# JMIR Medical Informatics

# Contents

## Reviews

## Original Papers

XSL•FO

**RenderX**

# Viewpoint

Blockchain-Based Digital Contact Tracing Apps for COVID-19 Pandemic Management: Issues, Challenges, Solutions, and Future Directions (e25245)

Review

# Electronic Medical Record–Based Case Phenotyping for the Charlson Conditions: Scoping Review

Seungwon Lee[1,2,3,4], MPH; Chelsea Doktorchik[1,2], MSc; Elliot Asher Martin[1,3], PhD; Adam Giles D'Souza[1,3], PhD; Cathy Eastwood[1,2], RN, PhD; Abdel Aziz Shaheen[1,2,5], MBChB, MSc; Christopher Naugler[2,6], MD; Joon Lee[1,2,4,7], PhD; Hude Quan[1,2], PhD, MD

[1]Centre for Health Informatics, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

[2]Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

[3]Alberta Health Services, Calgary, AB, Canada

[4]Data Intelligence for Health Lab, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

[5]Department of Medicine, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

[6]Department of Pathology and Laboratory Medicine, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

[7]Department of Cardiac Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

**Corresponding Author:**
Hude Quan, PhD, MD
Department of Community Health Sciences
Cumming School of Medicine
University of Calgary
Teaching, Research, & Wellness Building
3280 Hospital Dr NW
Calgary, AB, T2N 4Z6
Canada
Phone: 1 403 210 9317
Email: hquan@ucalgary.ca

## *Abstract*

**Background:** Electronic medical records (EMRs) contain large amounts of rich clinical information. Developing EMR-based case definitions, also known as EMR phenotyping, is an active area of research that has implications for epidemiology, clinical care, and health services research.

**Objective:** This review aims to describe and assess the present landscape of EMR-based case phenotyping for the Charlson conditions.

**Methods:** A scoping review of EMR-based algorithms for defining the Charlson comorbidity index conditions was completed. This study covered articles published between January 2000 and April 2020, both inclusive. Embase (Excerpta Medica database) and MEDLINE (Medical Literature Analysis and Retrieval System Online) were searched using keywords developed in the following 3 domains: terms related to EMR, terms related to case finding, and disease-specific terms. The manuscript follows the Preferred Reporting Items for Systematic reviews and Meta-analyses extension for Scoping Reviews (PRISMA) guidelines.

**Results:** A total of 274 articles representing 299 algorithms were assessed and summarized. Most studies were undertaken in the United States (181/299, 60.5%), followed by the United Kingdom (42/299, 14.0%) and Canada (15/299, 5.0%). These algorithms were mostly developed either in primary care (103/299, 34.4%) or inpatient (168/299, 56.2%) settings. Diabetes, congestive heart failure, myocardial infarction, and rheumatology had the highest number of developed algorithms. Data-driven and clinical rule–based approaches have been identified. EMR-based phenotype and algorithm development reflect the data access allowed by respective health systems, and algorithms vary in their performance.

**Conclusions:** Recognizing similarities and differences in health systems, data collection strategies, extraction, data release protocols, and existing clinical pathways is critical to algorithm development strategies. Several strategies to assist with phenotype-based case definitions have been proposed.

XSL•FO
RenderX

## Introduction

### Background

Recent advances in computational power, increased adoption of electronic medical records (EMRs), and the subsequent rise of big data analytics in health care have opened the door to precision medicine [1]. EMRs are systemized collections of patient health information and documentation, collected in real time and stored in a digital format. EMRs were originally designed to facilitate communication in support of clinical decision-making for individual patients and to improve the quality of care. Canada and other countries have heavily promoted EMR adoption [2,3]. Globally, EMR data have been used widely for secondary purposes, such as research.

Developing case definitions, a process known as phenotyping, has become an active area of research associated with EMRs. Establishing EMR data–based phenotyping is essential for setting up the operational framework toward pursuing precision medicine, which aims to tailor medical decisions and treatments to each patient in a timely manner. EMR phenotyping allows identification and surveillance of health conditions in a timely manner and can be integrated into existing clinical flows and infrastructure. Phenotyping comorbidities using EMR data have important implications on disease management. Comorbidity is a medical condition existing simultaneously with but independently from another condition in a patient. These diseases may be related to each other by some shared association [4]. The Charlson comorbidity index [4-6] is a measure that predicts 1-year mortality based on the presence or absence of specific chronic conditions. Typically, each condition is identified through the presence of specific International Classification of Diseases (ICD) codes and assigned a score depending on the risk of death. Scores are summed for each patient to provide a total score to predict mortality [7,8]. The Charlson [5] comorbidity algorithm is the most widely used comorbidity index at present and has demonstrated the importance of classifying conditions using health data [6,7], including risk adjustment analysis, developing patient safety indicators, and identifying specific disease cohorts for research and public health applications.

### Objectives

Few reviews [9-12] have been published on developing EMR case definitions or phenotyping algorithms for selected chronic conditions, but none specifically cover all of the Charlson comorbidities. Furthermore, these articles narrowed their scope to specific perspectives [10] or specific settings (eg, inpatient or primary care only) [9,11]. These reviews report few studies utilizing natural language processing (NLP) or machine learning (ML), which emphasizes the importance of data science techniques (eg, deep learning) in the present health research. The primary objective of this study is to provide an overview of EMR-based phenotyping algorithms for the Charlson conditions. The secondary objective is to provide recommendations for health systems considering the adoption of EMR-based case phenotyping.

## Methods

### Article Screening

The methodology follows the guidelines recommended by the Preferred Reporting Items for Systematic Reviews and Meta-analysis Extension Protocols for Scoping Reviews (PRISMA-ScR) [13]. The Excerpta Medica database (Embase), and Medical Literature Analysis and Retrieval System Online (MEDLINE) databases were searched from January 2000 to April 2020 to identify peer-reviewed papers. The search strategy covered the following 3 domains: (1) terms related to EMRs, (2) terms related to case finding, and (3) disease-specific terms. We initially used validated clinical text descriptions from ICD-10 to derive search terms for selected conditions (Multimedia Appendix 1). Boolean algorithms were developed for each specific condition using the domain keywords (Multimedia Appendix 2). The cancer categories of metastatic cancer and malignant cancer were excluded, as there is already an existing review on this topic [11].

Manual screening was performed according to the following established study guidelines. Peer-reviewed journal papers were included if they were published between January 2000 and April 2020, written in English, involved human subjects and EMR, and were retrieved by the Boolean search algorithm for at least one Charlson condition. This review study focused only on case phenotyping using EMR data, and therefore, papers were excluded if they only involved administrative databases. Administrative data studies that linked EMR data were included. The presence of the Charlson conditions in each study, if reported, was defined by the presence of ICD-9 or ICD-10 codes stated in the manuscript. The full PRISMA flow diagram was created (Multimedia Appendix 3). The final search results were exported to a reference software (EndNote, Clarivate Inc) [14], and duplicates were removed.

### Characterizing the Identified Literature

A data extraction form was developed. The extracted data components included article characteristics (year and country), health care type (eg, inpatient, outpatient, and emergency), specific name of the data source, whether diagnostic codes (eg, ICD) were used, types of EMR data (eg, structured, unstructured, or imaging), techniques (eg, epidemiology/biostatistics, ML, or NLP), and whether a validation methodology was employed. The extracted data types (categorical) were recoded as binary variables to indicate whether they were employed in the algorithm. The frequencies of the algorithms, EMR settings, and countries were calculated. The identified algorithms were substratified into the following 7 types in this review based on the types of data used: (1) diagnostic codes only; (2) codes and structured data (demographics, labs, and medications); (3) diagnostic codes and free-text data; (4) diagnostic codes, structured, and free-text data; (5) structured data only; (6)

free-text data only; and (7) free-text and structured data. The detailed operational definitions of case definitions used in the identified studies were also extracted. The extracted data were summarized using frequencies and graphs where applicable. STATA 14 software (StataCorp LLC) [15] was used for statistical analysis. We further summarized the used data elements, disease context, data linkage, and validation of phenotyping algorithms using the extracted tables.

## Results

### Article Screening

After 1097 duplicates were removed, a total of 3691 abstracts were identified from the electronic databases. A total of 3402 abstracts were excluded based on the title and abstract screening, resulting in 289 full-text articles for full article review. Of these, 39 articles were excluded because they did not include any Charlson conditions, and 22 articles could not be retrieved, leading to the exclusion of 61 articles. The remaining 228 articles were considered eligible for this review and analyzed. References of eligible full articles were screened, and additional articles were identified for inclusion (n=46), leading to a total of 274 articles for qualitative synthesis. Articles covering multiple disease phenotypes were counted once per phenotype, leading to a total of 299 disease phenotyping algorithms. The PRISMA diagram depicting this process is shown in Multimedia Appendix 3.

### Characteristics of the Identified Literature

The frequencies of the algorithms, EMR settings, and countries are shown in Table 1. The complete data extraction table is presented in Multimedia Appendix 4 [16-285]. A total of 274 articles representing 299 algorithms from 22 countries were identified in this review. The majority of this work was undertaken in the United States (181/299, 60.5%), followed by the United Kingdom (42/299, 14.0%) and Canada (15/299,

5.0%). Algorithm development has steadily increased over the years, with the majority of work published after 2016. The distributions of these algorithms by the year of publication and by country are shown in Figure 1. The breakdown of the disease areas of these algorithms is shown in Figure 2.

Table 2 provides a summary of the algorithm types used for each Charlson condition. The most common algorithm types were diagnostic codes and structured data (167/299, 55.9%), followed by diagnostic codes, structured and free-text data (51/299, 17.1%), and diagnostic codes only (40/299, 13.4%). Variations in the data sources used were observed based on disease context and data availability.

These algorithms were mostly developed either in primary care (103/299, 34.4%) or inpatient (168/299, 56.2%) settings. A total of 23 algorithms (23/299, 7.7%) used data sources from inpatient and outpatient EMR. This trend was consistent across the conditions assessed in this review. The United States had the highest algorithm count across most of the assessed conditions, followed by the United Kingdom, Canada, and other nations. Detailed information about the distribution of algorithms by disease, EMR setting, and country is shown in Table 1.

We abstracted study objectives and classified different purposes for which algorithms were developed for, as well as the setting of each study (Multimedia Appendix 4). Phenotyping algorithm development was not always the primary objective for the identified studies; sometimes, it was part of a larger process. The most commonly occurring objectives of the algorithms were (1) phenotyping algorithm development (193/299, 64.5%), (2) epidemiological analysis (70/299, 23.4%), and (3) predictive modeling (19/299, 6.4%). Other objectives included designing clinical decision support and implementation tools, genome analysis, and registry development. These objectives reflect the health system delivery and clinical practice contexts in which the studies were situated.

**Table 1.** Descriptive summary of the 299 Charlson algorithms.

| Disease | Algorithm count | EMR[a] settings | | | | Country |
|---|---|---|---|---|---|---|
| | | Inpatient | Inpatient and outpatient | Outpatient | Other | |
| Myocardial infarction [16-38] | 23 | 16 | 3 | 4 | 0 | • 16 United States<br>• 2 United Kingdom<br>• 5 Others |
| Congestive heart failure [19,39-75] | 38 | 22 | 1 | 14 | 1 | • 27 United States<br>• 3 Sweden<br>• 8 Others |
| Peripheral vascular disease [19,76-89] | 15 | 6 | 1 | 7 | 1 | • 9 United States<br>• 5 United Kingdom<br>• 1 Norway |
| Cerebrovascular disease [19,47,57,78,83,90-107] | 23 | 14 | 0 | 9 | 0 | • 10 United States<br>• 6 United Kingdom<br>• 7 Others |
| Hemiplegia and paraplegia | 0 | 0 | 0 | 0 | 0 | 0 |
| Dementia [19,84,108-130] | 25 | 13 | 1 | 10 | 1 | • 8 United States<br>• 2 United Kingdom<br>• 1 Netherlands<br>• 1 Canada<br>• 13 Others |
| Chronic pulmonary disease [129,131-160] | 31 | 14 | 1 | 16 | 0 | • 16 United States<br>• 6 United Kingdom<br>• 4 Canada<br>• 5 Others |
| Rheumatologic disease [161-185] | 25 | 15 | 1 | 9 | 0 | • 15 United States<br>• 7 United Kingdom<br>• 3 Others |
| Peptic ulcer disease [186-189] | 4 | 3 | 0 | 1 | 0 | • 3 United States<br>• 1 Singapore |
| Diabetes [19,28,34,47,48,84,128,129,140,150,166,190-234] | 56 | 30 | 6 | 20 | 0 | • 31 United States<br>• 8 Canada<br>• 4 United Kingdom<br>• 13 Others |
| Diabetes, with complications [57,235-242] | 9 | 6 | 1 | 2 | 0 | • 5 United States<br>• 2 United Kingdom<br>• 1 Israel<br>• 1 China |
| Renal disease [47,57,243-262] | 22 | 9 | 3 | 9 | 1 | • 16 United States<br>• 2 United Kingdom<br>• 2 Spain<br>• 2 Others |
| Mild liver disease [189,263-276] | 15 | 11 | 3 | 0 | 1 | • 11 United States<br>• 2 China<br>• 1 Australia<br>• 1 United Kingdom |
| Moderate/severe liver disease [244,275-280] | 7 | 5 | 0 | 2 | 0 | • 4 United States<br>• 1 United Kingdom<br>• 1 Netherlands<br>• 1 China |

| Disease | Algorithm count | EMR[a] settings | | | | Country |
|---------|-----------------|-----------------|---|---|---|---------|
| | | Inpatient | Inpatient and outpatient | Outpatient | Other | |
| HIV [137,281-285] | 6 | 4 | 2 | 0 | 0 | 6 United States |

[a]EMR: electronic medical record.

**Figure 1.** Distribution of published articles by country between January 2000 and April 2020.
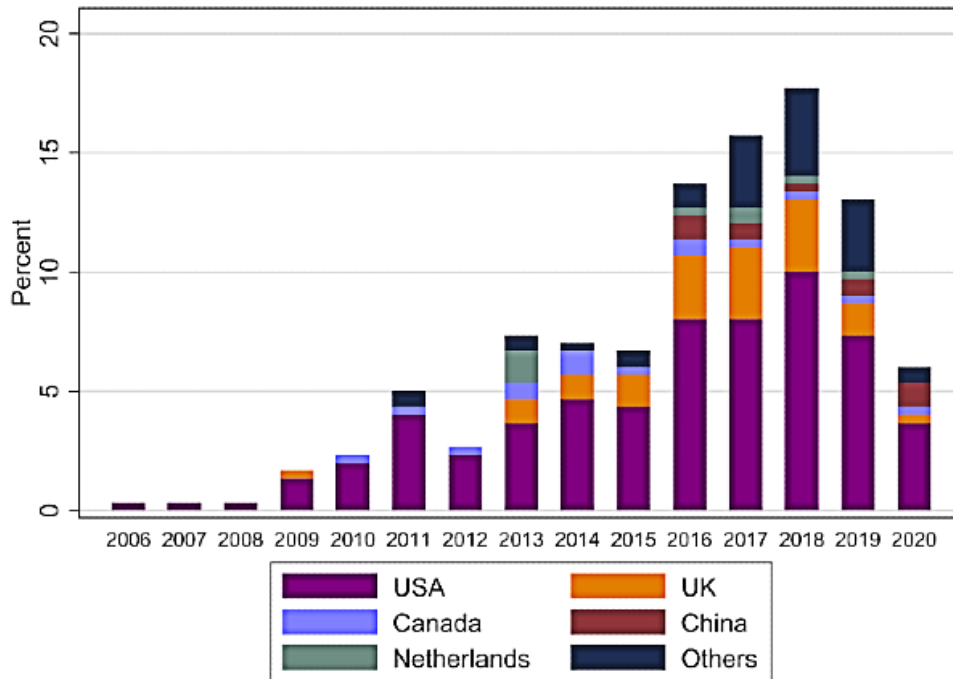


**Figure 2.** Distribution of electronic medical record data–based algorithms by Charlson disease area.
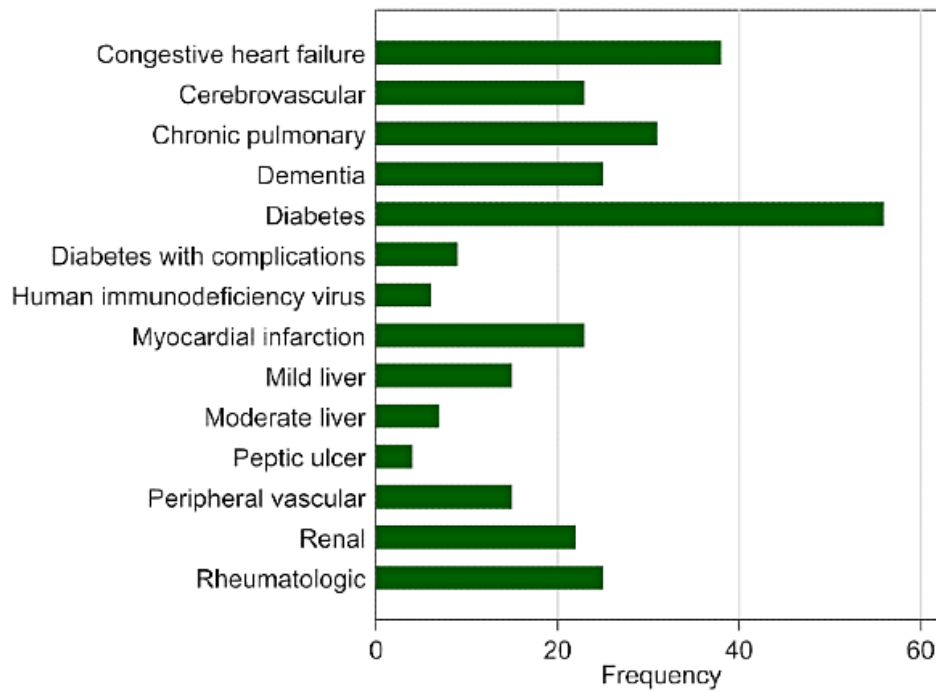
**Table 2.** The Charlson algorithm types identified in this scoping review.

| Charlson condition | Algorithm type | | | | | | |
|---|---|---|---|---|---|---|---|
| | Codes only | Codes and structured data | Codes and free-text data | Codes, structured, and free-text data | Structured data only | Free-text data only | Free-text and structured data |
| Myocardial infarction (n=23, 7.7%)) | 7 | 10 | 0 | 4 | 0 | 2 | 0 |
| Congestive heart failure (n=38, 12.7%)) | 7 | 19 | 2 | 9 | 0 | 0 | 1 |
| Peripheral vascular disease (n=15, 5.0%) | 1 | 6 | 1 | 2 | 2 | 2 | 1 |
| Cerebrovascular disease (n=23, 7.7%) | 6 | 14 | 0 | 1 | 1 | 1 | 0 |
| Dementia (n=25, 8.4%) | 4 | 8 | 4 | 4 | 1 | 1 | 3 |
| Chronic pulmonary disease (n=31, 10.4%) | 3 | 17 | 3 | 3 | 0 | 4 | 1 |
| Rheumatologic disease (n=25, 8.4%) | 2 | 9 | 2 | 12 | 0 | 0 | 0 |
| Peptic ulcer disease (n=4, 1.3%) | 0 | 1 | 1 | 2 | 0 | 0 | 0 |
| Diabetes (n=56, 18.7%) | 6 | 41 | 1 | 6 | 1 | 0 | 1 |
| Diabetes with complications (n=9, 3.0%) | 1 | 8 | 0 | 0 | 0 | 0 | 0 |
| Renal disease (n=22, 7.4%) | 2 | 15 | 1 | 2 | 1 | 0 | 1 |
| Mild liver disease (n=15, 5.0%) | 1 | 11 | 0 | 2 | 0 | 0 | 1 |
| Moderate/severe liver disease (n=7, 2.3%) | 0 | 3 | 0 | 3 | 0 | 0 | 1 |
| HIV (n=6, 2.0%) | 0 | 5 | 0 | 1 | 0 | 0 | 0 |
| Combined (n=299, 100.0%) | 40 | 167 | 15 | 51 | 6 | 10 | 10 |

## Data Elements: Structured Versus Unstructured

With regard to the EMR algorithms identified in this study, structured data most commonly consisted of demographics, diagnoses, procedures, vital signs, laboratory results, and medications. Structured data elements were the most common type of data employed by clinical rule–based algorithms and included basic demographics (eg, sex and age), medications, laboratory data, and diagnostic codes. A total of 233 out of 299 (77.9%) algorithms employed key laboratory diagnostic tests based on the present clinical practice.

