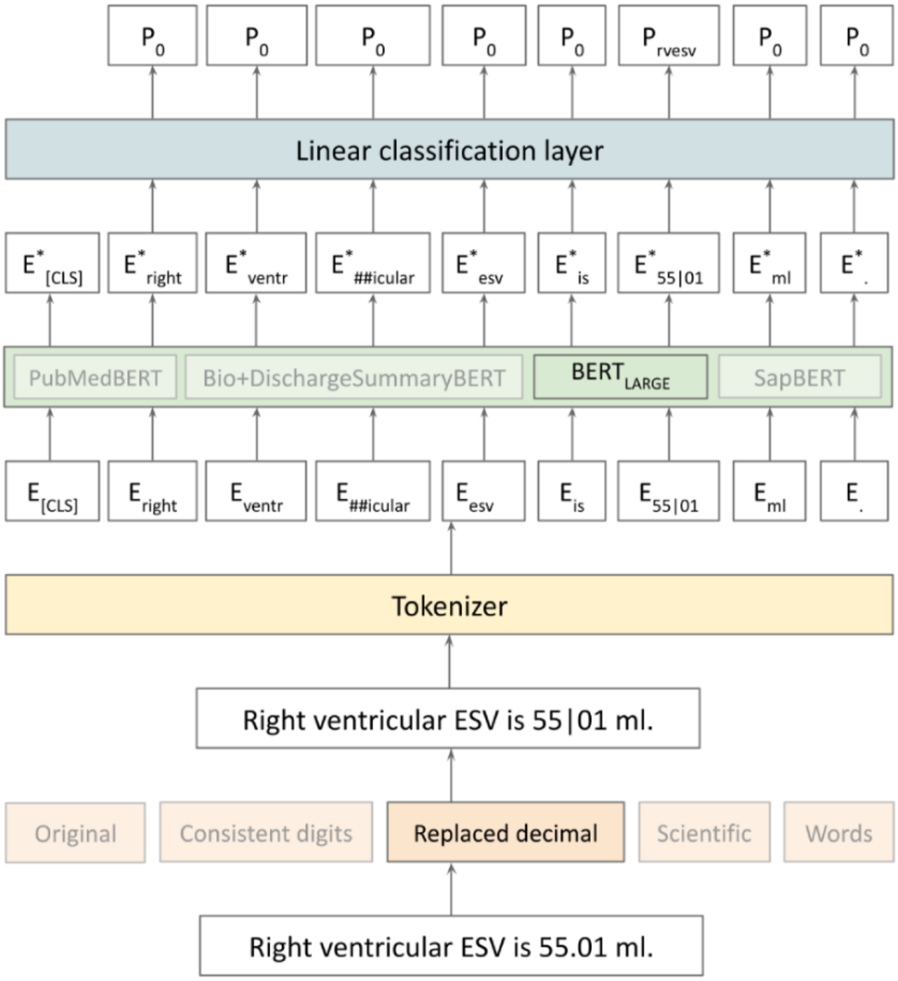
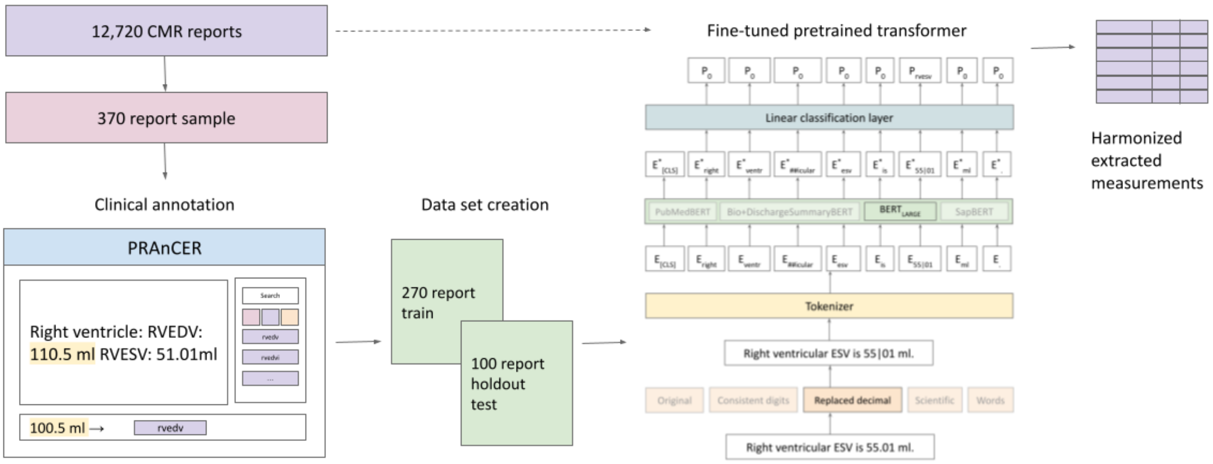


Figure 4. Natural language processing workflow for collecting clinical annotations, modeling, and extracting measurements from cardiac magnetic resonance imaging reports. BERT: Bidirectional Encoder Representations from Transformers; ESV: end systolic volume; CMR: cardiac magnetic resonance imaging; PRAnCER: Platform Enabling Rapid Annotation for Clinical Entity Recognition; RVEDV: right ventricular end diastolic volume; RVESV: right ventricular end systolic volume.