These structured EMR components are typically available across EMR systems. Algorithms based on diagnostic codes and structured data were used primarily (213/299, 71.2%) for chronic conditions such as diabetes, where laboratory tests and medication may be necessary and sufficient for clinical decision-making. The use of diagnostic codes depended on the EMR setting (ie, outpatient or inpatient) and the health services jurisdiction (eg, United Kingdom vs United States vs Canada) where the work took place (Multimedia Appendix 4). Types of diagnostic codes identified included ICD-9, ICD-10, Read, Oxford Medical Information System, and International Classification of Primary Care (ICPC). ICD codes were used predominantly within inpatient settings (148/168, 88.1%). These basic structured data-based definitions were enhanced by incorporating unstructured data such as free text and imaging for designing classification algorithms (Table 2) for complicated chronic conditions. In summary, the disease context determined the data elements that were used.

Unstructured free-text data (eg, discharge summaries, consult notes, and nursing notes) were incorporated in approximately

86 out of 299 (28.8%) case phenotyping algorithms. NLP techniques were used to analyze such unstructured free-text data. Many studies used controlled medical terminologies, such as the Unified Medical Language System [286] and the Systematized Nomenclature of Medicine Clinical Terms [287], in the processing of clinical notes. Both terminologies can be used by medical researchers. Many studies also employed custom vocabularies developed in consultation with clinicians or had clinicians manually annotate the free-text data to obtain the reference standard. Variations in the processing of the unstructured data were also noted. NLP processing programs such as clinical Text Analysis and Knowledge Extraction System [288], MedTagger [289], or in-house programs were employed using one of the terminologies mentioned above. This data processing converted unstructured free-text data into structured data. The converted data are often combined with existing structured data for phenotyping and disease prediction using a wide range of techniques in epidemiology, statistics, and ML. Cox regression modeling was used for survival analysis, along with incidence and prevalence in epidemiological studies. Supervised learning classification algorithms such as Naive Bayes, support vector machines, logistic regression, and neural networks are commonly used in the ML studies. The manually annotated notes or reference standard obtained from the chart review provided labels for supervised ML.

## Disease Context

Case phenotyping algorithms exhibited 2 distinct types of approaches: clinician-derived rule-based (ie, expert-driven) and data-driven approaches. Clinician-derived rule-based approaches for defining cases were based on clinical criteria dictated by guidelines or clinical practice. These rule-based methods are

generally easy to interpret and are accepted as clinically relevant. However, criteria were inconsistent within and across multiple diseases even for the clinical rule-based case phenotyping, implying that the interpretation of algorithm results may depend on choices made during the algorithm development process [290]. Despite these variations, common structured data elements were identified in each disease discipline within each context of patient care. In contrast, data-driven approaches to defining cases use information extracted from available data to determine the disease status of the patient, often with improved performance (eg, sensitivity, positive predictive value [PPV], and F1 score) compared with baseline rule-based algorithms. For example, feeding all available free-text and laboratory data for congestive heart failure (CHF) into a prediction model can classify the CHF status [73]. One study employed principal component analysis [34]. However, the association between the predictor variables and outcomes is often difficult to ascertain, and the model may be difficult to interpret.

The algorithms used various EMR data elements depending on the clinical disease context. For each disease area, unique diagnostic methods or clinical data elements were observed. Diabetes was the most commonly identified disease in our literature search (56/299, 18.7%) and will be used as an example. Case phenotyping for diabetes had fewer data element variations compared with other diseases, and algorithms involved hemoglobin $A_{1c}$ ($HbA_{1c}$), glucose levels, and fasting glucose as key laboratory tests and antidiabetic medications. Most diabetes algorithms did not define the severity of the disease but classified the conditions in terms of the presence or absence of type 1 or type 2 diabetes. Diabetes phenotyping studies designed patient cohort selection taking this into consideration. Developing phenotypes for identifying severe complications of diabetes did require additional data (ie, clinical narratives) and advanced methodological approaches (eg, NLP and ML), as structured data alone would not readily identify these unless diagnostic codes were included for such complications. EMR phenotypes for disease severity were sometimes developed, in the case of chronic conditions that have a widely accepted clinical severity definition. Using chronic kidney disease as an example, severity was defined according to the Kidney Disease Improving Global Outcomes [291] and the National Kidney Foundation [292] guidelines based on estimated glomerular filtration rate.

### Data Linkage

A subset of phenotyping algorithms (30/299, 10.0%) linked EMR data to disease registries or genomics data. A total of 24 out of 299 (8.0%) algorithms linked clinical and health administrative databases. All data linkage occurred in studies that used diagnostic codes. The most commonly occurring diagnostic codes were ICD-9 and ICD-10, with some regional or national diagnostic codes (eg, Read codes among UK studies). The EMR administrative data linkage context appeared mostly within primary care data-based algorithms (14/24). The UK Clinical Practice Research Datalink was linked with Hospital Episode Statistics and other administrative data to primary care EMR. The most commonly linked inpatient care data came from the Electronic Medical Records and Genomics (eMERGE) consortium [293], which provided additional validation between

clinical documentation and scientific (ie, genomic) observation. These data linkage studies were employed for epidemiological analyses (improved accuracy of incidence and prevalence estimates) of diseases at the population level [83,96,212].

### Validity of Phenotyping Algorithms

Studies varied in their reporting metrics for the validity of case definition algorithms. Commonly reported metrics were sensitivity, specificity, positive predictive value, negative predictive value, accuracy, and F1 score. A total of 185 algorithms (185/299, 62.1%) employed chart review as the reference standard to calculate some of the aforementioned validation metrics. Of these 185 algorithms, 9 employed ML, 39 employed NLP, and 17 employed both ML and NLP. Of the 114 algorithms that did not conduct a chart review, 17 incorporated ML, 14 incorporated NLP, and 7 employed both ML and NLP techniques. Including free-text data as a data source in phenotyping algorithms tended to yield higher performance, with an average sensitivity of 0.906 (SD 0.110) and PPV of 0.913 (SD 0.120) when compared with studies that did not use free-text or ML (average sensitivity of 0.825 (SD 0.214) and average PPV of 0.853 (SD 0.174)). Incorporation of ML as part of the data-driven phenotyping also led to similar performance in sensitivity but weaker PPV, with an average sensitivity of 0.832 (SD 0.095) and average PPV of 0.633 (SD 0.358). In total, 59 out of 166 (35.5%) inpatient algorithms employed NLP, whereas 10 out of 93 (10.8%) primary care algorithms employed NLP. Among the works that used NLP, terminology standards were based on either Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) or Unified Medical Language System (UMLS), although many developed their own in-house keywords. Coding standards within inpatients were based on either ICD-9 or ICD-10 depending on the timing of the study and the jurisdictions where each study took place. Similarly, primary care code standards also varied. For example, mostly Read or ICPC codes were used in the United Kingdom, whereas ICD codes were used in North America (United States and Canada). The additional data provide a specific list of ML techniques that were used in each study, if employed (Multimedia Appendix 4).

## Discussion

### EMR Phenotyping and Precision Medicine

Achieving precision medicine requires the right information to be delivered to the right personnel at the right time. Developing EMR data–based phenotypes and integrating them into existing health information systems is a pivotal step for building a learning health system. EMR phenotypes allow rapid detection of diseases and accelerate the delivery of information to clinicians who may need it to make informed clinical decisions, policymakers who may use them to obtain population information for making public health decisions, and health services organizations that may need such information for planning clinical operations or developing risk adjustment models for patient safety programs. The purposes of the case definitions identified in this review were largely achieving one of the stated objectives above.

EMR-based phenotype and algorithm development reflected the structure and data available within respective health systems. Diagnostic codes, such as ICD and present procedural terminology codes, are often used for billing purposes within inpatient and outpatient (ie, primary care) settings in certain countries (eg, the United States). These codes were also built into EMR systems (eg, problem lists). Consequently, these diagnostic codes were used extensively in algorithm development with the assumption that billing and problem list practices accurately reflect the provided care. In jurisdictional settings where ICD-based billing was not recorded directly in the EMR system during patient care (eg, inpatient care in Alberta), such assumptions could not be made and influenced the algorithm development process. Recognizing similarities and differences in data collection strategies, extraction, data release protocols, and existing clinical pathways is critical and will inform algorithm development strategies. ML and NLP techniques are increasingly being adopted in phenotyping algorithms. This is a testament to the fact that detailed records, available from free-text data, can assist with building high-performance classification algorithms.

## Data Extraction, Validity, and Quality

Developing data-driven case finding algorithms is not feasible without electronic data [294]. However, EMR data are not always easy to work with [295], as they are primarily intended to support clinical practice rather than research. EMR settings influence data collection and extraction strategies. Inpatient facilities often set up electronic data warehouses where EMR data are collected into centralized repositories, including free-text data. Primary care settings, in contrast, have variations in their systems, and studies based on primary care data often only use more common data elements such as laboratory data and demographics for multisite studies. Free-text data are less available when compared with inpatient facilities. Primary care clinics, including specialist clinics, are privately operated in many jurisdictions, whereas inpatient care may be publicly or privately operated. These different entities may not always be required to share health data or may have different data management protocols. These considerations influenced the algorithm development process, and a stark contrast in the used data elements can be observed between algorithms developed in outpatient and inpatient settings. To mitigate some of these issues, researchers conducted data linkage between data sources to expand the scope of the available data.

In addition, significant changes in the terminology and coding standards and practices in EMRs have occurred and are actively occurring. This often makes it difficult or impossible to compare or share algorithms developed for different EMR systems using different coding standards (eg, ICD-9, ICD-10, Read, SNOMED RT, SNOMED CT, and MEDCIN for diagnostic codes). Furthermore, many investigators noted that their studies were based on data from a single center, as they did not have access to external EMR data outside of their own institution. Thus, the potential lack of generalizability was a limitation for some studies. However, algorithms developed using commonly available data elements were often externally validated in multiple studies. In particular, simpler algorithms involving diagnostic codes or laboratory data appeared to be externally

validated more commonly. This trend was observed in diabetes and rheumatic conditions and occurred mostly in the United States.

Variation in reported metrics (eg, sensitivity, specificity, positive predictive value, negative predictive value, area under the receiver-operator characteristic curve, and F1 score) was observed in the identified literature. Standardized metrics used in health care should be reported, including sensitivity, specificity, positive predictive value, and negative predictive value. As there is a trade-off between sensitivity and positive predictive value and both are important, it is also useful to report the F1 score, which is the harmonic mean of these 2 quantities. In addition, as class imbalance is frequently a problem in the context of disease classification, with positive instances far less common than negative instances, studies are encouraged to report metrics that account for this, such as area under the precision-recall curve [296]. At present, there are no universally accepted EMR data quality assessment metrics available, although there are various proposed data quality assessment frameworks [297]. Data quality must be assessed based on the suitability of the data to achieve a specific research objective or downstream task. We discuss this later in the recommendations.

## Limitations

This study is not without limitations. First, it is possible that our search did not encompass all qualifying articles in the field. However, our search strategy was refined and improved by systematic review search experts and librarians, and we believe our search successfully captured a broad spectrum of articles on the Charlson conditions. Second, manual screening was carried out by one individual. The objectivity of the review may have been increased by including a second reviewer. Finally, our review did not discuss methods employed for assessing EMR data quality, which depends on the context and clinical application, and is a difficult concept to measure in general. To date, there is no universally accepted data quality metric developed for EMR data, and few of the papers in this review discuss whether or how data quality was assessed in their study. Further research is required to establish the scope of practice for EMR data quality assessment.

## Recommendations on the Basis of Findings

Our review identified that case phenotyping algorithms depends on the health delivery system and disease context. We present a few observed strategies to assist with refining phenotype case definitions using the following key strategies: (1) understanding the health system structure and setting (eg, outpatient vs inpatient, coding practice) will provide a general sense of the type of EMR data that may be available; (2) considering data linkage can increase the scope of data available for algorithm development, it is important to recognize that data may not be standardized or comparable between different data sources. Additional data processing such as data recoding or data imputation may be needed; (3) identifying the relevant clinical and/or health services pathway and involving respective specialty physicians and other stakeholders as part of the algorithm development process can assist with knowledge translation; 4) employing a common data model (eg,

observational medical outcomes partnership [298]) and using commonly available data elements to the possible extent can encourage widespread deployment and external validation. A common data model may differ between disease disciplines and health system areas; and (5) considering how to customize the algorithm to the needs of the end user. The needs are largely divided into clinical decision support through risk adjustment analysis, population-scale disease identification for public health initiatives, or developing methodologies to improve algorithm performance.

Health care is a unique environment, and a one-size-fits-all approach may not be appropriate. This review identified variations in EMR phenotyping, which were heavily influenced by the health care delivery setting and the disease context. To optimize performance, researchers should develop tailored algorithms that focus on the specific population of interest and the particular structure of the health system (eg, developing a primary care diabetes definition), while accounting for data issues such as variations in coding systems, clinical practice guidelines, and data quality. Once a locally developed algorithm is in place, health systems may consider implementing their case finding algorithms on standardized data models. This review identified several studies that either validated previously validated case definitions in a new setting or were refined to appropriately identify disease patients within a new setting. Having locally developed algorithms converted to standard data models will facilitate external validation and implementation, which can otherwise be a critical roadblock to the adoption of these algorithms, allowing for improved algorithm interoperability between health care systems.

The interoperability of algorithms across systems facilitates implementation within existing real-time clinical decision support systems. Easy access to developed code is also critical in validating and replicating published algorithms, after their computability has been confirmed. Analytical code and resources could be shared publicly (eg, on GitHub) to allow access for validation and implementation. The eMERGE consortium [293], CALIBER [299], and Canadian Primary Care Sentinel Surveillance Network [300], for example, have made their algorithms publicly available and have been widely adopted.

## Conclusions

We assessed EMR-based phenotyping of the Charlson conditions in health care settings. The phenotyping algorithms were locally developed and tailored to the needs and objectives of the individual studies. The health system structure and disease context determined data availability and type. The disease context dictated the common data types used for algorithm development. NLP with free-text data was employed for complex diseases that were difficult to identify with algorithms using readily available structured data. Supervised ML was employed in phenotyping algorithms, where applicable, which worked with reference standards obtained from medical chart review. Studies are encouraged to report standard health system metrics and metrics that account for class imbalance. Locally developed algorithms were validated or refined for adoption in the new setting. Locally developed disease- and setting-specific algorithms could be translated into a common data model for easier interoperability of algorithms across systems. Integrating EMR phenotyping algorithms within a health system could lead to the development of a clinical decision support system that makes use of refined existing risk adjustment scoring for risk stratification in clinical point-of-care and inform the public health and health system decision-making process, thus, leading to learning health systems.

## Conflicts of Interest

None declared.

Multimedia Appendix 1
Developed search terms (Medical Subject Headings) for scoping literature review.
[DOCX File , 17 KB - medinform_v9i2e23934_app1.docx ]

Multimedia Appendix 2
Embase and Medical Literature Analysis and Retrieval System Online search results of Charlson terms.
[DOCX File , 22 KB - medinform_v9i2e23934_app2.docx ]

Multimedia Appendix 3
Preferred Reporting Items for Systematic reviews and Meta-analyses flow diagram.
[PNG File , 171 KB - medinform_v9i2e23934_app3.png ]

Multimedia Appendix 4
Summary spreadsheet of identified articles between January 2000 and April 2020.
[XLSX File (Microsoft Excel File), 208 KB - medinform_v9i2e23934_app4.xlsx ]

## References

1. Jameson JL, Longo DL. Precision medicine--personalized, problematic, and promising. N Engl J Med 2015 Jun 04;372(23):2229-2234. [doi: 10.1056/NEJMsb1503104] [Medline: 26014593]

2. Adler-Milstein J, Jha AK. HITECH Act Drove Large Gains In Hospital Electronic Health Record Adoption. Health Aff (Millwood) 2017 Aug 01;36(8):1416-1422. [doi: 10.1377/hlthaff.2016.1651] [Medline: 28784734]

3. Gagnon M, Payne-Gagnon J, Breton E, Fortin J, Khoury L, Dolovich L, et al. Adoption of Electronic Personal Health Records in Canada: Perceptions of Stakeholders. Int J Health Policy Manag 2016 Jul 01;5(7):425-433 [FREE Full text] [doi: 10.15171/ijhpm.2016.36] [Medline: 27694670]

4. Valderas JM, Starfield B, Sibbald B, Salisbury C, Roland M. Defining comorbidity: implications for understanding health and health services. Ann Fam Med 2009;7(4):357-363 [FREE Full text] [doi: 10.1370/afm.983] [Medline: 19597174]

5. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40(5):373-383. [Medline: 3558716]

6. Quan H, Li B, Couris CM, Fushimi K, Graham P, Hider P, et al. Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. Am J Epidemiol 2011 Mar 15;173(6):676-682. [doi: 10.1093/aje/kwq433] [Medline: 21330339]

7. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi J, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care 2005 Nov;43(11):1130-1139. [Medline: 16224307]

8. Romano PS, Roos LL, Jollis JG. Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: differing perspectives. J Clin Epidemiol 1993 Oct;46(10):1075-9; discussion 1081. [Medline: 8410092]

9. McBrien KA, Souri S, Symonds NE, Rouhi A, Lethebe BC, Williamson TS, et al. Identification of validated case definitions for medical conditions used in primary care electronic medical record databases: a systematic review. J Am Med Inform Assoc 2018 Nov 01;25(11):1567-1578. [doi: 10.1093/jamia/ocy094] [Medline: 30137498]

10. Nissen F, Quint JK, Wilkinson S, Mullerova H, Smeeth L, Douglas IJ. Validation of asthma recording in electronic health records: a systematic review. Clin Epidemiol 2017;9:643-656 [FREE Full text] [doi: 10.2147/CLEP.S143718] [Medline: 29238227]

11. Wang P, Garza M, Zozus M. Cancer Phenotype Development: A Literature Review. Stud Health Technol Inform 2019;257:468-472. [Medline: 30741241]

12. Xu J, Rasmussen LV, Shaw PL, Jiang G, Kiefer RC, Mo H, et al. Review and evaluation of electronic health records-driven phenotype algorithm authoring tools for clinical and translational research. J Am Med Inform Assoc 2015 Dec;22(6):1251-1260 [FREE Full text] [doi: 10.1093/jamia/ocv070] [Medline: 26224336]

13. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med 2018 Oct 02;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]