Associations With Clinical Outcomes

Finally, to assess the clinical validity of model extractions, we evaluated whether selected extracted features demonstrated known relationships with clinical outcomes, including mortality, atrial fibrillation, and heart failure [27-29]. We first applied the highest performing model to extract left ventricular mass index (LVMI), LVEF, and right ventricular ejection fraction (RVEF) from all CMR reports in EWOC. Rather than choose a model score threshold for each label, we chose the label with the highest score for each token. For individuals with multiple reports containing a given feature, we used features extracted from the earliest report for the primary analysis.

We then assessed incidence rates of mortality, atrial fibrillation, and heart failure by quartile of extracted left ventricular mass. We also measured the incidence rate of mortality by abnormal and normal LVEF and RVEF, defined as LVEF <50% and RVEF <45%, respectively [1,30]. Clinical outcomes were defined using previously described groupings of diagnostic codes [31,32]. For incidence analysis, we omitted individuals with the primary outcome (ie, atrial fibrillation or heart failure) occurring prior to or on the same day as the CMR. For incident atrial fibrillation and heart failure analyses, follow-up time began at the time of the CMR and continued until occurrence of the primary outcome, death, or last clinical encounter. For mortality analysis, follow-up time began at the time of the CMR and continued until time of death or last clinical encounter. Confidence intervals were calculated by the exact method. We compared incidence rates using the 2-sample test of proportions [33]. In order to assess potential confounding of report timing on associations between extracted features and clinical outcomes, we also performed a sensitivity analysis where we selected features extracted from the last report.

Results

Model Performance

The training set included reports from 270 individuals with a median age of 65 (IQR 54-74) years at time of CMR of whom 34.2% (n=92) were female (Table 2). The test set included reports from 100 individuals with a median age of 58 (IQR 45-66) years at time of CMR of whom 33% (n=33) were female (Table 2).

All combinations of pretrained weights and numerical representations achieved excellent macroaveraged F_1-scores on the test set. Table 3 illustrates the maximum macroaveraged F_1-scores for all combinations of pretrained weight initializations and numerical representations. The best-performing combination was $\text{BERT}_{\text{LARGE}}$, fine-tuned on the replaced decimal numerical representation scheme, which achieved a maximum macroaveraged F_1-score of 0.957 after fine-tuning for 12 epochs. A plot of macroaveraged F_1-score on the test set over the training epochs is available in Figure S1 in Multimedia Appendix 1**,** and featurewise receiver operating characteristic curves are shown in Figure 5. The range of feature-level macroaveraged F_1-scores was 0.902 to 1.000, and all scores are reported in Table S5, Multimedia Appendix 1. To investigate the impact of labeling effort on model performance, we fine-tuned this combination of $\text{BERT}_{\text{LARGE}}$ pretraining and the replaced decimal numerical representation scheme on varying subsets of the training data, and plotted the macroaveraged F_1-score on the test set (Figure 6). This plot demonstrates consistently significant gains in performance when the number of training reports is iteratively increased from 45 to about 200 but starts to saturate after this point. We also correlated the number of annotations in the training sample with test F_1 performance for each measurement and did not find a strong relationship (Figure S2, Multimedia Appendix 1).

Table 3. Maximum macroaveraged F_1-scores and bootstrapped 95% CIs on gold-standard test labels by pretrained weight initialization and numerical representation.

Architecture	Numerical representation, maximum macroaveraged F_1-score (95% CI)				
	Original	Replaced decimal	Consistent digits	Scientific	Words
PubMedBERT[a]	0.954 (0.947-0.960)	0.952 (0.947-0.960)	0.950 (0.945-0.955)	0.955[b] (0.948-0.960)	0.953 (0.949-0.958)
SapBERT	0.955 (0.949-0.960)	0.954 (0.949-0.960)	0.955 (0.949-0.960)	0.955 (0.948-0.960)	0.956[b] (0.951-0.961)
Bio+Discharge SummaryBERT	0.950 (0.944-0.957)	0.953[b] (0.947-0.959)	0.953 (0.945-0.958)	0.952 (0.945-0.958)	0.946 (0.942-0.952)
$\text{BERT}_{\text{LARGE}}$	0.951 (0.945-0.957)	0.957[b] (0.951-0.962)	0.951 (0.945-0.957)	0.944 (0.938-0.951)	0.952 (0.947-0.957)

[a] BERT: Bidirectional Encoder Representations from Transformers.

[b] Best-performing numerical representation for each pretrained weight initialization.

Figure 5. Receiver operating characteristic curves for model predictions on the test set by cardiac magnetic resonance imaging measurement. AUC: area under the receiver operating characteristic curve.

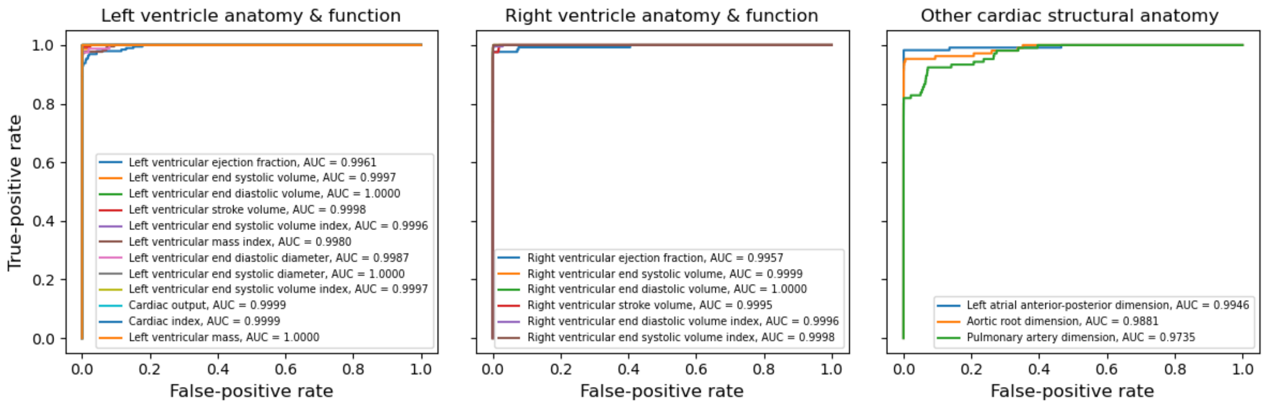
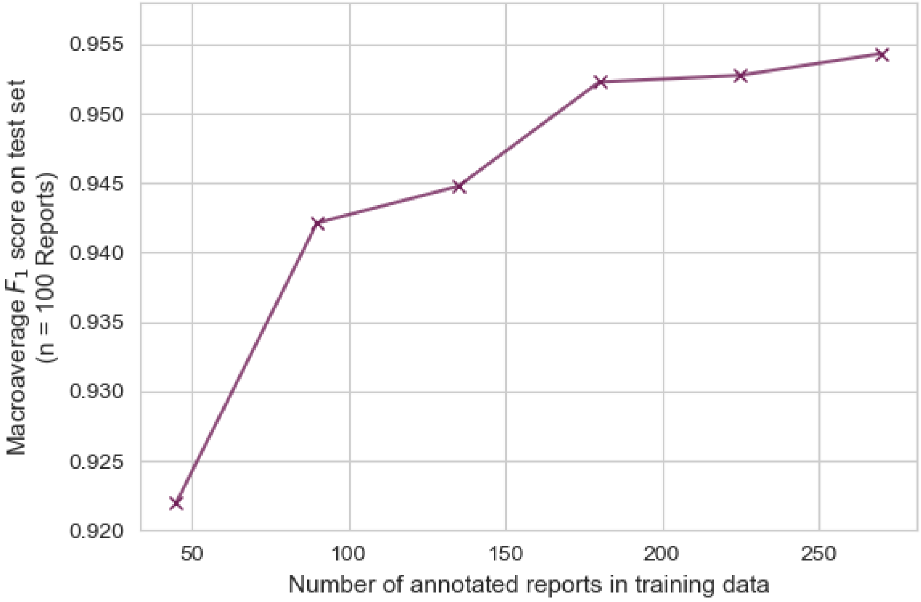


Figure 6. Fine-tuned $\text{BERT}_{\text{LARGE}}$ performance with replaced decimal numerical representations, as a function of number of annotated reports in the training set.