14. EndNote Version X8. https://endnote.com. Philadelphia, PA: Clarivate; 2013.

15. STATA SSR14. College Station, TX: StataCorp LLC; 2015. URL: https://www.stata.com

16. Ammann EM, Schweizer ML, Robinson JG, Eschol JO, Kafa R, Girotra S, et al. Chart validation of inpatient ICD-9-CM administrative diagnosis codes for acute myocardial infarction (AMI) among intravenous immune globulin (IGIV) users in the Sentinel Distributed Database. Pharmacoepidemiol Drug Saf 2018 Apr;27(4):398-404 [FREE Full text] [doi: 10.1002/pds.4398] [Medline: 29446185]

17. Ando T, Ooba N, Mochizuki M, Koide D, Kimura K, Lee SL, et al. Positive predictive value of ICD-10 codes for acute myocardial infarction in Japan: a validation study at a single center. BMC Health Serv Res 2018 Dec 26;18(1):895 [FREE Full text] [doi: 10.1186/s12913-018-3727-0] [Medline: 30477501]

18. Backenroth D, Chase H, Friedman C, Wei Y. Using Rich Data on Comorbidities in Case-Control Study Design with Electronic Health Record Data Improves Control of Confounding in the Detection of Adverse Drug Reactions. PLoS One 2016;11(10):e0164304 [FREE Full text] [doi: 10.1371/journal.pone.0164304] [Medline: 27716785]

19. Bent-Ennakhil N, Cécile Périer M, Sobocki P, Gothefors D, Johansson G, Milea D, et al. Incidence of cardiovascular diseases and type-2-diabetes mellitus in patients with psychiatric disorders. Nord J Psychiatry 2018 Oct;72(7):455-461. [doi: 10.1080/08039488.2018.1463392] [Medline: 30513230]

20. Bjerking LH, Hansen KW, Madsen M, Jensen JS, Madsen JK, Sørensen R, et al. Use of diagnostic coronary angiography in women and men presenting with acute myocardial infarction: a matched cohort study. BMC Cardiovasc Disord 2016 Jun 01;16:120 [FREE Full text] [doi: 10.1186/s12872-016-0248-9] [Medline: 27250115]

21. Coloma PM, Valkhoff VE, Mazzaglia G, Nielsson MS, Pedersen L, Molokhia M, EU-ADR Consortium. Identification of acute myocardial infarction from electronic healthcare records using different disease coding systems: a validation study in three European countries. BMJ Open 2013 Jun 20;3(6) [FREE Full text] [doi: 10.1136/bmjopen-2013-002862] [Medline: 23794587]

22. Cross DS, McCarty CA, Steinhubl SR, Carey DJ, Erlich PM. Development of a multi-institutional cohort to facilitate cardiovascular disease biomarker validation using existing biorepository samples linked to electronic health records. Clin Cardiol 2013 Aug;36(8):486-491 [FREE Full text] [doi: 10.1002/clc.22146] [Medline: 23740530]

23. Findlay I, Morris T, Zhang R, McCowan C, Shield S, Forbes B, et al. Linking hospital patient records for suspected or established acute coronary syndrome in a complex secondary care system: a proof-of-concept e-registry in National Health

Service Scotland. Eur Heart J Qual Care Clin Outcomes 2018 Jul 01;4(3):155-167. [doi: 10.1093/ehjqcco/qcy007] [Medline: 29462281]

24. FitzHenry F, Murff HJ, Matheny ME, Gentry N, Fielstein EM, Brown SH, et al. Exploring the frontier of electronic health record surveillance: the case of postoperative complications. Med Care 2013 Jul;51(6):509-516 [FREE Full text] [doi: 10.1097/MLR.0b013e31828d1210] [Medline: 23673394]

25. Floyd JS, Blondon M, Moore KP, Boyko EJ, Smith NL. Validation of methods for assessing cardiovascular disease using electronic health data in a cohort of Veterans with diabetes. Pharmacoepidemiol Drug Saf 2016 May;25(4):467-471 [FREE Full text] [doi: 10.1002/pds.3921] [Medline: 26555025]

26. Goldstein BA, Assimes T, Winkelmayer WC, Hastie T. Detecting clinically meaningful biomarkers with repeated measurements: An illustration with electronic health records. Biometrics 2015 Jul;71(2):478-486 [FREE Full text] [doi: 10.1111/biom.12283] [Medline: 25652566]

27. Herrett E, Bhaskaran K, Timmis A, Denaxas S, Hemingway H, Smeeth L. Association between clinical presentations before myocardial infarction and coronary mortality: a prospective population-based study using linked electronic records. Eur Heart J 2014 Oct 14;35(35):2363-2371 [FREE Full text] [doi: 10.1093/eurheartj/ehu286] [Medline: 25038774]

28. Hivert M, Grant RW, Shrader P, Meigs JB. Identifying primary care patients at risk for future diabetes and cardiovascular disease using electronic health records. BMC Health Serv Res 2009 Sep 22;9:170 [FREE Full text] [doi: 10.1186/1472-6963-9-170] [Medline: 19772639]

29. Mahler SA, Lenoir KM, Wells BJ, Burke GL, Duncan PW, Case LD, et al. Safely Identifying Emergency Department Patients With Acute Chest Pain for Early Discharge. Circulation 2018 Nov 27;138(22):2456-2468 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.118.036528] [Medline: 30571347]

30. Manemann SM, Gerber Y, Chamberlain AM, Dunlay SM, Bell MR, Jaffe AS, et al. Acute coronary syndromes in the community. Mayo Clin Proc 2015 May;90(5):597-605. [doi: 10.1016/j.mayocp.2015.02.015] [Medline: 25794453]

31. Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 2011 Aug 24;306(8):848-855. [doi: 10.1001/jama.2011.1204] [Medline: 21862746]

32. Persell SD, Dunne AP, Lloyd-Jones DM, Baker DW. Electronic health record-based cardiac risk assessment and identification of unmet preventive needs. Med Care 2009 May;47(4):418-424. [doi: 10.1097/MLR.0b013e31818dce21] [Medline: 19238100]

33. Reynolds K, Go AS, Leong TK, Boudreau DM, Cassidy-Bushrow AE, Fortmann SP, et al. Trends in Incidence of Hospitalized Acute Myocardial Infarction in the Cardiovascular Research Network (CVRN). Am J Med 2017 Mar;130(3):317-327 [FREE Full text] [doi: 10.1016/j.amjmed.2016.09.014] [Medline: 27751900]

34. Song W, Huang H, Zhang CZ, Bates DW, Wright A. Using whole genome scores to compare three clinical phenotyping methods in complex diseases. Sci Rep 2018 Jul 27;8(1):11360 [FREE Full text] [doi: 10.1038/s41598-018-29634-w] [Medline: 30054501]

35. Tien M, Kashyap R, Wilson GA, Hernandez-Torres V, Jacob AK, Schroeder DR, et al. Retrospective Derivation and Validation of an Automated Electronic Search Algorithm to Identify Post Operative Cardiovascular and Thromboembolic Complications. Appl Clin Inform 2015;6(3):565-576 [FREE Full text] [doi: 10.4338/ACI-2015-03-RA-0026] [Medline: 26448798]

36. Torabi A, Cleland JGF, Sherwi N, Atkin P, Panahi H, Kilpatrick E, et al. Influence of case definition on incidence and outcome of acute coronary syndromes. Open Heart 2016;3(2):e000487 [FREE Full text] [doi: 10.1136/openhrt-2016-000487] [Medline: 28123755]

37. Wang N, Li T, Du Q. Risk factors of upper gastrointestinal hemorrhage with acute coronary syndrome. Am J Emerg Med 2019 Apr;37(4):615-619. [doi: 10.1016/j.ajem.2018.06.055] [Medline: 30381146]

38. Zheng J, Yarzebski J, Ramesh BP, Goldberg RJ, Yu H. Automatically Detecting Acute Myocardial Infarction Events from EHR Text: A Preliminary Study. AMIA Annu Symp Proc 2014;2014:1286-1293 [FREE Full text] [Medline: 25954440]

39. Bielinski SJ, Pathak J, Carrell DS, Takahashi PY, Olson JE, Larson NB, et al. A Robust e-Epidemiology Tool in Phenotyping Heart Failure with Differentiation for Preserved and Reduced Ejection Fraction: the Electronic Medical Records and Genomics (eMERGE) Network. J Cardiovasc Transl Res 2015 Dec;8(8):475-483 [FREE Full text] [doi: 10.1007/s12265-015-9644-2] [Medline: 26195183]

40. Blecker S, Katz SD, Horwitz LI, Kuperman G, Park H, Gold A, et al. Comparison of Approaches for Heart Failure Case Identification From Electronic Health Record Data. JAMA Cardiol 2016 Dec 01;1(9):1014-1020 [FREE Full text] [doi: 10.1001/jamacardio.2016.3236] [Medline: 27706470]

41. Bosch L, Assmann P, de Grauw WJC, Schalk BWM, Biermans MCJ. Heart failure in primary care: prevalence related to age and comorbidity. Prim Health Care Res Dev 2019 Jul 29;20:e79 [FREE Full text] [doi: 10.1017/S1463423618000889] [Medline: 31868152]

42. Bosco-Lévy P, Duret S, Picard F, Dos Santos P, Puymirat E, Gilleron V, et al. Diagnostic accuracy of the International Classification of Diseases, Tenth Revision, codes of heart failure in an administrative database. Pharmacoepidemiol Drug Saf 2019 Feb;28(2):194-200. [doi: 10.1002/pds.4690] [Medline: 30395375]

43.  Byrd RJ, Steinhubl SR, Sun J, Ebadollahi S, Stewart WF. Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. Int J Med Inform 2014 Dec;83(12):983-992 [FREE Full text] [doi: 10.1016/j.ijmedinf.2012.12.005] [Medline: 23317809]

44.  Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc 2017 Dec 01;24(2):361-370 [FREE Full text] [doi: 10.1093/jamia/ocw112] [Medline: 27521897]

45.  Dai W, Brisimi TS, Adams WG, Mela T, Saligrama V, Paschalidis IC. Prediction of hospitalization due to heart diseases by supervised learning methods. Int J Med Inform 2015 Mar;84(3):189-197 [FREE Full text] [doi: 10.1016/j.ijmedinf.2014.10.002] [Medline: 25497295]

46.  Evans RS, Benuzillo J, Horne BD, Lloyd JF, Bradshaw A, Budge D, et al. Automated identification and predictive tools to help identify high-risk heart failure patients: pilot evaluation. J Am Med Inform Assoc 2016 Sep;23(5):872-878. [doi: 10.1093/jamia/ocv197] [Medline: 26911827]

47.  Frigaard M, Rubinsky A, Lowell L, Malkina A, Karliner L, Kohn M, et al. Validating laboratory defined chronic kidney disease in the electronic health record for patients in primary care. BMC Nephrol 2019 Jan 03;20(1):3 [FREE Full text] [doi: 10.1186/s12882-018-1156-2] [Medline: 30606109]

48.  Gini R, Schuemie MJ, Mazzaglia G, Lapi F, Francesconi P, Pasqua A, et al. Automatic identification of type 2 diabetes, hypertension, ischaemic heart disease, heart failure and their levels of severity from Italian General Practitioners' electronic medical records: a validation study. BMJ Open 2016 Dec 09;6(12):e012413 [FREE Full text] [doi: 10.1136/bmjopen-2016-012413] [Medline: 27940627]

49.  Huusko J, Purmonen T, Toppila I, Lassenius M, Ukkonen H. Real-world clinical diagnostics of heart failure patients with reduced or preserved ejection fraction. ESC Heart Fail 2020 Jul;7(3):1039-1048 [FREE Full text] [doi: 10.1002/ehf2.12665] [Medline: 32187879]

50.  Jonnalagadda SR, Adupa AK, Garg RP, Corona-Cox J, Shah SJ. Text Mining of the Electronic Health Record: An Information Extraction Approach for Automated Identification and Subphenotyping of HFpEF Patients for Clinical Trials. J Cardiovasc Transl Res 2017 Jul;10(3):313-321. [doi: 10.1007/s12265-017-9752-2] [Medline: 28585184]

51.  Kaspar M, Fette G, Güder G, Seidlmayer L, Ertl M, Dietrich G, et al. Underestimated prevalence of heart failure in hospital inpatients: a comparison of ICD codes and discharge letter information. Clin Res Cardiol 2018 Oct;107(9):778-787 [FREE Full text] [doi: 10.1007/s00392-018-1245-z] [Medline: 29667017]

52.  Koudstaal S, Pujades-Rodriguez M, Denaxas S, Gho JMIH, Shah AD, Yu N, et al. Prognostic burden of heart failure recorded in primary care, acute hospital admissions, or both: a population-based linked electronic health record cohort study in 2.1 million people. Eur J Heart Fail 2017 Sep;19(9):1119-1127. [doi: 10.1002/ejhf.709] [Medline: 28008698]

53.  Kurgansky KE, Schubert P, Parker R, Djousse L, Riebman JB, Gagnon DR, et al. Association of pulse rate with outcomes in heart failure with reduced ejection fraction: a retrospective cohort study. BMC Cardiovasc Disord 2020 Feb 26;20(1):92 [FREE Full text] [doi: 10.1186/s12872-020-01384-6] [Medline: 32101141]

54.  Lindmark K, Boman K, Olofsson M, Törnblom M, Levine A, Castelo-Branco A, et al. Epidemiology of heart failure and trends in diagnostic work-up: a retrospective, population-based cohort study in Sweden. Clin Epidemiol 2019;11:231-244. [doi: 10.2147/CLEP.S170873] [Medline: 30962724]

55.  Magnusson P, Palm A, Branden E, Mörner S. Misclassification of hypertrophic cardiomyopathy: validation of diagnostic codes. Clin Epidemiol 2017;9:403-410 [FREE Full text] [doi: 10.2147/CLEP.S139300] [Medline: 28848363]

56.  Panahiazar M, Taslimitehrani V, Pereira N, Pathak J. Using EHRs and Machine Learning for Heart Failure Survival Analysis. Stud Health Technol Inform 2015;216:40-44 [FREE Full text] [Medline: 26262006]

57.  Navaneethan SD, Jolly SE, Schold JD, Arrigain S, Saupe W, Sharp J, et al. Development and validation of an electronic health record-based chronic kidney disease registry. Clin J Am Soc Nephrol 2011 Jan;6(1):40-49. [doi: 10.2215/CJN.04230510] [Medline: 21051745]

58.  Ng K, Steinhubl SR, deFilippi C, Dey S, Stewart WF. Early Detection of Heart Failure Using Electronic Health Records: Practical Implications for Time Before Diagnosis, Data Diversity, Data Quantity, and Data Density. Circ Cardiovasc Qual Outcomes 2016 Nov;9(6):649-658. [doi: 10.1161/CIRCOUTCOMES.116.002797] [Medline: 28263940]

59.  Pakhomov S, Weston SA, Jacobsen SJ, Chute CG, Meverden R, Roger VL. Electronic medical records for clinical research: application to the identification of heart failure. Am J Manag Care 2007 Jun;13(6 Part 1):281-288 [FREE Full text] [Medline: 17567225]

60.  Patel YR, Robbins JM, Kurgansky KE, Imran T, Orkaby AR, McLean RR, et al. Development and validation of a heart failure with preserved ejection fraction cohort using electronic medical records. BMC Cardiovasc Disord 2018 Jun 28;18(1):128. [doi: 10.1186/s12872-018-0866-5] [Medline: 29954337]

61.  Pike MM, Decker PA, Larson NB, St Sauver JL, Takahashi PY, Roger VL, et al. Improvement in Cardiovascular Risk Prediction with Electronic Health Records. J Cardiovasc Transl Res 2016 Jun;9(3):214-222 [FREE Full text] [doi: 10.1007/s12265-016-9687-z] [Medline: 26960568]

62.  Rasmy L, Wu Y, Wang N, Geng X, Zheng WJ, Wang F, et al. A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. J Biomed Inform 2018 Aug;84:11-16 [FREE Full text] [doi: 10.1016/j.jbi.2018.06.011] [Medline: 29908902]

XSL•FO

RenderX

63. Shameer K, Johnson K, Yahi A. Predictive Modeling of Hospital Readmission Rates Using Electronic Medical Record-Wide Machine Learning: A Case-Study Using Mount Sinai Heart Failure Cohort. Pacific Symposium on Biocomputing;Pacific Symposium on Biocomputing 2016:22. [doi: 10.1142/9789813207813_0027] [Medline: 27896982]

64. Stålhammar J, Stern L, Linder R, Sherman S, Parikh R, Ariely R, et al. The burden of preserved ejection fraction heart failure in a real-world Swedish patient population. J Med Econ 2014 Jan;17(1):43-51. [doi: 10.3111/13696998.2013.848808] [Medline: 24066839]

65. Sun J, Hu J, Luo D, Markatou M, Wang F, Edabollahi S, et al. Combining knowledge and data driven insights for identifying risk factors using electronic health records. AMIA Annu Symp Proc 2012;2012:901-910 [FREE Full text] [Medline: 23304365]

66. Taslimitehrani V, Dong G, Pereira NL, Panahiazar M, Pathak J. Developing EHR-driven heart failure risk prediction models using CPXR(Log) with the probabilistic loss function. J Biomed Inform 2016 Apr;60:260-269. [doi: 10.1016/j.jbi.2016.01.009] [Medline: 26844760]

67. Thomas IC, Nishimura M, Ma J, Dickson SD, Alshawabkeh L, Adler E, et al. Clinical Characteristics and Outcomes of Patients With Heart Failure and Methamphetamine Abuse. J Card Fail 2020 Mar;26(3):202-209. [doi: 10.1016/j.cardfail.2019.10.002] [Medline: 31618697]

68. Tison GH, Chamberlain AM, Pletcher MJ, Dunlay SM, Weston SA, Killian JM, et al. Identifying heart failure using EMR-based algorithms. Int J Med Inform 2018 Dec;120:1-7 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.09.016] [Medline: 30409334]

69. Vijayakrishnan R, Steinhubl SR, Ng K, Sun J, Byrd RJ, Daar Z, et al. Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record. J Card Fail 2014 Jul;20(7):459-464 [FREE Full text] [doi: 10.1016/j.cardfail.2014.03.008] [Medline: 24709663]

70. Wang Y, Luo J, Hao S, Xu H, Shin AY, Jin B, et al. NLP based congestive heart failure case finding: A prospective analysis on statewide electronic medical records. Int J Med Inform 2015 Dec;84(12):1039-1047. [doi: 10.1016/j.ijmedinf.2015.06.007] [Medline: 26254876]

71. Wang Y, Ng K, Byrd R. Early detection of heart failure with varying prediction windows by structured and unstructured data in electronic health records. Conference proceedings : ;Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference 2015:2530-2533. [doi: 10.1109/embc.2015.7318907] [Medline: 26736807]

72. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care 2010 Jun;48(6 Suppl):S106-S113. [doi: 10.1097/MLR.0b013e3181de9e17] [Medline: 20473190]

73. Xu Y, Lee S, Martin E, D'souza AG, Doktorchik CTA, Jiang J, et al. Enhancing ICD-Code-Based Case Definition for Heart Failure Using Electronic Medical Record Data. J Card Fail 2020 Jul;26(7):610-617. [doi: 10.1016/j.cardfail.2020.04.003] [Medline: 32304875]

74. Yang X, Gong Y, Waheed N, March K, Bian J, Hogan WR, et al. Identifying Cancer Patients at Risk for Heart Failure Using Machine Learning Methods. AMIA Annu Symp Proc 2019;2019:933-941 [FREE Full text] [Medline: 32308890]

75. Zhang R, Ma S, Shanahan L, Munroe J, Horn S, Speedie S. Discovering and identifying New York heart association classification from electronic health records. BMC Med Inform Decis Mak 2018 Jul 23;18(Suppl 2):48. [doi: 10.1186/s12911-018-0625-7] [Medline: 30066653]