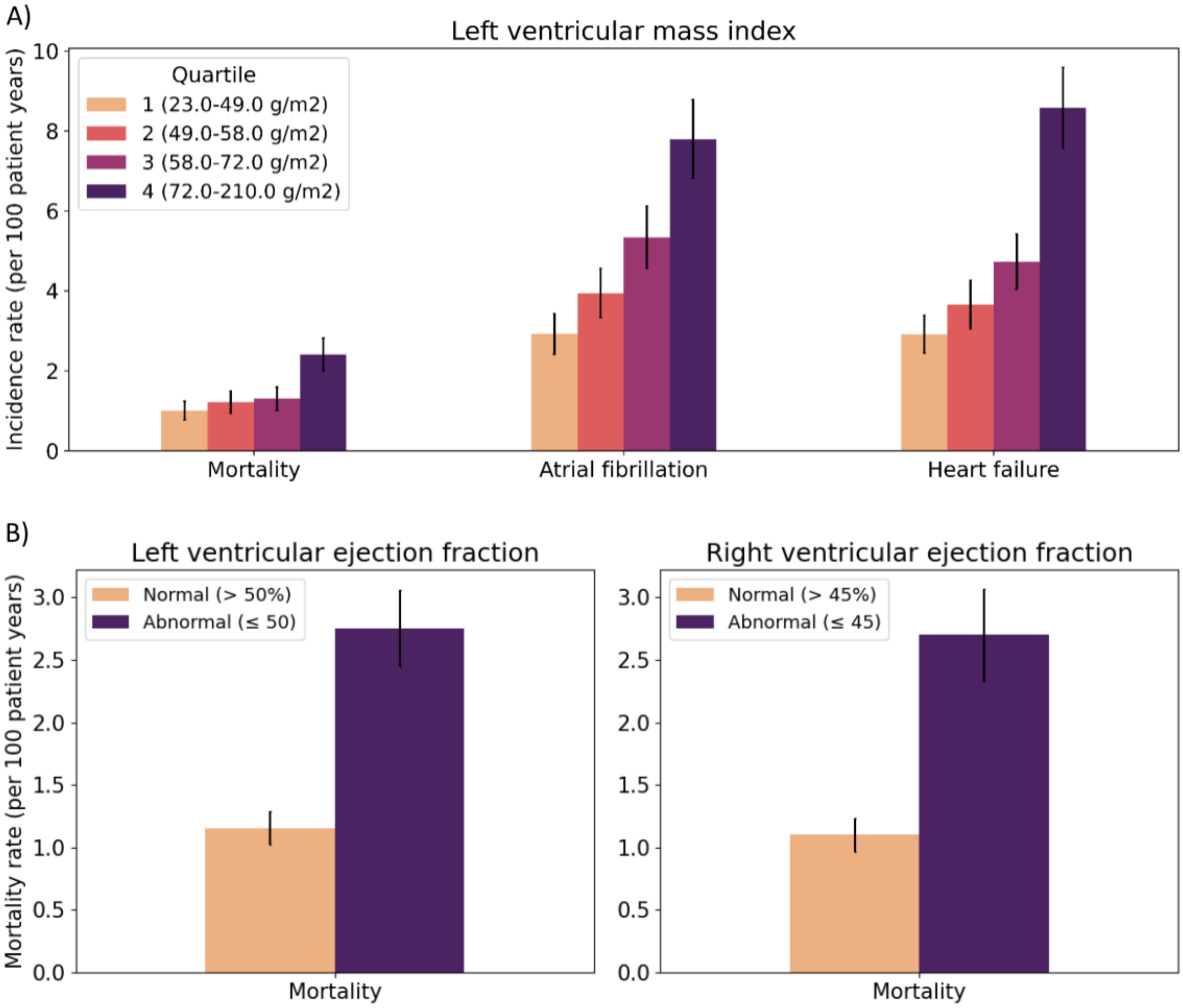


In EWOC, there were 12,720 CMR reports from 9280 individuals, which composed the CMR outcomes set (Figure 1). The median age of individuals in the outcomes set at the time of CMR was 57 (IQR 46-67) years, and 39.50% (3666/9280) were female (Table 1). After selecting the best model configuration, we applied the top-performing model to infer CMR values on all reports in this set. After running inference, we filtered by physiologic lower and upper bounds (Table S6, Multimedia Appendix 1) and extracted a total of 136,407 measurements. Counts for each extracted feature and distribution metrics are illustrated in Table S7 in Multimedia Appendix 1. We also compared the proportion of reports that contained model-predicted measurements in the CMR outcomes set and found them to be consistent with gold-standard annotation proportions in the test set (Table S8, Multimedia Appendix 1).

Associations With Clinical Outcomes

The median follow-up time of individuals in the CMR outcomes set was 5.3 (IQR 2.8-9.2). In the outcomes set, we observed 1520 incident heart failure events, 1488 incident atrial fibrillation events, and 909 deaths during follow-up. LVMI was extracted from 5015 of 9280 individuals (54.04%). In the outcomes set, increasing LVMI was associated with increasing incidence of mortality, atrial fibrillation, and heart failure with statistically significant differences in incidence rates between the lowest and highest quartiles (Figure 7). The mortality rate was 0.9 deaths per 100 person-years (PY; 95% CI 0.7-1.1) in the lowest quartile of extracted LVMI compared to 2.2 deaths per 100 PY (95% CI 1.9-2.6) in the highest quartile of extracted LVMI (*P*<.05; Figure 7). The incidence rate of atrial fibrillation was 3.0 events per 100 PY (95% CI2.5-3.5) in the lowest quartile of extracted LVMI compared to 7.9 events per 100 PY (95% CI 6.8-8.7) in the highest quartile of extracted LVMI (*P*<.05). The incidence rate of heart failure was 3.2 events per 100 PY (95% CI 2.7-3.7) in the lowest quartile of extracted LVMI compared to 8.1 events per 100 PY (95% CI 7.2-9.1) in the highest quartile of extracted LVMI (*P*<.05).

Figure 7. Association of extracted left ventricular mass index, left ventricular ejection fraction, and right ventricular ejection fraction with clinical outcomes.



LVEF was extracted from 7389 of 9280 individuals (79.62%), and 2297 met the criteria for abnormal LV systolic dysfunction (LVEF <50%). RVEF was extracted from 6324 of 9280 individuals (68.15%), and 1626 met criteria for abnormal RV systolic function (RVEF <45%; Figure 7). Both abnormal LVEF and RVEF were significantly associated with increased incidence of mortality compared to normal ventricular function (*P*<.05 for both measures). In the abnormal LVEF group, the mortality rate was 2.5 deaths per 100 PY (95% CI 2.2-2.8) compared to 1.1 deaths per 100 PY (95% CI 0.9-1.2) in the normal LVEF group (*P*<.05). In the abnormal RVEF group, the mortality rate was 2.5 deaths per 100 PY (95% CI 2.1-2.8) compared to 1.0 deaths per 100 PY (95% CI 0.9-1.2) in the normal RVEF group (*P*<.05).

We also performed a sensitivity analysis where the last CMR report was used for feature extraction of LVMI, LVEF, and RVEF. There were 687 of 5015 (13.70%) individuals with more than 1 extracted LVMI, 1268 of 7389 (17.16%) individuals with more than 1 extracted LVEF, and 1038 of 6324 (16.41%) individuals with more than 1 extracted RVEF. The mean time difference between the first and last reports for LVMI was 2.4 (SD 2.2) years, the LVEF was 2.9 (SD 2.9) years, and the RVEF was 2.7 (SD 2.6) years. Similar to the primary analysis, we observed increasing rates of mortality, atrial fibrillation, and heart failure with increasing LVMI; and significantly higher mortality rates in individuals with abnormal LVEF or RVEF compared to individuals with normal LVEF or RVEF (Figure S3, Multimedia Appendix 1).

Discussion

Principal Results

In this study, we report the results of an accurate and practical NLP-based approach for simultaneously extracting 21 quantitative measurements from CMR reports. Our final model, which yielded a macroaveraged F_1-score of 0.957, was derived from a workflow leveraging open-source frameworks for collecting gold-standard clinician labels and publicly available transformer model weights. We also highlight the clinical validity of our approach by demonstrating known associations of extracted CMR measurements with outcomes such as atrial fibrillation, heart failure, and mortality (Figure 7) [30,34].

We found that $\text{BERT}_{\text{LARGE}}$ demonstrated excellent performance when compared to model initializations based on clinically oriented pretraining, indicating that clinical pretraining does not have a significant impact on clinical numerical value extraction (Table 3). BERT$_{\text{LARGE}}$ is larger than the available clinically oriented models, and model complexity may play a role in comparable performance, indicating that larger clinically pretrained models represent a direction for future work. We also experimented with 4 different alternative representations of numerical measurements and found the test performance to be similar to that of the default representation (Table 3). Our findings suggest that for the particular case of extracting numerical quantities, transformer-based models do not require clinical pretraining or alternative numerical representations. Through experiments with limited training set sizes, we found that excellent performance can be achieved with fewer than 50 labeled reports. Furthermore, a training set with 175 reports was sufficient to train a model with performance that was within the 95% CI of a model trained with 270 reports (Figure 6).

Measurements extracted by our model potentially facilitate the automated characterization of a range of important cardiac diseases, which we leave to future work. We expect that our proposed workflow can be easily used by others to extract arbitrary measurements from clinical text. The PRAnCER platform is open source and can be easily adapted to label clinical measurements of interest. Our software for fine-tuning and evaluating NLP models is also open source [34], and model training is possible using a standard graphic processing unit–equipped machine. We expect it to be possible to extract an arbitrary number of clinical measurements with a practical amount of labeling effort and computational requirements in clinical domains not limited to CMRs.

Attention-Based Exploration of Error Modes

The characterization of error modes can be instructive toward having confidence in model predictions and for finding ways to improve a model by future researchers. Despite the overall high accuracy of our best model across all the types of measurements that we considered, the most common error mode involved the model assigning a "0" label to values that should have been labeled as measurements. In many cases that we examined, a measurement such as "aortic root dimension" would be correctly labeled in one report and not labeled in another report despite a similar sequence of tokens surrounding the value to be labeled. By examining the attention weights for the token to be labeled in both reports, we discovered that the correctly labeled value most heavily weighted the word "dimension" in the preceding "aortic root dimension" phrase. For the incorrectly labeled value, 3 of the 4 most-attended tokens were separate instances of the word "dimension," one of which was part of the correct phrase, with the other instances appearing in the remainder of the text. All of the attention weights were much lower than the attention paid to the word "dimension" by the correctly labeled example. This may indicate that an opportunity for further improvement could involve providing more training examples with sections of text that are absent from most reports in our data set or by augmenting existing labeled text with synthetic text containing critical tokens.

Additionally, we recognize that while our models perform well, extraction errors are inevitable. The clinical consequences of these errors depend on the specific feature. For example, incorrect LVEF extraction could misclassify a patient with heart failure as reduced ejection fraction or preserved ejection fraction and thereby impact treatment choices. Similarly, incorrect RVEF could misclassify a patient with right-sided heart failure. Incorrect aortic root size could misclassify an aortic root aneurysm. False-positive errors may be particularly difficult to detect as the final postprocessing stop of physiologic filtering means that false positives will still be within the expected range. Therefore, careful evaluation of model performance is necessary, especially if applying such a model to new data sets.

Comparison With Prior Work

To our knowledge, this is the first example of using a transformer-based model (without pretraining from scratch) fine-tuned on clinician labels to extract numerical measurements from diagnostic text. We previously demonstrated the value of extracting 4 vital sign measurements from clinical text based on a large number of weak labels that were generated using a rule-based approach [16]. Our previous approach was based on the assumption that it would be impractical to accrue a sufficient quantity of gold-standard annotations in order to fine-tune a transformer-based approach. However, we found that a single clinician required at most 15 hours to produce sufficient gold-standard annotations for 21 types of quantitative measurements, thereby eliminating the need for rule-based approaches and enabling easy scaling to a large number of relevant measurements.