76. Afzal N, Mallipeddi VP, Sohn S, Liu H, Chaudhry R, Scott CG, et al. Natural language processing of clinical notes for identification of critical limb ischemia. Int J Med Inform 2018 Mar;111:83-89 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.12.024] [Medline: 29425639]

77. Afzal N, Sohn S, Abram S, Scott CG, Chaudhry R, Liu H, et al. Mining peripheral arterial disease cases from narrative clinical notes using natural language processing. J Vasc Surg 2017 Jun;65(6):1753-1761 [FREE Full text] [doi: 10.1016/j.jvs.2016.11.031] [Medline: 28189359]

78. Archangelidi O, Pujades-Rodriguez M, Timmis A, Jouven X, Denaxas S, Hemingway H. Clinically recorded heart rate and incidence of 12 coronary, cardiac, cerebrovascular and peripheral arterial diseases in 233,970 men and women: A linked electronic health record study. Eur J Prev Cardiol 2018 Sep;25(14):1485-1495. [doi: 10.1177/2047487318785228] [Medline: 29966429]

79. Arruda-Olson AM, Afzal N, Priya Mallipeddi V, Said A, Moussa Pacha H, Moon S, et al. Leveraging the Electronic Health Record to Create an Automated Real-Time Prognostic Tool for Peripheral Arterial Disease. J Am Heart Assoc 2018 Dec 04;7(23):e009680 [FREE Full text] [doi: 10.1161/JAHA.118.009680] [Medline: 30571601]

80. Caleyachetty R, Thomas GN, Toulis KA, Mohammed N, Gokhale KM, Balachandran K, et al. Metabolically Healthy Obese and Incident Cardiovascular Disease Events Among 3.5 Million Men and Women. J Am Coll Cardiol 2017 Oct 19;70(12):1429-1437 [FREE Full text] [doi: 10.1016/j.jacc.2017.07.763] [Medline: 28911506]

81. Daskivich T, Abedi G, Kaplan S. Electronic Health Record Problem Lists: Accurate Enough for Risk Adjustment? Am J Manag Care 2018;24(1):A. [Medline: 29350512]

82. Emdin CA, Anderson SG, Callender T, Conrad N, Salimi-Khorshidi G, Mohseni H, et al. Usual blood pressure, peripheral arterial disease, and vascular risk: cohort study of 4.2 million adults. BMJ 2015 Oct 29;351:h4865 [FREE Full text] [doi: 10.1136/bmj.h4865] [Medline: 26419648]

83. George J, Rapsomaniki E, Pujades-Rodriguez M, Shah AD, Denaxas S, Herrett E, et al. How Does Cardiovascular Disease First Present in Women and Men? Incidence of 12 Cardiovascular Diseases in a Contemporary Cohort of 1,937,360 People. Circulation 2015 Oct 06;132(14):1320-1328 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.114.013797] [Medline: 26330414]

84. Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, et al. Electronic medical records for genetic research: results of the eMERGE consortium. Sci Transl Med 2011 Apr 20;3(79):79re1 [FREE Full text] [doi: 10.1126/scitranslmed.3001807] [Medline: 21508311]

85. Kullo IJ, Shameer K, Jouni H, Lesnick TG, Pathak J, Chute CG, et al. The ATXN2-SH2B3 locus is associated with peripheral arterial disease: an electronic medical record-based genome-wide association study. Front Genet 2014;5:166 [FREE Full text] [doi: 10.3389/fgene.2014.00166] [Medline: 25009551]

86. Man A, Zhu Y, Zhang Y, Dubreuil M, Rho YH, Peloquin C, et al. The risk of cardiovascular disease in systemic sclerosis: a population-based cohort study. Ann Rheum Dis 2013 Jul;72(7):1188-1193 [FREE Full text] [doi: 10.1136/annrheumdis-2012-202007] [Medline: 22904260]

87. Ross EG, Jung K, Dudley JT, Li L, Leeper NJ, Shah NH. Predicting Future Cardiovascular Events in Patients With Peripheral Artery Disease Using Electronic Health Record Data. Circ Cardiovasc Qual Outcomes 2019 Mar;12(3):e004741 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.118.004741] [Medline: 30857412]

88. Savova GK, Fan J, Ye Z, Murphy SP, Zheng J, Chute CG, et al. Discovering peripheral arterial disease cases from radiology notes using natural language processing. AMIA Annu Symp Proc 2010 Nov 13;2010:722-726 [FREE Full text] [Medline: 21347073]

89. Wolfson J, Vock DM, Bandyopadhyay S, Kottke T, Vazquez-Benitez G, Johnson P, et al. Use and Customization of Risk Scores for Predicting Cardiovascular Events Using Electronic Health Record Data. J Am Heart Assoc 2017 May 24;6(4) [FREE Full text] [doi: 10.1161/JAHA.116.003670] [Medline: 28438733]

90. Ammann EM, Leira EC, Winiecki SK, Nagaraja N, Dandapat S, Carnahan RM, et al. Chart validation of inpatient ICD-9-CM administrative diagnosis codes for ischemic stroke among IGIV users in the Sentinel Distributed Database. Medicine (Baltimore) 2017 Dec;96(52):e9440 [FREE Full text] [doi: 10.1097/MD.0000000000009440] [Medline: 29384925]

91. Bell S, Daskalopoulou M, Rapsomaniki E, George J, Britton A, Bobak M, et al. Association between clinically recorded alcohol consumption and initial presentation of 12 cardiovascular diseases: population based cohort study using linked health records. BMJ 2017 Mar 22;356:j909. [doi: 10.1136/bmj.j909] [Medline: 28331015]

92. Castro VM, Dligach D, Finan S, Yu S, Can A, Abd-El-Barr M, et al. Large-scale identification of patients with cerebral aneurysms using natural language processing. Neurology 2017 Jan 10;88(2):164-168 [FREE Full text] [doi: 10.1212/WNL.0000000000003490] [Medline: 27927935]

93. Esteban S, Rodríguez Tablado M, Ricci RI, Terrasa S, Kopitowski K. A rule-based electronic phenotyping algorithm for detecting clinically relevant cardiovascular disease cases. BMC Res Notes 2017 Jul 14;10(1):281 [FREE Full text] [doi: 10.1186/s13104-017-2600-2] [Medline: 28705240]

94. Garg R, Oh E, Naidech A, Kording K, Prabhakaran S. Automating Ischemic Stroke Subtype Classification Using Machine Learning and Natural Language Processing. J Stroke Cerebrovasc Dis 2019 Jul;28(7):2045-2051. [doi: 10.1016/j.jstrokecerebrovasdis.2019.02.004] [Medline: 31103549]

95. Gon Y, Kabata D, Yamamoto K, Shintani A, Todo K, Mochizuki H, et al. Validation of an algorithm that determines stroke diagnostic code accuracy in a Japanese hospital-based cancer registry using electronic medical records. BMC Med Inform Decis Mak 2017 Dec 04;17(1):157 [FREE Full text] [doi: 10.1186/s12911-017-0554-x] [Medline: 29202795]

96. Gulliford MC, Charlton J, Ashworth M, Rudd AG, Toschke AM, eCRT Research Team. Selection of medical diagnostic codes for analysis of electronic patient records. Application to stroke in a primary care database. PLoS One 2009 Oct 24;4(9):e7168 [FREE Full text] [doi: 10.1371/journal.pone.0007168] [Medline: 19777060]

97. Imran TF, Posner D, Honerlaw J, Vassy JL, Song RJ, Ho Y, et al. A phenotyping algorithm to identify acute ischemic stroke accurately from a national biobank: the Million Veteran Program. Clin Epidemiol 2018;10:1509-1521 [FREE Full text] [doi: 10.2147/CLEP.S160764] [Medline: 30425582]

98. Kivimäki M, Batty GD, Singh-Manoux A, Britton A, Brunner EJ, Shipley MJ. Validity of Cardiovascular Disease Event Ascertainment Using Linkage to UK Hospital Records. Epidemiology 2017 Sep;28(5):735-739 [FREE Full text] [doi: 10.1097/EDE.0000000000000688] [Medline: 28570383]

99. Kogan E, Twyman K, Heap J, Milentijevic D, Lin JH, Alberts M. Assessing stroke severity using electronic health record data: a machine learning approach. BMC Med Inform Decis Mak 2020 Jan 08;20(1):8 [FREE Full text] [doi: 10.1186/s12911-019-1010-x] [Medline: 31914991]

100. Kreuger AL, Middelburg RA, Beckers EAM, de Vooght KMK, Zwaginga JJ, Kerkhoffs JH, et al. The identification of cases of major hemorrhage during hospitalization in patients with acute leukemia using routinely recorded healthcare data. PLoS One 2018;13(8):e0200655 [FREE Full text] [doi: 10.1371/journal.pone.0200655] [Medline: 30110326]

101. Ljubisavljevic S, Milosevic V, Stojanov A, Ljubisavljevic M, Dunjic O, Zivkovic M. Identification of clinical and paraclinical findings predictive for headache occurrence during spontaneous subarachnoid hemorrhage. Clin Neurol Neurosurg 2017 Jul;158:40-45. [doi: 10.1016/j.clineuro.2017.04.017] [Medline: 28458057]

XSL•FO

RenderX

102. Ni Y, Alwell K, Moomaw CJ, Woo D, Adeoye O, Flaherty ML, et al. Towards phenotyping stroke: Leveraging data from a large-scale epidemiological study to detect stroke diagnosis. PLoS One 2018;13(2):e0192586 [FREE Full text] [doi: 10.1371/journal.pone.0192586] [Medline: 29444182]

103. Øie LR, Madsbu MA, Giannadakis C, Vorhaug A, Jensberg H, Salvesen , et al. Validation of intracranial hemorrhage in the Norwegian Patient Registry. Brain Behav 2018 Feb;8(2):e00900 [FREE Full text] [doi: 10.1002/brb3.900] [Medline: 29484261]

104. Oostema JA, Konen J, Chassee T, Nasiri M, Reeves MJ. Clinical predictors of accurate prehospital stroke recognition. Stroke 2015 Jun;46(6):1513-1517. [doi: 10.1161/STROKEAHA.115.008650] [Medline: 25922507]

105. Pouwels KB, Voorham J, Hak E, Denig P. Identification of major cardiovascular events in patients with diabetes using primary care data. BMC Health Serv Res 2016 May 02;16:110 [FREE Full text] [doi: 10.1186/s12913-016-1361-2] [Medline: 27038959]

106. Weinstein R, Ess K, Sirdar B, Song S, Cutting S. Primary Intraventricular Hemorrhage: Clinical Characteristics and Outcomes. J Stroke Cerebrovasc Dis 2017 May;26(5):995-999. [doi: 10.1016/j.jstrokecerebrovasdis.2016.11.114] [Medline: 28111008]

107. Wheater E, Mair G, Sudlow C, Alex B, Grover C, Whiteley W. A validated natural language processing algorithm for brain imaging phenotypes from radiology reports in UK electronic health records. BMC Med Inform Decis Mak 2019 Sep 09;19(1):184. [doi: 10.1186/s12911-019-0908-7] [Medline: 31500613]

108. Amra S, O'Horo JC, Singh TD, Wilson GA, Kashyap R, Petersen R, et al. Derivation and validation of the automated search algorithms to identify cognitive impairment and dementia in electronic health records. J Crit Care 2017 Feb;37:202-205. [doi: 10.1016/j.jcrc.2016.09.026] [Medline: 27969571]

109. Anzaldi LJ, Davison A, Boyd CM, Leff B, Kharrazi H. Comparing clinician descriptions of frailty and geriatric syndromes using electronic health records: a retrospective cohort study. BMC Geriatr 2017 Dec 25;17(1):248 [FREE Full text] [doi: 10.1186/s12877-017-0645-7] [Medline: 29070036]

110. Barnes DE, Zhou J, Walker RL, Larson EB, Lee SJ, Boscardin WJ, et al. Development and Validation of eRADAR: A Tool Using EHR Data to Detect Unrecognized Dementia. J Am Geriatr Soc 2020 Jan;68(1):103-111. [doi: 10.1111/jgs.16182] [Medline: 31612463]

111. Boustani M, Perkins AJ, Khandker RK, Duong S, Dexter PR, Lipton R, et al. Passive Digital Signature for Early Identification of Alzheimer's Disease and Related Dementia. J Am Geriatr Soc 2020 Mar;68(3):511-518. [doi: 10.1111/jgs.16218] [Medline: 31784987]

112. Corradi JP, Chhabra J, Mather JF, Waszynski CM, Dicks RS. Analysis of multi-dimensional contemporaneous EHR data to refine delirium assessments. Comput Biol Med 2016 Aug 01;75:267-274. [doi: 10.1016/j.compbiomed.2016.06.013] [Medline: 27340924]

113. Ernecoff NC, Wessell KL, Gabriel S, Carey TS, Hanson LC. A Novel Screening Method to Identify Late-Stage Dementia Patients for Palliative Care Research and Practice. J Pain Symptom Manage 2018 Apr;55(4):1152-1158.e1 [FREE Full text] [doi: 10.1016/j.jpainsymman.2017.12.480] [Medline: 29288881]

114. Ford E, Rooney P, Oliver S, Hoile R, Hurley P, Banerjee S, et al. Identifying undetected dementia in UK primary care patients: a retrospective case-control study comparing machine-learning and standard epidemiological approaches. BMC Med Inform Decis Mak 2019 Dec 02;19(1):248 [FREE Full text] [doi: 10.1186/s12911-019-0991-9] [Medline: 31791325]

115. Halpern R, Seare J, Tong J, Hartry A, Olaoye A, Aigbogun MS. Using electronic health records to estimate the prevalence of agitation in Alzheimer disease/dementia. Int J Geriatr Psychiatry 2019 Mar;34(3):420-431 [FREE Full text] [doi: 10.1002/gps.5030] [Medline: 30430642]

116. Jaakkimainen RL, Bronskill SE, Tierney MC, Herrmann N, Green D, Young J, et al. Identification of Physician-Diagnosed Alzheimer's Disease and Related Dementias in Population-Based Administrative Data: A Validation Study Using Family Physicians' Electronic Medical Records. J Alzheimers Dis 2016 Aug 10;54(1):337-349. [doi: 10.3233/JAD-160105] [Medline: 27567819]

117. Kharrazi H, Anzaldi LJ, Hernandez L, Davison A, Boyd CM, Leff B, et al. The Value of Unstructured Electronic Health Record Data in Geriatric Syndrome Case Identification. J Am Geriatr Soc 2018 Aug;66(8):1499-1507. [doi: 10.1111/jgs.15411] [Medline: 29972595]

118. Lewis G, Werbeloff N, Hayes JF, Howard R, Osborn DPJ. Diagnosed depression and sociodemographic factors as predictors of mortality in patients with dementia. Br J Psychiatry 2018 Aug;213(2):471-476 [FREE Full text] [doi: 10.1192/bjp.2018.86] [Medline: 29898791]

119. McCoy TH, Han L, Pellegrini AM, Tanzi RE, Berretta S, Perlis RH. Stratifying risk for dementia onset using large-scale electronic health record data: A retrospective cohort study. Alzheimers Dement 2020 Mar;16(3):531-540. [doi: 10.1016/j.jalz.2019.09.084] [Medline: 31859230]

120. Perera G, Pedersen L, Ansel D, Alexander M, Arrighi HM, Avillach P, et al. Dementia prevalence and incidence in a federation of European Electronic Health Record databases: The European Medical Informatics Framework resource. Alzheimers Dement 2018 Feb;14(2):130-139 [FREE Full text] [doi: 10.1016/j.jalz.2017.06.2270] [Medline: 28734783]

121. Pham TM, Petersen I, Walters K, Raine R, Manthorpe J, Mukadam N, et al. Trends in dementia diagnosis rates in UK ethnic groups: analysis of UK primary care data. Clin Epidemiol 2018;10:949-960 [FREE Full text] [doi: 10.2147/CLEP.S152647] [Medline: 30123007]

122. Ponjoan A, Garre-Olmo J, Blanch J, Fages E, Alves-Cabratosa L, Martí-Lluch R, et al. How well can electronic health records from primary care identify Alzheimer's disease cases? Clin Epidemiol 2019;11:509-518 [FREE Full text] [doi: 10.2147/CLEP.S206770] [Medline: 31456649]

123. Ponjoan A, Garre-Olmo J, Blanch J, Fages E, Alves-Cabratosa L, Martí-Lluch R, et al. Epidemiology of dementia: prevalence and incidence estimates using validated electronic health records from primary care. Clin Epidemiol 2019;11:217-228 [FREE Full text] [doi: 10.2147/CLEP.S186590] [Medline: 30881138]

124. Pujades-Rodriguez M, Assi V, Gonzalez-Izquierdo A, Wilkinson T, Schnier C, Sudlow C, et al. The diagnosis, burden and prognosis of dementia: A record-linkage cohort study in England. PLoS One 2018;13(6):e0199026 [FREE Full text] [doi: 10.1371/journal.pone.0199026] [Medline: 29944675]

125. Reuben DB, Hackbarth AS, Wenger NS, Tan ZS, Jennings LA. An Automated Approach to Identifying Patients with Dementia Using Electronic Medical Records. J Am Geriatr Soc 2017 Mar;65(3):658-659. [doi: 10.1111/jgs.14744] [Medline: 28152164]

126. Sommerlad A, Perera G, Mueller C, Singh-Manoux A, Lewis G, Stewart R, et al. Hospitalisation of people with dementia: evidence from English electronic health records from 2008 to 2016. Eur J Epidemiol 2019 Jul;34(6):567-577 [FREE Full text] [doi: 10.1007/s10654-019-00481-x] [Medline: 30649705]

127. van Bussel EF, Richard E, Arts DL, Nooyens ACJ, Coloma PM, de Waal MWM, et al. Dementia incidence trend over 1992-2014 in the Netherlands: Analysis of primary care data. PLoS Med 2017 Mar;14(3):e1002235 [FREE Full text] [doi: 10.1371/journal.pmed.1002235] [Medline: 28267788]

128. Wei W, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc 2016 Apr;23(e1):e20-e27 [FREE Full text] [doi: 10.1093/jamia/ocv130] [Medline: 26338219]

129. Williamson T, Green ME, Birtwhistle R, Khan S, Garies S, Wong ST, et al. Validating the 8 CPCSSN case definitions for chronic disease surveillance in a primary care database of electronic health records. Ann Fam Med 2014 Jul;12(4):367-372 [FREE Full text] [doi: 10.1370/afm.1644] [Medline: 25024246]

130. Wu C, Kuo C, Su C, Wang S, Dai H. Using text mining to extract depressive symptoms and to validate the diagnosis of major depressive disorder from electronic health records. J Affect Disord 2020 Jan 01;260:617-623. [doi: 10.1016/j.jad.2019.09.044] [Medline: 31541973]

131. Afzal Z, Engelkes M, Verhamme KMC, Janssens HM, Sturkenboom MCJM, Kors JA, et al. Automatic generation of case-detection algorithms to identify children with asthma from large electronic health record databases. Pharmacoepidemiol Drug Saf 2013 Aug;22(8):826-833. [doi: 10.1002/pds.3438] [Medline: 23592573]

132. Akgün KM, Sigel K, Cheung K, Kidwai-Khan F, Bryant AK, Brandt C, et al. Extracting lung function measurements to enhance phenotyping of chronic obstructive pulmonary disease (COPD) in an electronic health record using automated tools. PLoS One 2020;15(1):e0227730 [FREE Full text] [doi: 10.1371/journal.pone.0227730] [Medline: 31945115]