Recent work [15] used a combination of embeddings produced by pretraining a BERT model and a FLAIR model from scratch on domain-specific data. Embeddings were then used as input to a combination of a bidirectional long short-term memory with a conditional random field layer to label tokens of interest, including numerical measurements. This approach worked well and achieved comparable performance to our approach with a similar amount of labeling effort. We demonstrate with our work that pretraining a model from scratch on domain-specific data is not necessary to achieve a high level of accuracy. The days, or perhaps even weeks, of computation required to pretrain a model from scratch on clinical data can be avoided. Furthermore, our work examines the impact of the number of annotations on performance.

Other approaches for extracting numerical measurements from clinical text have also achieved reasonable accuracy, but we suggest that our approach minimizes labeling effort, is more robust, and is sufficiently computationally efficient to serve as a practical solution for accelerating EHR-based clinical research. Rule-based approaches, while potentially accurate, generally require multiple iterations of development and validation to ensure accuracy given the wide variability of clinical text [4]. Prior work has also shown that rule-based approaches may not be easily portable to other EHRs outside of where they were developed. In their work evaluating the portability of a rule-based model for extraction of echocardiogram measurements, Adekkanattu et al [7] report variable F_1-scores that differ by clinical site. We demonstrate that transformer-based models pretrained on clinical text can be fine-tuned on a practical number of labels to learn to extract measurements in a way that is flexible to variability in how such measurements are expressed in clinical text.

Limitations and Directions for Future Work

Our study must be interpreted in the context of its limitations. Our test set consisted of a relatively small sample of 100 reports, but an analysis to randomly resample the test set of the same size yielded models with a markedly close range of macro F_1-scores (0.947-0.970 across 10 samples), which indicates the robustness of our approach. Our approach required a minimal degree of postprocessing and mainly involved imposing physiologic ranges for values extracted by the model. Although relatively few values were filtered this way, these may represent model false positives. Another aspect of postprocessing involved extending model predictions to include missed significant digits, which happened very rarely. Our experiments with numerical representations and pretrained models enabled high extraction accuracy, but further work is required to understand how to best use transformer-based models in handling arbitrary numerical values [35]. In addition, CMR reports were taken from a large heterogeneous health care system, and while our model was able to handle significant variability in the presentation of relevant measurements, further work is required to show that our modeling approach is portable to other institutions.

Similar to other artificial intelligence models with health care applications, clinical implementation of our model is stymied by several barriers [36]. The first is deployment of a model within an EHR environment, which involves both accessing siloed clinical data and integrating modeling results into the electronic environment for presentation. The second is ensuring that the model is adaptable to changes in report structure either between institutions or prospectively over the lifetime of the model. Last, monitoring and regular quality control is essential to ensuring patient safety. Although few models have successfully overcome these numerous challenges, we hypothesize that our work offers a modeling strategy that is adaptable to changes in report structure and provides a framework for developing new quantitative models aimed at other important clinical tasks. Future work should test the performance of models like these in real-time settings to prove generalizability to new environments and data structures.

Conclusions

We present a powerful natural language workflow for simultaneously extracting 21 types of numerical measurements from CMR free-text reports. We found that general pretrained transformer-based language models require a relatively small number of gold-standard annotations, necessitate minimal data processing, and are robust to significant variability in the context and presentation of numerical measurements. We observed expected associations between extracted CMR measurements and known clinical outcomes like heart failure, atrial fibrillation, and mortality. Our workflow is reproducible and is likely applicable to many other types of clinical data.

Acknowledgments

We would like to thank Monica Agrawal and David Sontag for their assistance with adapting the Platform Enabling Rapid Annotation for Clinical Entity Recognition (PRAnCER) platform to label cardiac magnetic resonance imaging (CMR) reports.

Conflicts of Interest

SAL is a full-time employee of Novartis as of July 18, 2022. SAL previously received support from NIH grants R01HL139731 and R01HL157635, and American Heart Association 18SFRN34250007. SAL has received sponsored research support from Bristol Myers Squibb, Pfizer, Boehringer Ingelheim, Fitbit, Medtronic, Premier, and IBM, and has consulted for Bristol Myers Squibb, Pfizer, Blackstone Life Sciences, and Invitae. JEH has received sponsored research support from Bayer AG. PB receives sponsored research support from Bayer AG and IBM and has consulted for Novartis and Prometheus Biosciences. CDA receives sponsored research support from Bayer AG and has consulted for ApoPharma. The other authors report no potential conflicts of interest.