133. Almoguera B, Vazquez L, Mentch F, Connolly J, Pacheco JA, Sundaresan AS, et al. Identification of Four Novel Loci in Asthma in European American and African American Populations. Am J Respir Crit Care Med 2017 Mar 15;195(4):456-463. [doi: 10.1164/rccm.201604-0861OC] [Medline: 27611488]

134. Asche C, Said Q, Joish V, Hall CO, Brixner D. Assessment of COPD-related outcomes via a national electronic medical record database. Int J Chron Obstruct Pulmon Dis 2008;3(2):323-326 [FREE Full text] [doi: 10.2147/copd.s1857] [Medline: 18686742]

135. Borlée F, Yzermans CJ, Krop E, Aalders B, Rooijackers J, Zock J, et al. Spirometry, questionnaire and electronic medical record based COPD in a population survey: Comparing prevalence, level of agreement and associations with potential risk factors. PLoS One 2017;12(3):e0171494 [FREE Full text] [doi: 10.1371/journal.pone.0171494] [Medline: 28273094]

136. Cave AJ, Davey C, Ahmadi E, Drummond N, Fuentes S, Kazemi-Bajestani SMR, et al. Development of a validated algorithm for the diagnosis of paediatric asthma in electronic medical records. NPJ Prim Care Respir Med 2016 Dec 24;26:16085 [FREE Full text] [doi: 10.1038/npjpcrm.2016.85] [Medline: 27882997]

137. Crothers K, Rodriguez CV, Nance RM, Akgun K, Shahrir S, Kim J, et al. Accuracy of electronic health record data for the diagnosis of chronic obstructive pulmonary disease in persons living with HIV and uninfected persons. Pharmacoepidemiol Drug Saf 2019 Feb;28(2):140-147. [doi: 10.1002/pds.4567] [Medline: 29923258]

138. DiSantostefano RL, Sampson T, Le HV, Hinds D, Davis KJ, Bakerly ND. Risk of pneumonia with inhaled corticosteroid versus long-acting bronchodilator regimens in chronic obstructive pulmonary disease: a new-user cohort study. PLoS One 2014;9(5):e97149 [FREE Full text] [doi: 10.1371/journal.pone.0097149] [Medline: 24878543]

139. Hsu J, Pacheco JA, Stevens WW, Smith ME, Avila PC. Accuracy of phenotyping chronic rhinosinusitis in the electronic health record. Am J Rhinol Allergy 2014;28(2):140-144 [FREE Full text] [doi: 10.2500/ajra.2014.28.4012] [Medline: 24717952]

140. Kadhim-Saleh A, Green M, Williamson T, Hunter D, Birtwhistle R. Validation of the diagnostic algorithms for 5 chronic conditions in the Canadian Primary Care Sentinel Surveillance Network (CPCSSN): a Kingston Practice-based Research

Network (PBRN) report. J Am Board Fam Med 2013;26(2):159-167 [FREE Full text] [doi: 10.3122/jabfm.2013.02.120183] [Medline: 23471929]

141. Kurmi OP, Vaucher J, Xiao D, Holmes MV, Guo Y, Davis KJ, et al. Validity of COPD diagnoses reported through nationwide health insurance systems in the People's Republic of China. Int J Chron Obstruct Pulmon Dis 2016;11:419-430 [FREE Full text] [doi: 10.2147/COPD.S100736] [Medline: 27042034]

142. Lee TM, Tu K, Wing LL, Gershon AS. Identifying individuals with physician-diagnosed chronic obstructive pulmonary disease in primary care electronic medical records: a retrospective chart abstraction study. NPJ Prim Care Respir Med 2017 May 15;27(1):34 [FREE Full text] [doi: 10.1038/s41533-017-0035-9] [Medline: 28507288]

143. Liang H, Tsui BY, Ni H, Valentim CCS, Baxter SL, Liu G, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med 2019 Mar;25(3):433-438. [doi: 10.1038/s41591-018-0335-9] [Medline: 30742121]

144. Nissen F, Morales DR, Mullerova H, Smeeth L, Douglas IJ, Quint JK. Validation of asthma recording in the Clinical Practice Research Datalink (CPRD). BMJ Open 2017 Aug 11;7(8):e017474 [FREE Full text] [doi: 10.1136/bmjopen-2017-017474] [Medline: 28801439]

145. Nissen F, Morales DR, Mullerova H, Smeeth L, Douglas IJ, Quint JK. Concomitant diagnosis of asthma and COPD: a quantitative study in UK primary care. Br J Gen Pract 2018 Dec;68(676):e775-e782. [doi: 10.3399/bjgp18X699389] [Medline: 30249612]

146. Pacheco JA, Avila PC, Thompson JA, Law M, Quraishi JA, Greiman AK, et al. A highly specific algorithm for identifying asthma cases and controls for genome-wide association studies. AMIA Annu Symp Proc 2009 Dec 14;2009:497-501. [Medline: 20351906]

147. Pennington AF, Strickland MJ, Freedle KA, Klein M, Drews-Botsch C, Hansen C, et al. Evaluating early-life asthma definitions as a marker for subsequent asthma in an electronic medical record setting. Pediatr Allergy Immunol 2016 Sep;27(6):591-596. [doi: 10.1111/pai.12586] [Medline: 27116587]

148. Rothnie KJ, Chandan JS, Goss HG, Müllerová H, Quint JK. Validity and interpretation of spirometric recordings to diagnose COPD in UK primary care. Int J Chron Obstruct Pulmon Dis 2017;12:1663-1668 [FREE Full text] [doi: 10.2147/COPD.S133891] [Medline: 28652719]

149. Rothnie KJ, Müllerová H, Hurst JR, Smeeth L, Davis K, Thomas SL, et al. Validation of the Recording of Acute Exacerbations of COPD in UK Primary Care Electronic Healthcare Records. PLoS One 2016;11(3):e0151357 [FREE Full text] [doi: 10.1371/journal.pone.0151357] [Medline: 26959820]

150. Schulz S, Seddig T, Hanser S, Zaiss A, Daumke P. Checking coding completeness by mining discharge summaries. Stud Health Technol Inform 2011;169:594-598. [Medline: 21893818]

151. Seol HY, Rolfes MC, Chung W, Sohn S, Ryu E, Park MA, et al. Expert artificial intelligence-based natural language processing characterises childhood asthma. BMJ Open Resp Res 2020 Feb 04;7(1):e000524. [doi: 10.1136/bmjresp-2019-000524]

152. Sohn S, Wang Y, Wi C, Krusemark EA, Ryu E, Ali MH, et al. Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions. J Am Med Inform Assoc 2017 Nov 30. [doi: 10.1093/jamia/ocx138] [Medline: 29202185]

153. Sohn S, Wi C, Wu ST, Liu H, Ryu E, Krusemark E, et al. Ascertainment of asthma prognosis using natural language processing from electronic medical records. J Allergy Clin Immunol 2018 Jun;141(6):2292-2294.e3 [FREE Full text] [doi: 10.1016/j.jaci.2017.12.1003] [Medline: 29438770]

154. Sperrin M, Webb DJ, Patel P, Davis KJ, Collier S, Pate A, et al. Chronic obstructive pulmonary disease exacerbation episodes derived from electronic health record data validated using clinical trial data. Pharmacoepidemiol Drug Saf 2019 Oct;28(10):1369-1376 [FREE Full text] [doi: 10.1002/pds.4883] [Medline: 31385428]

155. Sundaresan AS, Schneider G, Reynolds J, Kirchner HL. Identifying Asthma Exacerbation-Related Emergency Department Visit Using Electronic Medical Record and Claims Data. Appl Clin Inform 2018 Jul;9(3):528-540 [FREE Full text] [doi: 10.1055/s-0038-1666994] [Medline: 30040112]

156. Vazquez Guillamet R, Ursu O, Iwamoto G, Moseley PL, Oprea T. Chronic obstructive pulmonary disease phenotypes using cluster analysis of electronic medical records. Health Informatics J 2018 Dec;24(4):394-409 [FREE Full text] [doi: 10.1177/1460458216675661] [Medline: 27856785]

157. Wi C, Sohn S, Ali M, Krusemark E, Ryu E, Liu H, et al. Natural Language Processing for Asthma Ascertainment in Different Practice Settings. J Allergy Clin Immunol Pract 2018;6(1):126-131 [FREE Full text] [doi: 10.1016/j.jaip.2017.04.041] [Medline: 28634104]

158. Wi C, Sohn S, Rolfes MC, Seabright A, Ryu E, Voge G, et al. Application of a Natural Language Processing Algorithm to Asthma Ascertainment. An Automated Chart Review. Am J Respir Crit Care Med 2017 Aug 15;196(4):430-437 [FREE Full text] [doi: 10.1164/rccm.201610-2006OC] [Medline: 28375665]

159. Wu ST, Sohn S, Ravikumar KE, Wagholikar K, Jonnalagadda SR, Liu H, et al. Automated chart review for asthma cohort identification using natural language processing: an exploratory study. Ann Allergy Asthma Immunol 2013 Dec;111(5):364-369 [FREE Full text] [doi: 10.1016/j.anai.2013.07.022] [Medline: 24125142]

160. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006;6:30 [FREE Full text] [doi: 10.1186/1472-6947-6-30] [Medline: 16872495]

161. Barnado A, Casey C, Carroll RJ, Wheless L, Denny JC, Crofford LJ. Developing Electronic Health Record Algorithms That Accurately Identify Patients With Systemic Lupus Erythematosus. Arthritis Care Res (Hoboken) 2017 May;69(5):687-693 [FREE Full text] [doi: 10.1002/acr.22989] [Medline: 27390187]

162. Carroll RJ, Eyler AE, Denny JC. Naïve Electronic Health Record phenotype identification for Rheumatoid arthritis. AMIA Annu Symp Proc 2011;2011:189-196. [Medline: 22195070]

163. Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. J Am Med Inform Assoc 2012 Jun;19(e1):e162-e169 [FREE Full text] [doi: 10.1136/amiajnl-2011-000583] [Medline: 22374935]

164. Cote J, Berger A, Kirchner LH, Bili A. Low vitamin D level is not associated with increased incidence of rheumatoid arthritis. Rheumatol Int 2014 Oct;34(10):1475-1479. [doi: 10.1007/s00296-014-3019-x] [Medline: 24748471]

165. de Abreu MM, Maiorano AC, Tedeschi SK, Yoshida K, Lin T, Solomon DH. Outcomes of lupus and rheumatoid arthritis patients with primary dengue infection: A seven-year report from Brazil. Semin Arthritis Rheum 2018 Apr;47(5):749-755. [doi: 10.1016/j.semarthrit.2017.09.001] [Medline: 29061443]

166. Escudié J, Rance B, Malamut G, Khater S, Burgun A, Cellier C, et al. A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities in patients with celiac disease. BMC Med Inform Decis Mak 2017 Oct 29;17(1):140 [FREE Full text] [doi: 10.1186/s12911-017-0537-y] [Medline: 28962565]

167. Ford E, Carroll J, Smith H, Davies K, Koeling R, Petersen I, et al. What evidence is there for a delay in diagnostic coding of RA in UK general practice records? An observational study of free text. BMJ Open 2016 Jun 28;6(6):e010393 [FREE Full text] [doi: 10.1136/bmjopen-2015-010393] [Medline: 27354069]

168. Ford E, Nicholson A, Koeling R, Tate A, Carroll J, Axelrod L, et al. Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: what information is hidden in free text? BMC Med Res Methodol 2013 Aug 21;13:105 [FREE Full text] [doi: 10.1186/1471-2288-13-105] [Medline: 23964710]

169. Jamian L, Wheless L, Crofford LJ, Barnado A. Rule-based and machine learning algorithms identify patients with systemic sclerosis accurately in the electronic health record. Arthritis Res Ther 2019 Dec 30;21(1):305. [doi: 10.1186/s13075-019-2092-7] [Medline: 31888720]

170. Jorge A, Castro VM, Barnado A, Gainer V, Hong C, Cai T, et al. Identifying lupus patients in electronic health records: Development and validation of machine learning algorithms and application of rule-based algorithms. Semin Arthritis Rheum 2019 Aug;49(1):84-90 [FREE Full text] [doi: 10.1016/j.semarthrit.2019.01.002] [Medline: 30665626]

171. Kronzer VL, Wang L, Liu H, Davis JM, Sparks JA, Crowson CS. Investigating the impact of disease and health record duration on the eMERGE algorithm for rheumatoid arthritis. J Am Med Inform Assoc 2020 May 01;27(4):601-605. [doi: 10.1093/jamia/ocaa014] [Medline: 32134444]

172. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res (Hoboken) 2010 Aug;62(8):1120-1127 [FREE Full text] [doi: 10.1002/acr.20184] [Medline: 20235204]

173. Lin C, Karlson EW, Canhao H, Miller TA, Dligach D, Chen PJ, et al. Automatic prediction of rheumatoid arthritis disease activity from the electronic medical records. PLoS One 2013;8(8):e69932 [FREE Full text] [doi: 10.1371/journal.pone.0069932] [Medline: 23976944]

174. Muller S, Hider SL, Raza K, Stack RJ, Hayward RA, Mallen CD. An algorithm to identify rheumatoid arthritis in primary care: a Clinical Practice Research Datalink study. BMJ Open 2015 Dec 23;5(12):e009309 [FREE Full text] [doi: 10.1136/bmjopen-2015-009309] [Medline: 26700281]

175. Murray SG, Avati A, Schmajuk G, Yazdany J. Automated and flexible identification of complex disease: building a model for systemic lupus erythematosus using noisy labeling. J Am Med Inform Assoc 2019 Jan 01;26(1):61-65. [doi: 10.1093/jamia/ocy154] [Medline: 30476175]

176. Nicholson A, Ford E, Davies KA, Smith HE, Rait G, Tate AR, et al. Optimising use of electronic health records to describe the presentation of rheumatoid arthritis in primary care: a strategy for developing code lists. PLoS One 2013;8(2):e54878 [FREE Full text] [doi: 10.1371/journal.pone.0054878] [Medline: 23451024]

177. Nielen MMJ, Ursum J, Schellevis FG, Korevaar JC. The validity of the diagnosis of inflammatory arthritis in a large population-based primary care database. BMC Fam Pract 2013 Jul 07;14:79 [FREE Full text] [doi: 10.1186/1471-2296-14-79] [Medline: 24128086]

178. Nikiphorou E, de Lusignan S, Mallen CD, Khavandi K, Bedarida G, Buckley CD, et al. Cardiovascular risk factors and outcomes in early rheumatoid arthritis: a population-based study. Heart 2020 Mar 24 [FREE Full text] [doi: 10.1136/heartjnl-2019-316193] [Medline: 32209618]

179. Partington RJ, Muller S, Helliwell T, Mallen CD, Abdul Sultan A. Incidence, prevalence and treatment burden of polymyalgia rheumatica in the UK over two decades: a population-based study. Ann Rheum Dis 2018 Dec;77(12):1750-1756. [doi: 10.1136/annrheumdis-2018-213883] [Medline: 30297332]

XSL•FO

RenderX

180. Redd D, Frech TM, Murtaugh MA, Rhiannon J, Zeng QT. Informatics can identify systemic sclerosis (SSc) patients at risk for scleroderma renal crisis. Comput Biol Med 2014 Oct;53:203-205 [FREE Full text] [doi: 10.1016/j.compbiomed.2014.07.022] [Medline: 25168254]

181. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet 2010 Apr 09;86(4):560-572 [FREE Full text] [doi: 10.1016/j.ajhg.2010.03.003] [Medline: 20362271]

182. Turner CA, Jacobs AD, Marques CK, Oates JC, Kamen DL, Anderson PE, et al. Word2Vec inversion and traditional text classifiers for phenotyping lupus. BMC Med Inform Decis Mak 2017 Aug 22;17(1):126 [FREE Full text] [doi: 10.1186/s12911-017-0518-1] [Medline: 28830409]

183. Verma A, Basile AO, Bradford Y, Kuivaniemi H, Tromp G, Carey D, et al. Phenome-Wide Association Study to Explore Relationships between Immune System Related Genetic Loci and Complex Traits and Diseases. PLoS One 2016;11(8):e0160573 [FREE Full text] [doi: 10.1371/journal.pone.0160573] [Medline: 27508393]

184. Wang L, Rastegar-Mojarad M, Ji Z, Liu S, Liu K, Moon S, et al. Detecting Pharmacovigilance Signals Combining Electronic Medical Records With Spontaneous Reports: A Case Study of Conventional Disease-Modifying Antirheumatic Drugs for Rheumatoid Arthritis. Front Pharmacol 2018;9:875. [doi: 10.3389/fphar.2018.00875] [Medline: 30131701]

185. Zhou S, Fernandez-Gutierrez F, Kennedy J, Cooksey R, Atkinson M, Denaxas S, UK Biobank Follow-upOutcomes Group, et al. Defining Disease Phenotypes in Primary Care Electronic Health Records by a Machine Learning Approach: A Case Study in Identifying Rheumatoid Arthritis. PLoS One 2016;11(5):e0154515 [FREE Full text] [doi: 10.1371/journal.pone.0154515] [Medline: 27135409]

186. Gill JM, Mainous AG, Koopman RJ, Player MS, Everett CJ, Chen YX, et al. Impact of EHR-based clinical decision support on adherence to guidelines for patients on NSAIDs: a randomized controlled trial. Ann Fam Med 2011;9(1):22-30 [FREE Full text] [doi: 10.1370/afm.1172] [Medline: 21242557]

187. Salmasian H, Freedberg DE, Abrams JA, Friedman C. An automated tool for detecting medication overuse based on the electronic health records. Pharmacoepidemiol Drug Saf 2013 Feb;22(2):183-189 [FREE Full text] [doi: 10.1002/pds.3387] [Medline: 23233423]

188. Shelat VG, Ahmed S, Chia CLK, Cheah YL. Strict Selection Criteria During Surgical Training Ensures Good Outcomes in Laparoscopic Omental Patch Repair (LOPR) for Perforated Peptic Ulcer (PPU). Int Surg 2015 Mar;100(2):370-375. [doi: 10.9738/INTSURG-D-13-00241.1] [Medline: 25692444]

189. Singh B, Singh A, Ahmed A, Wilson GA, Pickering BW, Herasevich V, et al. Derivation and validation of automated electronic search strategies to extract Charlson comorbidities from electronic medical records. Mayo Clin Proc 2012 Sep;87(9):817-824 [FREE Full text] [doi: 10.1016/j.mayocp.2012.04.015] [Medline: 22958988]

190. Ahmad FS, Chan C, Rosenman MB, Post WS, Fort DG, Greenland P, et al. Validity of Cardiovascular Data From Electronic Sources: The Multi-Ethnic Study of Atherosclerosis and HealthLNK. Circulation 2017 Oct 26;136(13):1207-1216. [doi: 10.1161/CIRCULATIONAHA.117.027436] [Medline: 28687707]

191. Chi GC, Li X, Tartof SY, Slezak JM, Koebnick C, Lawrence JM. Validity of ICD-10-CM codes for determination of diabetes type for persons with youth-onset type 1 and type 2 diabetes. BMJ Open Diabetes Res Care 2019;7(1):e000547 [FREE Full text] [doi: 10.1136/bmjdrc-2018-000547] [Medline: 30899525]

192. Coleman N, Halas G, Peeler W, Casalang N, Williamson T, Katz A. From patient care to research: a validation study examining the factors contributing to data quality in a primary care electronic medical record database. BMC Fam Pract 2015 Feb 05;16:11 [FREE Full text] [doi: 10.1186/s12875-015-0223-z] [Medline: 25649201]

193. Crawford AG, Cote C, Couto J, Daskiran M, Gunnarsson C, Haas K, et al. Prevalence of obesity, type II diabetes mellitus, hyperlipidemia, and hypertension in the United States: findings from the GE Centricity Electronic Medical Record database. Popul Health Manag 2010 Jul;13(3):151-161. [doi: 10.1089/pop.2009.0039] [Medline: 20521902]

194. Esteban S, Rodríguez Tablado M, Peper F, Mahumud YS, Ricci RI, Kopitowski K, et al. Development and Validation of Various Phenotyping Algorithms for Diabetes Mellitus Using Data from Electronic Health Records. Stud Health Technol Inform 2017;245:366-369. [Medline: 29295117]

195. Gjelsvik B, Tran AT, Berg TJ, Bakke , Mdala I, Nøkleby K, et al. Exploring the relationship between coronary heart disease and type 2 diabetes: a cross-sectional study of secondary prevention among diabetes patients. BJGP Open 2019 May;3(1):bjgpopen18X101636. [doi: 10.3399/bjgpopen18X101636] [Medline: 31049416]

196. Harris SB, Glazier RH, Tompkins JW, Wilton AS, Chevendra V, Stewart MA, et al. Investigating concordance in diabetes diagnosis between primary care charts (electronic medical records) and health administrative data: a retrospective cohort study. BMC Health Serv Res 2010 Dec 23;10:347. [doi: 10.1186/1472-6963-10-347] [Medline: 21182790]

197. Henderson J, Barnett S, Ghosh A, Pollack AJ, Hodgkins A, Win KT, et al. Validation of electronic medical data: Identifying diabetes prevalence in general practice. Health Inf Manag 2019 Jan;48(1):3-11. [doi: 10.1177/1833358318798123] [Medline: 30278786]

198. Ho ML, Lawrence N, van Walraven C, Manuel D, Keely E, Malcolm J, et al. The accuracy of using integrated electronic health care data to identify patients with undiagnosed diabetes mellitus. J Eval Clin Pract 2012 Jul;18(3):606-611. [doi: 10.1111/j.1365-2753.2011.01633.x] [Medline: 21332609]

199. Kadhim-Saleh A, Green M, Williamson T, Hunter D, Birtwhistle R. Validation of the diagnostic algorithms for 5 chronic conditions in the Canadian Primary Care Sentinel Surveillance Network (CPCSSN): a Kingston Practice-based Research Network (PBRN) report. J Am Board Fam Med 2013;26(2):159-167 [FREE Full text] [doi: 10.3122/jabfm.2013.02.120183] [Medline: 23471929]

200. Ke C, Stukel TA, Luk A, Shah BR, Jha P, Lau E, et al. Development and validation of algorithms to classify type 1 and 2 diabetes according to age at diagnosis using electronic health records. BMC Med Res Methodol 2020 Feb 24;20(1):35 [FREE Full text] [doi: 10.1186/s12874-020-00921-3] [Medline: 32093635]

201. 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]

202. Khokhar B, Quan H, Kaplan GG, Butalia S, Rabi D. Exploring novel diabetes surveillance methods: a comparison of administrative, laboratory and pharmacy data case definitions using THIN. J Public Health (Oxf) 2018 Sep 01;40(3):652-658. [doi: 10.1093/pubmed/fdx096] [Medline: 28977382]

203. Klompas M, Eggleston E, McVetta J, Lazarus R, Li L, Platt R. Automated detection and classification of type 1 versus type 2 diabetes using electronic health record data. Diabetes Care 2013 Apr;36(4):914-921 [FREE Full text] [doi: 10.2337/dc12-0964] [Medline: 23193215]

204. Kosowan L, Wicklow B, Queenan J, Yeung R, Amed S, Singer A. Enhancing Health Surveillance: Validation of a Novel Electronic Medical Records-Based Definition of Cases of Pediatric Type 1 and Type 2 Diabetes Mellitus. Can J Diabetes 2019 Aug;43(6):392-398. [doi: 10.1016/j.jcjd.2019.02.005] [Medline: 30956098]

205. Kudyakov R, Bowen J, Ewen E, West SL, Daoud Y, Fleming N, et al. Electronic health record use to classify patients with newly diagnosed versus preexisting type 2 diabetes: infrastructure for comparative effectiveness research and population health management. Popul Health Manag 2012 Mar;15(1):3-11. [doi: 10.1089/pop.2010.0084] [Medline: 21877923]

206. Lawrence JM, Black MH, Zhang JL, Slezak JM, Takhar HS, Koebnick C, et al. Validation of pediatric diabetes case identification approaches for diagnosed cases by using information in the electronic health records of a large integrated managed health care organization. Am J Epidemiol 2014 Jan 01;179(1):27-38. [doi: 10.1093/aje/kwt230] [Medline: 24100956]

207. Lipscombe LL, Hwee J, Webster L, Shah BR, Booth GL, Tu K. Identifying diabetes cases from administrative data: a population-based validation study. BMC Health Serv Res 2018 May 02;18(1):316. [doi: 10.1186/s12913-018-3148-0] [Medline: 29720153]

208. Makam AN, Nguyen OK, Moore B, Ma Y, Amarasingham R. Identifying patients with diabetes and the earliest date of diagnosis in real time: an electronic health record case-finding algorithm. BMC Med Inform Decis Mak 2013 Aug 01;13:81 [FREE Full text] [doi: 10.1186/1472-6947-13-81] [Medline: 23915139]

209. Moreno-Iribas C, Sayon-Orea C, Delfrade J, Ardanaz E, Gorricho J, Burgui R, et al. Validity of type 2 diabetes diagnosis in a population-based electronic health record database. BMC Med Inform Decis Mak 2017 Apr 08;17(1):34. [doi: 10.1186/s12911-017-0439-z] [Medline: 28390396]

210. Nichols GA, Desai J, Elston LJ, Lawrence JM, O'Connor PJ, Pathak RD, et al. Construction of a multisite DataLink using electronic health records for the identification, surveillance, prevention, and management of diabetes mellitus: the SUPREME-DM project. Prev Chronic Dis 2012;9:E110 [FREE Full text] [Medline: 22677160]

211. Nichols GA, Schroeder EB, Karter AJ, Gregg EW, Desai J, Lawrence JM, et al. Trends in diabetes incidence among 7 million insured adults, 2006-2011: the SUPREME-DM project. Am J Epidemiol 2015 Jan 1;181(1):32-39. [doi: 10.1093/aje/kwu255] [Medline: 25515167]

212. Young JB, Gauthier-Loiselle M, Bailey RA, Manceur AM, Lefebvre P, Greenberg M, et al. Development of predictive risk models for major adverse cardiovascular events among patients with type 2 diabetes mellitus using health insurance claims data. Cardiovasc Diabetol 2018 Aug 24;17(1):118 [FREE Full text] [doi: 10.1186/s12933-018-0759-z] [Medline: 30143045]

213. Pacheco JA, Thompson W, Kho A. Automatically detecting problem list omissions of type 2 diabetes cases using electronic medical records. AMIA Annu Symp Proc 2011;2011:1062-1069 [FREE Full text] [Medline: 22195167]

214. Pantalone KM, Misra-Hebert AD, Hobbs TM, Wells BJ, Kong SX, Chagin K, et al. Effect of glycemic control on the Diabetes Complications Severity Index score and development of complications in people with newly diagnosed type 2 diabetes. J Diabetes 2018 Mar;10(3):192-199. [doi: 10.1111/1753-0407.12613] [Medline: 28976724]

215. Paul SK, Shaw JE, Montvida O, Klein K. Weight gain in insulin-treated patients by body mass index category at treatment initiation: new evidence from real-world data in patients with type 2 diabetes. Diabetes Obes Metab 2016 Dec;18(12):1244-1252. [doi: 10.1111/dom.12761] [Medline: 27502528]

216. Richesson RL, Rusincovitch SA, Wixted D, Batch BC, Feinglos MN, Miranda ML, et al. A comparison of phenotype definitions for diabetes mellitus. J Am Med Inform Assoc 2013 Dec;20(e2):e319-e326 [FREE Full text] [doi: 10.1136/amiajnl-2013-001952] [Medline: 24026307]

217. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet 2010 Apr 09;86(4):560-572 [FREE Full text] [doi: 10.1016/j.ajhg.2010.03.003] [Medline: 20362271]

218. Rodgers LR, Weedon MN, Henley WE, Hattersley AT, Shields BM. Cohort profile for the MASTERMIND study: using the Clinical Practice Research Datalink (CPRD) to investigate stratification of response to treatment in patients with type 2 diabetes. BMJ Open 2017 Oct 12;7(10):e017989 [FREE Full text] [doi: 10.1136/bmjopen-2017-017989] [Medline: 29025846]

219. Schroeder EB, Donahoo WT, Goodrich GK, Raebel MA. Validation of an algorithm for identifying type 1 diabetes in adults based on electronic health record data. Pharmacoepidemiol Drug Saf 2018 Oct;27(10):1053-1059 [FREE Full text] [doi: 10.1002/pds.4377] [Medline: 29292555]

220. Sharma M, Petersen I, Nazareth I, Coton SJ. An algorithm for identification and classification of individuals with type 1 and type 2 diabetes mellitus in a large primary care database. Clin Epidemiol 2016;8:373-380 [FREE Full text] [doi: 10.2147/CLEP.S113415] [Medline: 27785102]

221. Spratt SE, Pereira K, Granger BB, Batch BC, Phelan M, Pencina M, et al. Assessing electronic health record phenotypes against gold-standard diagnostic criteria for diabetes mellitus. J Am Med Inform Assoc 2017 May 01;24(e1):e121-e128 [FREE Full text] [doi: 10.1093/jamia/ocw123] [Medline: 27616701]

222. Teltsch DY, Fazeli Farsani S, Swain RS, Kaspers S, Huse S, Cristaldi C, et al. Development and validation of algorithms to identify newly diagnosed type 1 and type 2 diabetes in pediatric population using electronic medical records and claims data. Pharmacoepidemiol Drug Saf 2019 Feb;28(2):234-243. [doi: 10.1002/pds.4728] [Medline: 30677205]

223. Tu K, Manuel D, Lam K, Kavanagh D, Mitiku TF, Guo H. Diabetics can be identified in an electronic medical record using laboratory tests and prescriptions. J Clin Epidemiol 2011 May;64(4):431-435. [doi: 10.1016/j.jclinepi.2010.04.007] [Medline: 20638237]

224. Upadhyaya SG, Murphree DH, Ngufor CG, Knight AM, Cronk DJ, Cima RR, et al. Automated Diabetes Case Identification Using Electronic Health Record Data at a Tertiary Care Facility. Mayo Clin Proc Innov Qual Outcomes 2017 Jul;1(1):100-110 [FREE Full text] [doi: 10.1016/j.mayocpiqo.2017.04.005] [Medline: 30225406]

225. Wei W, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, Chai HS, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. J Am Med Inform Assoc 2012;19(2):219-224 [FREE Full text] [doi: 10.1136/amiajnl-2011-000597] [Medline: 22249968]

226. Wiese AD, Roumie CL, Buse JB, Guzman H, Bradford R, Zalimeni E, et al. Performance of a computable phenotype for identification of patients with diabetes within PCORnet: The Patient-Centered Clinical Research Network. Pharmacoepidemiol Drug Saf 2019 May;28(5):632-639 [FREE Full text] [doi: 10.1002/pds.4718] [Medline: 30680840]

227. Williams BA, Geba D, Cordova JM, Shetty SS. A risk prediction model for heart failure hospitalization in type 2 diabetes mellitus. Clin Cardiol 2020 Mar;43(3):275-283 [FREE Full text] [doi: 10.1002/clc.23298] [Medline: 31837035]

228. Wysham CH, Lefebvre P, Pilon D, Lafeuille M, Emond B, Kamstra R, et al. An investigation into the durability of glycemic control in patients with type II diabetes initiated on canagliflozin or sitagliptin: A real-world analysis of electronic medical records. J Diabetes Complications 2019 Feb;33(2):140-147 [FREE Full text] [doi: 10.1016/j.jdiacomp.2018.10.016] [Medline: 30554838]

229. Yang F, Ma Q, Liu J, Ma B, Guo M, Liu F, et al. Prevalence and major risk factors of type 2 diabetes mellitus among adult psychiatric inpatients from 2005 to 2018 in Beijing, China: a longitudinal observational study. BMJ Open Diabetes Res Care 2020 Mar;8(1) [FREE Full text] [doi: 10.1136/bmjdrc-2019-000996] [Medline: 32139600]

230. Yue X, Wu J, Ruan Z, Wolden ML, Li L, Lin Y. The Burden of Hypoglycemia in Patients With Insulin-Treated Diabetes Mellitus in China: Analysis of Electronic Medical Records From 4 Tertiary Hospitals. Value Health Reg Issues 2020 May;21:17-21. [doi: 10.1016/j.vhri.2019.06.003] [Medline: 31634792]

231. Zheng L, Wang Y, Hao S, Shin AY, Jin B, Ngo AD, et al. Web-based Real-Time Case Finding for the Population Health Management of Patients With Diabetes Mellitus: A Prospective Validation of the Natural Language Processing-Based Algorithm With Statewide Electronic Medical Records. JMIR Med Inform 2016 Nov 11;4(4):e37 [FREE Full text] [doi: 10.2196/medinform.6328] [Medline: 27836816]

232. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform 2017 Dec;97:120-127 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.09.014] [Medline: 27919371]

233. Zhong VW, Obeid JS, Craig JB, Pfaff ER, Thomas J, Jaacks LM, et al. An efficient approach for surveillance of childhood diabetes by type derived from electronic health record data: the SEARCH for Diabetes in Youth Study. J Am Med Inform Assoc 2016 Nov;23(6):1060-1067 [FREE Full text] [doi: 10.1093/jamia/ocv207] [Medline: 27107449]

234. Zhong VW, Pfaff ER, Beavers DP, Thomas J, Jaacks LM, Bowlby DA, Search for Diabetes in Youth Study Group. Use of administrative and electronic health record data for development of automated algorithms for childhood diabetes case ascertainment and type classification: the SEARCH for Diabetes in Youth Study. Pediatr Diabetes 2014 Dec;15(8):573-584 [FREE Full text] [doi: 10.1111/pedi.12152] [Medline: 24913103]

235. Agrawal S, Kremsdorf R, Uysal S, Fredette ME, Topor LS. Nephrolithiasis: A complication of pediatric diabetic ketoacidosis. Pediatr Diabetes 2018 Mar;19(2):329-332. [doi: 10.1111/pedi.12559] [Medline: 28737266]

236. Cahn A, Altaras T, Agami T, Liran O, Touaty CE, Drahy M, et al. Validity of diagnostic codes and estimation of prevalence of diabetic foot ulcers using a large electronic medical record database. Diabetes Metab Res Rev 2019 Feb;35(2):e3094. [doi: 10.1002/dmrr.3094] [Medline: 30378240]

237. Dong Y, Gao W, Zhang L, Wei J, Hammar N, Cabrera CS, et al. Patient characteristics related to metabolic disorders and chronic complications in type 2 diabetes mellitus patients hospitalized at the Qingdao Endocrine and Diabetes Hospital from 2006 to 2012 in China. Diab Vasc Dis Res 2017 Jan;14(1):24-32 [FREE Full text] [doi: 10.1177/1479164116675489] [Medline: 27941053]

238. DuBrava S, Mardekian J, Sadosky A, Bienen EJ, Parsons B, Hopps M, et al. Using Random Forest Models to Identify Correlates of a Diabetic Peripheral Neuropathy Diagnosis from Electronic Health Record Data. Pain Med 2017 Dec 01;18(1):107-115. [doi: 10.1093/pm/pnw096] [Medline: 27252307]

239. Lee CS, Lee AY, Baughman D, Sim D, Akelere T, Brand C, UK DR EMR Users Group. The United Kingdom Diabetic Retinopathy Electronic Medical Record Users Group: Report 3: Baseline Retinopathy and Clinical Features Predict Progression of Diabetic Retinopathy. Am J Ophthalmol 2017 Aug;180:64-71 [FREE Full text] [doi: 10.1016/j.ajo.2017.05.020] [Medline: 28572062]

240. Martín-Merino E, Fortuny J, Rivero-Ferrer E, García-Rodríguez LA. Incidence of retinal complications in a cohort of newly diagnosed diabetic patients. PLoS One 2014;9(6):e100283 [FREE Full text] [doi: 10.1371/journal.pone.0100283] [Medline: 24963628]

241. Song X, Waitman LR, Hu Y, Yu ASL, Robins D, Liu M. Robust clinical marker identification for diabetic kidney disease with ensemble feature selection. J Am Med Inform Assoc 2019 Mar 01;26(3):242-253. [doi: 10.1093/jamia/ocy165] [Medline: 30602020]

242. VanderWeele J, Pollack T, Oakes DJ, Smyrniotis C, Illuri V, Vellanki P, et al. Validation of data from electronic data warehouse in diabetic ketoacidosis: Caution is needed. J Diabetes Complications 2018 Jul;32(7):650-654. [doi: 10.1016/j.jdiacomp.2018.05.004] [Medline: 29903409]

243. Abhyankar S, Demner-Fushman D, Callaghan FM, McDonald CJ. Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. J Am Med Inform Assoc 2014;21(5):801-807 [FREE Full text] [doi: 10.1136/amiajnl-2013-001915] [Medline: 24384230]

244. Afzal Z, Schuemie MJ, van Blijderveen JC, Sen EF, Sturkenboom MCJM, Kors JA. Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records. BMC Med Inform Decis Mak 2013 Mar 02;13:30 [FREE Full text] [doi: 10.1186/1472-6947-13-30] [Medline: 23452306]

245. Alcober-Morte L, Barrio-Ruiz C, Parellada-Esquius N, Subirana I, Comín-Colet J, Grau M, et al. Heart failure admission across glomerular filtration rate categories in a community cohort of 125,053 individuals over 60 years of age. Hypertens Res 2019 Dec;42(12):2013-2020. [doi: 10.1038/s41440-019-0315-6] [Medline: 31477871]

246. Broder A, Mowrey WB, Izmirly P, Costenbader KH. Validation of Systemic Lupus Erythematosus Diagnosis as the Primary Cause of Renal Failure in the US Renal Data System. Arthritis Care Res (Hoboken) 2017 Apr;69(4):599-604 [FREE Full text] [doi: 10.1002/acr.22972] [Medline: 27390299]

247. Crawford DC, Bailey JNC, Miskimen K, Miron P, McCauley JL, Sedor JR, et al. Somatic T-cell Receptor Diversity in a Chronic Kidney Disease PatientPopulation Linked to Electronic Health Records. AMIA Jt Summits Transl Sci Proc 2018;2017:63-71 [FREE Full text] [Medline: 29888042]

248. Ernecoff NC, Wessell KL, Hanson LC, Lee AM, Shea CM, Dusetzina SB, et al. Electronic Health Record Phenotypes for Identifying Patients with Late-Stage Disease: a Method for Research and Clinical Application. J Gen Intern Med 2019 Dec;34(12):2818-2823 [FREE Full text] [doi: 10.1007/s11606-019-05219-9] [Medline: 31396813]

249. Fraccaro P, van der Veer S, Brown B, Prosperi M, O'Donoghue D, Collins GS, et al. An external validation of models to predict the onset of chronic kidney disease using population-based electronic health records from Salford, UK. BMC Med 2016 Jul 12;14:104 [FREE Full text] [doi: 10.1186/s12916-016-0650-2] [Medline: 27401013]