Multimedia Appendix 1
Supplemental material.
[DOC File , 711 KB - medinform_v10i9e38178_app1.doc]

References

1. McMurray J, Adamopoulos S, Anker S, Auricchio A, Böhm M, Dickstein K, ESC Committee for Practice Guidelines. ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: The Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2012 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association (HFA) of the ESC. Eur Heart J 2012 Jul;33(14):1787-1847. [doi: 10.1093/eurheartj/ehs104] [Medline: 22611136]
2. Simon MA. Assessment and treatment of right ventricular failure. Nat Rev Cardiol 2013 Apr;10(4):204-218. [doi: 10.1038/nrcardio.2013.12] [Medline: 23399974]
3. Isselbacher EM. Thoracic and Abdominal Aortic Aneurysms. Circulation 2005 Feb 15;111(6):816-828. [doi: 10.1161/01.cir.0000154569.08857.7a]

4. Cai T, Zhang L, Yang N, Kumamaru KK, Rybicki FJ, Cai T, et al. EXTraction of EMR numerical data: an efficient and generalizable tool to EXTEND clinical research. BMC Med Inform Decis Mak 2019 Nov 15;19(1):226 [FREE Full text] [doi: 10.1186/s12911-019-0970-1] [Medline: 31730484]
5. Schwartz JL, Tseng E, Maruthur NM, Rouhizadeh M. Identification of prediabetes discussions in unstructured clinical documentation: validation of a natural language processing algorithm. JMIR Med Inform 2022 Mar 24;10(2):e29803 [FREE Full text] [doi: 10.2196/29803] [Medline: 35200154]
6. Nath C, Albaghdadi MS, Jonnalagadda SR. A natural language processing tool for large-scale data extraction from echocardiography reports. PLoS One 2016;11(4):e0153749 [FREE Full text] [doi: 10.1371/journal.pone.0153749] [Medline: 27124000]
7. Adekkanattu P, Jiang G, Luo Y, Kingsbury P, Xu Z, Rasmussen L, et al. Evaluating the portability of an NLP system for processing echocardiograms: a retrospective, multi-site observational study. AMIA Annu Symp Proc 2019;2019:190-199 [FREE Full text] [Medline: 32308812]
8. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. 2017 Presented at: Advances in Neural Information Processing Systems; Dec 4, 2017; Long Beach, CA URL: http://arxiv.org/abs/1706.03762
9. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); June 2, 2019; Minneapolis, MN URL: http://arxiv.org/abs/1810.04805 [doi: https://doi.org/10.18653/v1/N19-1423]
10. Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, Hill F, et al. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. 2019 Presented at: Advances in Neural Information Processing Systems; Dec 8, 2019; Vancouver, BC URL: http://arxiv.org/abs/1905.00537 [doi: https://doi.org/10.48550/arXiv.1905.00537]
11. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans. Comput. Healthcare 2022 Jan 31;3(1):1-23. [doi: 10.1145/3458754]
12. Liu F, Shareghi E, Meng Z, Basaldella M, Collier N. Self-alignment pretraining for biomedical entity representations. 2020 Presented at: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 6, 2021; Online URL: http://arxiv.org/abs/2010.11784 [doi: 10.18653/v1/2021.naacl-main.334]
13. Alsentzer E, Murphy J, Boag W, Weng W, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. 2019 Presented at: Proceedings of the 2nd Clinical Natural Language Processing Workshop; June 7, 2019; Minneapolis, MA URL: http://arxiv.org/abs/1904.03323 [doi: 10.18653/v1/w19-1909]
14. Zaman S, Petri C, Vimalesvaran K, Howard J, Bharath A, Francis D, et al. Automatic diagnosis labeling of cardiovascular mri by using semisupervised natural language processing of text reports. Radiol Artif Intell 2022 Jan;4(1):e210085 [FREE Full text] [doi: 10.1148/ryai.210085] [Medline: 35146435]
15. Syed S, Angel A, Syeda H, Jennings C, VanScoy J, Syed M, et al. The h-ANN model: comprehensive colonoscopy concept compilation using combined contextual embeddings. Biomed Eng Syst Technol Int Jt Conf BIOSTEC Revis Sel Pap 2022 Mar;5:189-200 [FREE Full text] [doi: 10.5220/0010903300003123] [Medline: 35373222]
16. Khurshid S, Reeder C, Harrington L, Singh P, Sarma G, Friedman S, et al. Cohort design and natural language processing to reduce bias in electronic health records research. NPJ Digit Med 2022 Apr 08;5(1):47 [FREE Full text] [doi: 10.1038/s41746-022-00590-0] [Medline: 35396454]
17. GitHub. JEDI. URL: https://github.com/broadinstitute/jedi-public [accessed 2022-01-01]
18. Moon S, Sagheb E, Liu S, Chen D, Bos M, Geske J, et al. Abstract 13811: An automated natural language processing algorithm to classify magnetic resonance imaging reports containing positive diagnoses of hypertrophic cardiomyopathy. Circulation 2019;140:A13811.
19. Github. PRAnCER: Platform enabling Rapid Annotation for Clinical Entity Recognition. URL: https://github.com/clinicalml/prancer [accessed 2022-01-01]
20. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 01;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
21. Nogueira R, Jiang Z, Lin J. Investigating the limitations of transformers with simple arithmetic tasks. 2021 Presented at: Mathematical Reasoning in General Artificial Intelligence Workshop, ICLR 2021; May 07, 2021; Online URL: http://arxiv.org/abs/2102.13019
22. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's Transformers: State-of-the-art natural language processing. 2020 Presented at: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; Nov 16-20, 2020; Online URL: http://arxiv.org/abs/1910.03771 [doi: 10.18653/v1/2020.emnlp-demos.6]
23. Models. Hugging Face. URL: https://huggingface.co/models [accessed 2022-03-14]
24. Kawel-Boehm N, Hetzel SJ, Ambale-Venkatesh B, Captur G, Francois CJ, Jerosch-Herold M, et al. Reference ranges ("normal values") for cardiovascular magnetic resonance (CMR) in adults and children: 2020 update. J Cardiovasc Magn Reson 2020 Dec 14;22(1):87 [FREE Full text] [doi: 10.1186/s12968-020-00683-3] [Medline: 33308262]