250. Hao S, Fu T, Wu Q, Jin B, Zhu C, Hu Z, et al. Estimating One-Year Risk of Incident Chronic Kidney Disease: Retrospective Development and Validation Study Using Electronic Medical Record Data From the State of Maine. JMIR Med Inform 2017 Jul 26;5(3):e21 [FREE Full text] [doi: 10.2196/medinform.7954] [Medline: 28747298]

251. Kitsos A, Peterson GM, Jose MD, Khanam MA, Castelino RL, Radford JC. Variation in Documenting Diagnosable Chronic Kidney Disease in General Medical Practice: Implications for Quality Improvement and Research. J Prim Care Community Health 2019;10:2150132719833298 [FREE Full text] [doi: 10.1177/2150132719833298] [Medline: 30879383]

252. Koyner JL, Adhikari R, Edelson DP, Churpek MM. Development of a Multicenter Ward-Based AKI Prediction Model. Clin J Am Soc Nephrol 2016 Nov 07;11(11):1935-1943 [FREE Full text] [doi: 10.2215/CJN.00280116] [Medline: 27633727]

253. Magvanjav O, Cooper-DeHoff RM, McDonough CW, Gong Y, Segal MS, Hogan WR, et al. Antihypertensive therapy prescribing patterns and correlates of blood pressure control among hypertensive patients with chronic kidney disease. J Clin Hypertens (Greenwich) 2019 Jan;21(1):91-101 [FREE Full text] [doi: 10.1111/jch.13429] [Medline: 30427124]

254. Malas MS, Wish J, Moorthi R, Grannis S, Dexter P, Duke J, et al. A comparison between physicians and computer algorithms for form CMS-2728 data reporting. Hemodial Int 2017 Jan;21(1):117-124. [doi: 10.1111/hdi.12445] [Medline: 27353890]

255. Malas MS, Wish J, Moorthi R, Grannis S, Dexter P, Duke J, et al. A comparison between physicians and computer algorithms for form CMS-2728 data reporting. Hemodial Int 2017 Jan;21(1):117-124. [doi: 10.1111/hdi.12445] [Medline: 27353890]

256.  Meyers JL, Candrilli SD, Kovacs B. Type 2 diabetes mellitus and renal impairment in a large outpatient electronic medical records database: rates of diagnosis and antihyperglycemic medication dose adjustment. Postgrad Med 2011 May;123(3):133-143. [doi: 10.3810/pgm.2011.05.2291] [Medline: 21566423]

257.  Nadkarni GN, Gottesman O, Linneman JG, Chase H, Berg RL, Farouk S, et al. Development and validation of an electronic phenotyping algorithm for chronic kidney disease. AMIA Annu Symp Proc 2014;2014:907-916 [FREE Full text] [Medline: 25954398]

258.  Robertson LM, Denadai L, Black C, Fluck N, Prescott G, Simpson W, et al. Is routine hospital episode data sufficient for identifying individuals with chronic kidney disease? A comparison study with laboratory data. Health Informatics J 2016 Jun;22(2):383-396 [FREE Full text] [doi: 10.1177/1460458214562286] [Medline: 25552482]

259.  Salvador-González B, Rodríguez-Latre LM, Güell-Miró R, Álvarez-Funes V, Sanz-Ródenas H, Tovillas-Morán FJ. Estimation of glomerular filtration rate by MDRD-4 IDMS and CKD-EPI in individuals of 60 years of age or older in primary care. Nefrologia 2013;33(4):552-563 [FREE Full text] [doi: 10.3265/Nefrologia.pre2013.Apr.11929] [Medline: 23897188]

260.  Schroeder EB, Powers JD, O'Connor PJ, Nichols GA, Xu S, Desai JR, SUPREME-DM Study Group. Prevalence of chronic kidney disease among individuals with diabetes in the SUPREME-DM Project, 2005-2011. J Diabetes Complications 2015 Jul;29(5):637-643. [doi: 10.1016/j.jdiacomp.2015.04.007] [Medline: 25936953]

261.  Semler MW, Rice TW, Shaw AD, Siew ED, Self WH, Kumar AB, et al. Identification of Major Adverse Kidney Events Within the Electronic Health Record. J Med Syst 2016 Jul;40(7):167 [FREE Full text] [doi: 10.1007/s10916-016-0528-z] [Medline: 27234478]

262.  Sun AZ, Shu Y, Harrison TN, Hever A, Jacobsen SJ, O'Shaughnessy MM, et al. Identifying Patients with Rare Disease Using Electronic Health Record Data: The Kaiser Permanente Southern California Membranous Nephropathy Cohort. Perm J 2020;24 [FREE Full text] [doi: 10.7812/TPP/19.126] [Medline: 32069207]

263.  Anand V, Hyun C, Khan QM, Hall C, Hessefort N, Sonnenberg A, et al. Identification and Fibrosis Staging of Hepatitis C Patients Using the Electronic Medical Record System. J Clin Gastroenterol 2016 Sep;50(8):664-669. [doi: 10.1097/MCG.0000000000000519] [Medline: 26974763]

264.  Atiemo K, Skaro A, Maddur H, Zhao L, Montag S, VanWagner L, et al. Mortality Risk Factors Among Patients With Cirrhosis and a Low Model for End-Stage Liver Disease Sodium Score (≤15): An Analysis of Liver Transplant Allocation Policy Using Aggregated Electronic Health Record Data. Am J Transplant 2017 Oct;17(9):2410-2419 [FREE Full text] [doi: 10.1111/ajt.14239] [Medline: 28226199]

265.  Bateman-Steel CR, Smedley EJ, Kong M, Ferson MJ. Hepatitis C enhanced surveillance: results from a southeastern Sydney pilot program. Public Health Res Pract 2015 Mar 30;25(2):e2521520 [FREE Full text] [doi: 10.17061/phrp2521520] [Medline: 25848738]

266.  Corey KE, Kartoun U, Zheng H, Shaw SY. Development and Validation of an Algorithm to Identify Nonalcoholic Fatty Liver Disease in the Electronic Medical Record. Dig Dis Sci 2016 Mar;61(3):913-919 [FREE Full text] [doi: 10.1007/s10620-015-3952-x] [Medline: 26537487]

267.  Cuthbert JA, Arslanlar S, Yepuri J, Montrose M, Ahn CW, Shah JP. Predicting short-term mortality and long-term survival for hospitalized US patients with alcoholic hepatitis. Dig Dis Sci 2014 Jul;59(7):1594-1602 [FREE Full text] [doi: 10.1007/s10620-013-3020-3] [Medline: 24445730]

268.  Fialoke S, Malarstig A, Miller MR, Dumitriu A. Application of Machine Learning Methods to Predict Non-Alcoholic Steatohepatitis (NASH) in Non-Alcoholic Fatty Liver (NAFL) Patients. AMIA Annu Symp Proc 2018;2018:430-439 [FREE Full text] [Medline: 30815083]

269.  Kaplan DE, Dai F, Aytaman A, Baytarian M, Fox R, Hunt K, VOCAL Study Group. Development and Performance of an Algorithm to Estimate the Child-Turcotte-Pugh Score From a National Electronic Healthcare Database. Clin Gastroenterol Hepatol 2015 Dec;13(13):2333-41.e1 [FREE Full text] [doi: 10.1016/j.cgh.2015.07.010] [Medline: 26188137]

270.  Kartoun U, Corey KE, Simon TG, Zheng H, Aggarwal R, Ng K, et al. The MELD-Plus: A generalizable prediction risk score in cirrhosis. PLoS One 2017;12(10):e0186301 [FREE Full text] [doi: 10.1371/journal.pone.0186301] [Medline: 29069090]

271.  Lai JC, Wong GL, Yip TC, Tse Y, Lam KL, Lui GC, et al. Chronic Hepatitis B Increases Liver-Related Mortality of Patients With Acute Hepatitis E: A Territorywide Cohort Study From 2000 to 2016. Clin Infect Dis 2018 Sep 28;67(8):1278-1284. [doi: 10.1093/cid/ciy234] [Medline: 30265321]

272.  Loomis AK, Kabadi S, Preiss D, Hyde C, Bonato V, St Louis M, et al. Body Mass Index and Risk of Nonalcoholic Fatty Liver Disease: Two Electronic Health Record Prospective Studies. J Clin Endocrinol Metab 2016 Mar;101(3):945-952 [FREE Full text] [doi: 10.1210/jc.2015-3444] [Medline: 26672639]

273.  Lu M, Chacra W, Rabin D, Rupp LB, Trudeau S, Li J, et al. Validity of an automated algorithm using diagnosis and procedure codes to identify decompensated cirrhosis using electronic health records. Clin Epidemiol 2017;9:369-376 [FREE Full text] [doi: 10.2147/CLEP.S136134] [Medline: 28744162]

274.  Nguyen TA, DeShazo JP, Thacker LR, Puri P, Sanyal AJ. The Worsening Profile of Alcoholic Hepatitis in the United States. Alcohol Clin Exp Res 2016 Jun;40(6):1295-1303 [FREE Full text] [doi: 10.1111/acer.13069] [Medline: 27147285]

275. Singal AG, Rahimi RS, Clark C, Ma Y, Cuthbert JA, Rockey DC, et al. An automated model using electronic medical record data identifies patients with cirrhosis at high risk for readmission. Clin Gastroenterol Hepatol 2013 Oct;11(10):1335-1341.e1 [FREE Full text] [doi: 10.1016/j.cgh.2013.03.022] [Medline: 23591286]

276. Xu Y, Li N, Lu M, Myers RP, Dixon E, Walker R, et al. Development and validation of method for defining conditions using Chinese electronic medical record. BMC Med Inform Decis Mak 2016 Aug 20;16:110 [FREE Full text] [doi: 10.1186/s12911-016-0348-6] [Medline: 27542973]

277. Jamil K, Huang X, Lovelace B, Pham AT, Lodaya K, Wan G. The burden of illness of hepatorenal syndrome (HRS) in the United States: a retrospective analysis of electronic health records. J Med Econ 2019 May;22(5):421-429. [doi: 10.1080/13696998.2019.1580201] [Medline: 30724682]

278. Koola JD, Davis SE, Al-Nimri O, Parr SK, Fabbri D, Malin BA, et al. Development of an automated phenotyping algorithm for hepatorenal syndrome. J Biomed Inform 2018 Apr;80:87-95 [FREE Full text] [doi: 10.1016/j.jbi.2018.03.001] [Medline: 29530803]

279. Lin C, Karlson EW, Dligach D, Ramirez MP, Miller TA, Mo H, et al. Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. J Am Med Inform Assoc 2015 May;22(e1):e151-e161 [FREE Full text] [doi: 10.1136/amiajnl-2014-002642] [Medline: 25344930]

280. Wing K, Bhaskaran K, Smeeth L, van Staa TP, Klungel OH, Reynolds RF, et al. Optimising case detection within UK electronic health records: use of multiple linked databases for detecting liver injury. BMJ Open 2016 Sep 02;6(9):e012102 [FREE Full text] [doi: 10.1136/bmjopen-2016-012102] [Medline: 27591023]

281. Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad N. Using Clinical Notes and Natural Language Processing for Automated HIV Risk Assessment. J Acquir Immune Defic Syndr 2018 Feb 01;77(2):160-166 [FREE Full text] [doi: 10.1097/QAI.0000000000001580] [Medline: 29084046]

282. Felsen UR, Bellin EY, Cunningham CO, Zingman BS. Development of an electronic medical record-based algorithm to identify patients with unknown HIV status. AIDS Care 2014;26(10):1318-1325 [FREE Full text] [doi: 10.1080/09540121.2014.911813] [Medline: 24779521]

283. Goetz MB, Hoang T, Kan VL, Rimland D, Rodriguez-Barradas M. Development and validation of an algorithm to identify patients newly diagnosed with HIV infection from electronic health records. AIDS Res Hum Retroviruses 2014 Jul;30(7):626-633. [doi: 10.1089/AID.2013.0287] [Medline: 24564256]

284. McInnes DK, Shimada SL, Midboe AM, Nazi KM, Zhao S, Wu J, et al. Patient Use of Electronic Prescription Refill and Secure Messaging and Its Association With Undetectable HIV Viral Load: A Retrospective Cohort Study. J Med Internet Res 2017 Feb 15;19(2):e34 [FREE Full text] [doi: 10.2196/jmir.6932] [Medline: 28202428]

285. Paul DW, Neely NB, Clement M, Riley I, Al-Hegelan M, Phelan M, et al. Development and validation of an electronic medical record (EMR)-based computed phenotype of HIV-1 infection. J Am Med Inform Assoc 2018 Feb 01;25(2):150-157 [FREE Full text] [doi: 10.1093/jamia/ocx061] [Medline: 28645207]

286. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]

287. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. Stud Health Technol Inform 2006;121:279-290. [Medline: 17095826]

288. 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;17(5):507-513 [FREE Full text] [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]

289. Torii M, Wagholikar K, Liu H. Using machine learning for concept extraction on clinical documents from multiple data sources. J Am Med Inform Assoc 2011;18(5):580-587 [FREE Full text] [doi: 10.1136/amiajnl-2011-000155] [Medline: 21709161]

290. Doering TA, Plapp F, Crawford JM. Establishing an evidence base for critical laboratory value thresholds. Am J Clin Pathol 2014 Dec;142(5):617-628. [doi: 10.1309/AJCPDI0FYZ4UNWEQ] [Medline: 25319976]

291. Kidney Disease: Improving Global Outcomes (KDIGO) CKD-MBD Work Group. KDIGO clinical practice guideline for the diagnosis, evaluation, prevention, and treatment of Chronic Kidney Disease-Mineral and Bone Disorder (CKD-MBD). Kidney Int Suppl 2009 Aug(113):S1-130. [doi: 10.1038/ki.2009.188] [Medline: 19644521]

292. Levey AS, Coresh J, Balk E, Kausz AT, Levin A, Steffes MW, National Kidney Foundation. National Kidney Foundation practice guidelines for chronic kidney disease: evaluation, classification, and stratification. Ann Intern Med 2003 Jul 15;139(2):137-147. [Medline: 12859163]

293. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, eMERGE Team. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics 2011 Jan 26;4:13 [FREE Full text] [doi: 10.1186/1755-8794-4-13] [Medline: 21269473]

294. McCue ME, McCoy AM. The Scope of Big Data in One Medicine: Unprecedented Opportunities and Challenges. Front Vet Sci 2017;4:194 [FREE Full text] [doi: 10.3389/fvets.2017.00194] [Medline: 29201868]

295. Sweet LE, Moulaison HL. Electronic Health Records Data and Metadata: Challenges for Big Data in the United States. Big Data 2013 Dec;1(4):245-251. [doi: 10.1089/big.2013.0023] [Medline: 27447257]

XSL•FO
RenderX

296. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10(3):e0118432 [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]

297. Reimer AP, Milinovich A, Madigan EA. Data quality assessment framework to assess electronic medical record data for use in research. Int J Med Inform 2016 Jul;90:40-47 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.03.006] [Medline: 27103196]

298. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19(1):54-60 [FREE Full text] [doi: 10.1136/amiajnl-2011-000376] [Medline: 22037893]

299. Denaxas S, Gonzalez-Izquierdo A, Fitzpatrick N, Direk K, Hemingway H. Phenotyping UK Electronic Health Records from 15 Million Individuals for Precision Medicine: The CALIBER Resource. Stud Health Technol Inform 2019 Jul 04;262:220-223. [doi: 10.3233/SHTI190058] [Medline: 31349307]

300. Birtwhistle RV. Canadian Primary Care Sentinel Surveillance Network: a developing resource for family medicine and public health. Can Fam Physician 2011 Oct;57(10):1219-1220 [FREE Full text] [Medline: 21998241]

## Abbreviations

**CHF:** congestive heart failure
**eMERGE:** Electronic Medical Records and Genomics
**EMR:** electronic medical records
**ICD:** International Classification of Diseases
**ICPC:** International Classification of Primary Care
**ML:** machine learning
**NLP:** natural language processing
**PPV:** positive predictive value
**PRISMA-ScR:** Preferred Reporting Items for Systematic reviews and Meta-analyses extension for Scoping Reviews

XSL•FO
RenderX

<u>Original Paper</u>

# Automated Identification of Common Disease-Specific Outcomes for Comparative Effectiveness Research Using ClinicalTrials.gov: Algorithm Development and Validation Study

Anas Elghafari[1], MSc; Joseph Finkelstein[1], MD, PhD

Center for Biomedical and Population Health Informatics, Icahn School of Medicine at Mount Sinai, New York, NY, United States

**Corresponding Author:**

Joseph Finkelstein, MD, PhD
Center for Biomedical and Population Health Informatics
Icahn School of Medicine at Mount Sinai
1425 Madison Ave, Rm L2-36
New York, NY, 10029
United States
Phone: 1 212 659 9596
Fax: 1 212 423 2998
Email: Joseph.Finkelstein@mssm.edu

## *Abstract*

**Background:**   Common disease-specific outcomes are vital for ensuring comparability of clinical trial data and enabling meta analyses and interstudy comparisons. Traditionally, the process of deciding which outcomes should be recommended as common for a particular disease relied on assembling and surveying panels of subject-matter experts. This is usually a time-consuming and laborious process.

**Objective:**   The objectives of this work were to develop and evaluate a generalized pipeline that can automatically identify common outcomes specific to any given disease by finding, downloading, and analyzing data of previous clinical trials relevant to that disease.

**Methods:**   An automated pipeline to interface with ClinicalTrials.gov's application programming interface and download the relevant trials for the input condition was designed. The primary and secondary outcomes of those trials were parsed and grouped based on text similarity and ranked based on frequency. The quality and usefulness of the pipeline's output were assessed by comparing the top outcomes identified by it for chronic obstructive pulmonary disease (COPD) to a list of 80 outcomes manually abstracted from the most frequently cited and comprehensive reviews delineating clinical outcomes for COPD.

**Results:**   The common disease-specific outcome pipeline successfully downloaded and processed 3876 studies related to COPD. Manual verification indicated that the pipeline was downloading and processing the same number of trials as were obtained from the self-service ClinicalTrials.gov portal. Evaluating the automatically identified outcomes against the manually abstracted ones showed that the pipeline achieved a recall of 92% and precision of 79%. The precision number indicated that the pipeline was identifying many outcomes that were not covered in the literature reviews. Assessment of those outcomes indicated that they are relevant to COPD and could be considered in future research.

**Conclusions:**   An automated evidence-based pipeline can identify common clinical trial outcomes of comparable breadth and quality as the outcomes identified in comprehensive literature reviews. Moreover, such an approach can highlight relevant outcomes for further consideration.

## *Introduction*

Over the past decade, comparative effectiveness research has taken center stage as a major vehicle to facilitate informed decisions on optimal treatment regimens as well as a means to improve health care at both the individual and population levels [1]. Comparative effectiveness research has been defined by the Institute of Medicine committee as "the generation and

XSL•FO

**RenderX**

synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat, and monitor a clinical condition or to improve the delivery of care. The purpose of comparative effectiveness research is to assist consumers, clinicians, purchasers, and policy makers to make informed decisions that will improve health care at both the individual and population levels" [2]. Randomized controlled trials (RCTs) are considered the gold standard for clinical effectiveness research [3].