25. Olivotto I, Maron MS, Autore C, Lesser JR, Rega L, Casolo G, et al. Assessment and significance of left ventricular mass by cardiovascular magnetic resonance in hypertrophic cardiomyopathy. Journal of the American College of Cardiology 2008 Aug;52(7):559-566. [doi: 10.1016/j.jacc.2008.04.047]
26. Hombach V, Merkle N, Torzewski J, Kraus JM, Kunze M, Zimmermann O, et al. Electrocardiographic and cardiac magnetic resonance imaging parameters as predictors of a worse outcome in patients with idiopathic dilated cardiomyopathy. European Heart Journal 2009 Jul 24;30(16):2011-2018. [doi: 10.1093/eurheartj/ehp293]
27. de Simone G, Gottdiener J, Chinali M, Maurer M. Left ventricular mass predicts heart failure not related to previous myocardial infarction: the Cardiovascular Health Study. Eur Heart J 2008 Mar;29(6):741-747. [doi: 10.1093/eurheartj/ehm605] [Medline: 18204091]
28. Vakili BA, Okin PM, Devereux RB. Prognostic implications of left ventricular hypertrophy. American Heart Journal 2001 Mar;141(3):334-341. [doi: 10.1067/mhj.2001.113218]
29. Verdecchia P, Reboldi G, Gattobigio R, Bentivoglio M, Borgioni C, Angeli F, et al. Atrial fibrillation in hypertension. Hypertension 2003 Feb;41(2):218-223. [doi: 10.1161/01.hyp.0000052830.02773.e4]
30. Surkova E, Muraru D, Genovese D, Aruta P, Palermo C, Badano LP. Relative prognostic importance of left and right ventricular ejection fraction in patients with cardiac diseases. J Am Soc Echocardiogr 2019 Nov;32(11):1407-1415.e3. [doi: 10.1016/j.echo.2019.06.009] [Medline: 31400846]
31. Goff DC, Pandey DK, Chan FA, Ortiz C, Nichaman MZ. Congestive heart failure in the United States: is there more than meets the I(CD code)? The Corpus Christi Heart Project. Arch Intern Med 2000 Jan 24;160(2):197-202. [doi: 10.1001/archinte.160.2.197] [Medline: 10647758]
32. Khurshid S, Keaney J, Ellinor PT, Lubitz SA. A simple and portable algorithm for identifying atrial fibrillation in the electronic medical record. Am J Cardiol 2016 Jan 15;117(2):221-225 [FREE Full text] [doi: 10.1016/j.amjcard.2015.10.031] [Medline: 26684516]
33. Han C. Comparing two independent incidence rates using conditional and unconditional exact tests. Pharm Stat 2008;7(3):195-201. [doi: 10.1002/pst.289] [Medline: 17506083]
34. Nagata Y, Wu VC, Kado Y, Otani K, Lin F, Otsuji Y, et al. Prognostic value of right ventricular ejection fraction assessed by transthoracic 3D echocardiography. Circ Cardiovasc Imaging 2017 Feb;10(2):e005384. [doi: 10.1161/CIRCIMAGING.116.005384] [Medline: 28174197]
35. Thawani A, Pujara J, Ilievski F, Szekely P. Representing numbers in NLP: a survey and a vision. : Association for Computational Linguistics; 2021 Presented at: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 06, 2021; Online p. 644-656. [doi: 10.18653/v1/2021.naacl-main.53]
36. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019 Oct 29;17(1):195 [FREE Full text] [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
CMR: cardiac magnetic resonance imaging
EHR: electronic health record
EWOC: Enterprise Warehouse of Cardiology
LVEF: left ventricular ejection fraction
LVMI: left ventricular mass index
NIH: National Institutes of Health
NLP: natural language processing
PRAnCER: Platform Enabling Rapid Annotation for Clinical Entity Recognition
PY: person-years
RVEF: right ventricular ejection fraction

Edited by T Hao; submitted 21.03.22; peer-reviewed by A Arruda-Olson, G Lim, M Syed, R Abeysinghe; comments to author 28.05.22; revised version received 22.07.22; accepted 11.08.22; published 16.09.22.

Please cite as:
Singh P, Haimovich J, Reeder C, Khurshid S, Lau ES, Cunningham JW, Philippakis A, Anderson CD, Ho JE, Lubitz SA, Batra P
One Clinician Is All You Need–Cardiac Magnetic Resonance Imaging Measurement Extraction: Deep Learning Algorithm Development
JMIR Med Inform 2022;10(9):e38178
URL: https://medinform.jmir.org/2022/9/e38178
doi:10.2196/38178
PMID:35960155

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org