RCTs are conducted to determine whether an intervention is effective by comparing outcomes between different arms of a study that are chosen to reflect beneficial and harmful effects [4]. Results of outcome comparison are used by decision makers to make evidence-based health care choices [5]. Thus, it is critical that study outcomes used in RCTs are relevant for the decision makers and allow cross-trial comparison especially when used to assess the same condition [6]. Recent studies demonstrated inconsistencies in choices of RCT outcomes, which limit potential cross-trial comparison and affect the reproducibility and overall usefulness of RCTs [7]. For example, a comprehensive review of oncology research found that more than 25,000 outcomes were reported only once or twice in oncology trials [8]. Differences in outcome definitions and measurements make it difficult or even impossible to synthesize results of different RCTs [9]. An analysis of missing data in systematic clinical trial reviews found that 71% of reviews could not obtain relevant key outcomes from the included trials [10]. Significant variation in outcome reporting has been noted by a recent systematic review of 109 RCTs assessing interventions for genitourinary symptoms associated with menopause [11].

Misalignment in clinical trial reporting could be addressed by the introduction of an agreed upon collection of common data elements (CDEs) [12]. The importance of developing CDEs for clinical trials, including common disease-specific trial outcomes, has been emphasized by researchers in various fields of medicine and public health [13-16]. Common trial outcomes can help researchers conduct cross-study aggregations and comparisons, facilitate meta-analyses, and increase reproducibility and efficiency. Sheehan et al [15] emphasized the importance of developing CDEs for clinical research generally and noted the current absence of "formal international specifications governing the construction or use of Common Data Elements." Thurmond et al [13] discussed a multiagency scientific initiative to develop CDEs for traumatic brain injury and psychological health and noted that the "use of different measures to assess similar study variables and/or assess outcomes may limit important advances in (...) research. Without a set of common data elements (CDEs; to include variable definitions and recommended measures for the purpose of this discussion), comparison of findings across studies is challenging."

With regard to clinical outcomes specifically (as the primary class of CDEs clinical trials are concerned with), a lot of emphasis has been placed on developing standardization approaches and addressing potential gaps. Ioannidis et al [17] examined the gaps in outcomes reported by clinical trials. In their survey of 174 systematic reviews with 1041 trials on preterm infants, they found that most trials were missing information on serious common outcomes for this population,

and concluded by recommending the "use of standardized clinical outcomes that would have to be collected and reported by default in all trials in a given specialty."

The traditional approach for the development of common outcomes for a particular field involves assembling panels of subject matter experts, who will then embark on an iterative multiphased deliberation process to identify the set of outcomes and agree on definitions and time frames. Redeker et al [16] offer a window into this, describing a process that involves "convening a working group, subdividing the working group based on areas of need, holding an introductory meeting, developing CDEs for assigned areas by subgroups, reviewing the work of all the subgroups, revising the CDEs based on feedback, obtaining public review of the identified CDEs, revising the CDEs based on feedback, and posting the first versions of the CDEs on the website." Typically, this time-consuming and labor-intensive process does not employ automated or data-driven methods to systematically utilize information from ClinicalTrials.gov on the thousands of clinical trials relevant to the conditions under consideration.

ClinicalTrials.gov is the most comprehensive international clinical trial registry that contains over 350,000 research studies from 216 countries [18]. Registration with ClinicalTrials.gov includes submission of verified, detailed, and structured information pertinent to clinical trial design, study timeline, inclusion/exclusion criteria, primary and secondary outcomes, subject follow-up, and trial results. Data from ClinicalTrials.gov have found a variety of innovative uses in biomedical informatics research. For example, Huser and Cimino have worked to link ClinicalTrials.gov to PubMed to analyze the proportion of trials that reported results through publication [19] and to understand the quality and completeness of the links [20]. Anderson et al [21] used ClinicalTrials.gov data to study level of compliance with result reporting requirements. Bourgeois et al [22] used ClinicalTrials.gov data to compare industry-funded trials to nonindustry-funded trials in terms of the likelihood of reporting positive outcomes, while Hartung et al [23] investigated the discrepancies between results submitted to ClinicalTrials.gov's results database and those published in peer-reviewed journals.

ClinicalTrials.gov data mining has been used to analyze the characteristics of oncology clinical trials [8], trends in clinical trials for stroke treatment [24], disparities in racial and ethnic representation in stem cell clinical trials [25], nonpublication rates of registered digital health trials [26], and relationships between mutations and phenotypes [27]. With regard to outcomes and other CDEs, Huser et al [19,20] examined the use of CDEs in real data sets and showed how the CDEs identified change by changing the threshold of commonness. Vodicka et al [28] analyzed the proportion and characteristics of ClinicalTrials.gov trials that included patient-reported outcomes. Luo et al [29] proposed a semiautomatic approach for identifying inclusion criteria CDEs. Mayer et al [30] collected variables from 15 HIV clinical trial dictionaries and clustered them using the Unified Medical Language System (UMLS). These efforts demonstrate the power of automated evidence-based approaches. However, the potential of ClinicalTrials.gov data to inform the development of clinical

trial outcomes has received very little attention in the biomedical informatics literature.

Our aim in this work was to address the lack of automated evidence-based tools in the development of clinical outcomes by introducing our outcome identification pipeline and evaluating the technical correctness of its operations, as well as the quality and relevance of the clinical outcomes identified.

## Methods

### Data Source: ClinicalTrials.gov

The Food and Drug Administration Amendments Act of 2007 (FDAAA) (Section 801 of Public Law 110-85) requires an entity or individual who is responsible for an applicable clinical trial to submit the clinical trial information to the Clinical Trial Registry Data Bank (CTRDB). For the purposes of the FDAAA, ClinicalTrials.gov plays the role of the CTRDB maintained by the National Library of Medicine. ClinicalTrials.gov serves as a mandatory repository for clinical trials conducted under US regulatory oversight. Registration in ClinicalTrials.gov or a similar registry is a prerequisite for publishing clinical trial results in the majority of peer-reviewed journals.

The ClinicalTrials.gov portal supports self-service queries of registered clinical trials through a user interface at the website's main page. The interface allows the user to search for a particular condition or disease by inputting its name into the "condition or disease" field or into the "other terms" field. In the former case, only trials that focus on the condition are returned. In the latter case, more results are returned, but they may not all be relevant to the condition. In parallel to the website, ClinicalTrials.gov offers a RESTful application programming interface (API) that allows automated submission of search queries (eg, from a computer program) and returns results in a computation-friendly format (eg, XML) for further processing.

We implemented the ClinicalTrials.gov query pipeline using Python 3.7 [31] with libraries *URLLIB.request*, *Pandas*, and *Xml.etree*. In what follows, we provide a technical description of the components of the pipeline, representing the logical steps from the input query to the final output, and the list of collated and ranked clinical outcomes. A visual summary of the workflow is provided in Figure 1.

**Figure 1.** Pipeline workflow.
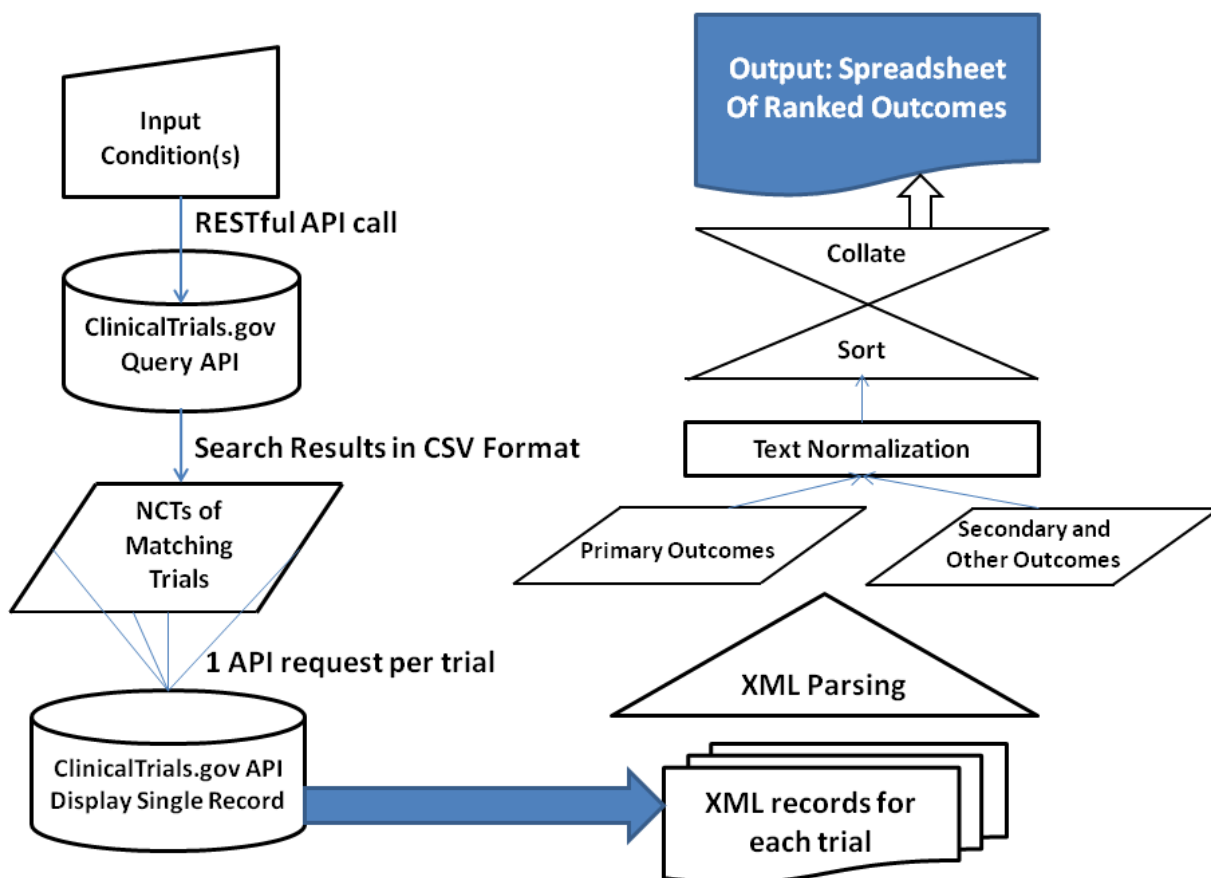
### Step 1: Interfacing With the ClinicalTrials.gov API Search Endpoint

The input term, representing a condition such as "chronic obstructive pulmonary disease (COPD)", is embedded into a URL that is used by the *URLLIB.request* module to interface ClinicalTrials.gov's RESTful API at https://clinicaltrials.gov/ct2/results/download_fields?cond=COPD.

Other parameters of the API call include the following: the number of results to be returned per call (*down_chunk*), the results page (if the total number of matching results exceeds the number of results per page, the results will be broken into several chunks, and each must be accessed with a separate API call that references that page's number), and the format of the results table (*down_fmt*, which can be specified as XML, CSV, PDF, etc).

This call mimics the search for the term using the "*condition or disease*" field on the user interface. By substituting "term" for "cond" in the URL, the call would return the same results obtained by querying for the term using the "*other terms*" field on the web page.

In our implementation, we set the number of *down_count* to 10,000 (the maximum that ClinicalTrials.gov's API allows) and *down_chunk* to 1, which guarantees that most queries will return results contained in one chunk. For the few queries that yield over 10,000 matching trials (eg, "cardiovascular diseases" yields 39,704 results), the pipeline continues incrementing the *down_chunk* parameter and generating a new API call with the updated chunk number until the results are exhausted.

Each call returns a table of results in CSV format. Python's Pandas library is used to parse the table into a Data Frame object. The output from this component is a list of Data Frames, with one for each chunk of each term's results.

### Step 2: Aggregating and Deduplicating the Results

The output of the interface with the search endpoint is a list of Data Frames (tables). Each table stores the details about the trials that match the input condition, and those details include the National Clinical Trial (NCT) number of each trial. The NCT number functions as a unique identifier for a registered study and can be used to download the full record of that study. In the case of multiple tables (due to multiple input terms or multiple pages of results returned by the API), it is necessary to aggregate the NCT numbers from all the tables and remove duplicate NCT numbers if they occur (this happens when the input terms are related, eg, "emphysema" and "COPD," as many trials match both conditions). The output of this step is a list of unique NCT numbers that identify the trials matching the input conditions.

### Step 3: Interfacing With ClinicalTrials.gov to Download the Trial Records

Having arrived at the list of NCT numbers for all the trials in the results, the next step involves interfacing with ClinicalTrials.gov again to download each result in XML. ClinicalTrials.gov allows obtaining a single record in XML by calling https://clinicaltrials.gov/ct2/show/NCT_Number?displayxml=true.

XML is a widely used markup format that most programming languages can work with and parse. Obtaining the trial data in XML sidesteps the challenges of parsing the exact text from a web page. Instead, the XML tree can be searched for the nodes with certain labels (eg, "primary_outcome" and "secondary_outcome"), and the values of those nodes are then immediately accessible in a structured manner. For each NCT number in our aggregated set of results, a URL call is made and the XML record of that study is saved for parsing in the next step.

### Step 4: Parsing the XML for Primary, Secondary, and Other Outcomes and Time Frames

Having downloaded the XML records for the trials that match the input terms, the next step is to parse the clinical outcome names, descriptions, and time frames from the XML.

Our implementation uses Python's built-in *xml.etree* module to parse the XML string into a tree. Then, the *iter()* function is called on the name of the nodes whose values will be extracted (*primary_outcome*, *secondary_outcome*, and *other_outcome*). Each node has further children that record the name of the outcome (*measure*), description, and time frames. The pipeline parses out those elements and stores them along with the NCT number.

### Step 5: Normalizing Outcome Texts and Building a Frequency Table

Following the parsing of outcomes in the previous step, the next step in the pipeline is to normalize those texts of those outcomes, group them, and rank them by frequency. The text normalization step is needed to handle the numerous heterogeneous ways for writing the same outcome name. Given an outcome string, the pipeline applies the following transformations:

1. If the outcome string ends with an abbreviation (letters between parentheses matching the initials of the words, eg, "Quality of life [QoL]"), remove the abbreviation.
2. Change the string to all lower case.
3. Replace all punctuation marks with a space.
4. Replace every occurrence of two or more consecutive space marks with one space mark, and strip the spaces from both ends of the string.

Then, the normalized form of the outcome text is stored in a hash table that maps each outcome string to the list of trials in which it is used. After all the outcome strings are normalized, the table is sorted by frequency of occurrences.

### Step 6: Generating an Output Spreadsheet

In the final step, the pipeline uses the constructed frequency tables to generate a readable CSV spreadsheet of the clinical outcomes for the input condition. The spreadsheet consists of the following three columns: outcome name, number of trials in which it is used, and the NCT IDs of those trials (allowing the user to further explore trials).

### Evaluation Methods

Our evaluation of the pipeline consisted of the following two parts: (1) a technical evaluation that compares the pipeline's output to the data accessible through ClinicalTrials.gov's

website to verify that the downloading and parsing steps are implemented correctly and (2) a systematic evaluation of the outcomes identified by the pipeline for COPD in comparison to the outcomes identified in published widely cited reviews.

### Evaluating the Technical Correctness of the Download and Parsing Processes

We verified the technical correctness of the implementation by comparing the trials downloaded via the API to the trials that can be obtained from the ClinicalTrials.gov website's self-service query interface. This involved verifying that for each query condition, the pipeline was downloading the number of trials as the number that appeared on the website when manually searching for that condition.

In addition, we verified that the XML parsing and collating were correct by sampling pairs of outcomes and NCT identifiers from the resulting spreadsheet, accessing the ClinicalTrials.gov webpages of those trials, and verifying that all the outcomes listed in the output are present on that page. We similarly evaluated the completeness of the pipeline's output relative to the website by sampling in the other direction. Starting from the results obtainable from the website for a given condition, we sampled various trials and verified that each trial in the sample appeared in the output of the pipeline along with all the outcomes listed for it.

### Evaluating the Quality of Pipeline-Identified Clinical Outcomes

To assess the quality and usefulness of the clinical outcomes that can be automatically identified by the pipeline, we selected COPD as a testing use case. COPD was chosen because of the availability of several frequently cited expert reviews delineating COPD-specific clinical outcomes for clinical research, which could serve as a gold standard for assessing the relevance of clinical outcomes generated by the automated pipeline from ClinicalTrials.gov data.

Clinical outcomes from four published systematic reviews [32-35] were manually abstracted by the authors, resulting in a total of 80 outcomes for COPD clinical trials. These four reviews represent the most widely cited publications systematically analyzing outcome measures in COPD trials during the last 15 years. These reviews were conducted manually by internationally recognized expert teams, and they were based on overall 389 articles referenced in these publications. The automated pipeline used four query terms related to the condition (*COPD*, *chronic obstructive pulmonary disease*, *emphysema*, and *chronic bronchitis*) to generate pipeline-identified outcomes

that were compared to the outcomes manually abstracted from the expert reviews.

### Evaluation Metrics

The quality of the automated ClinicalTrials.gov pipeline for clinical outcome generation was assessed using recall and precision. Treating the literature review outcomes as the gold standard, every pipeline-identified outcome that appeared in the gold-standard set was a "true positive" (TP) prediction, every pipeline-identified outcome not appearing in the gold standard set was a "false positive" (FP) prediction, and every outcome from the literature reviews not identified by the pipeline was a "false negative" (FN) prediction for the pipeline. Recall was considered the ratio of TP to (TP + FN), while precision was the ratio of TP to (TP + FP). Intuitively, recall measures the coverage of our pipeline relative to the benchmark, and a low recall would mean the pipeline is failing to identify many benchmark outcomes. Precision, on the other hand, measures how many of the pipeline outcomes are the same as the benchmark outcomes, and a low precision indicates that the pipeline is identifying many outcomes that do not appear in the benchmark set.

## Results

### Correctness of the Data Downloading and Parsing

In evaluating the technical correctness of the output, we employed a number of testing conditions and terms, and compared the result count from both the pipeline and the website. There was a perfect match between the two in all cases, indicating no loss of data that the pipeline is obtaining from the API as compared to the website. Our evaluation of samples of the outcomes and trials similarly indicated a perfect match between the data obtained from the website and the output of the pipeline, with the only difference being the intentional normalization by the pipeline of the outcome texts described in step 5 of the pipeline operation.

Table 1 provides general statistics related to our application of the pipeline for COPD-related terms. The number of trials collected for each term was the same as can be seen on the ClinicalTrials.gov website on January 22, 2020 (number of trials for a given condition can increase over time as new studies are registered). As can be seen in Table 1, the number of trials generated by querying the "other terms" field was higher than that generated by querying the "condition" field, as the former includes a search of additional fields.

**Table 1.** General statistics.

| Variable | Querying using the "condition" field | Querying using the "other terms" field |
|---|---|---|
| Number of trials downloaded using the query term "COPD" | 3596 | 4201 |
| Number of trials downloaded using the query term "chronic obstructive pulmonary disease" | 2993 | 3341 |
| Number of trials downloaded using the query term "emphysema" | 414 | 592 |
| Number of trials downloaded using the query term "chronic bronchitis" | 182 | 244 |
| Number of unique trials (after removing duplicates) | 3876 | 4450 |
| Number of unique trials with outcomes listed | 3734 | 4299 |
| Percentage of trials listing outcomes | 96.3% | 96.6% |
| Number of primary outcomes parsed | 5856 | 7033 |
| Number of secondary/other outcomes parsed | 16,016 | 18,872 |
| Time required by the pipeline to download and parse the trials | 10 minutes 34 seconds | 13 minutes 35 seconds |

### Comparing the Automatically Identified Clinical Outcomes to Published Reviews

On comparing the outcomes identified automatically by the pipeline to the 80 outcomes abstracted from four widely cited reviews [32-35], we found matches for 74 of the 80 manually abstracted ones, giving the pipeline an overall recall of 92%. Tables 2 and 3 list the top primary and secondary pipeline outcomes, while Table 4 lists the four reviews' outcomes that appeared in more than two reviews. Multimedia Appendix 1 shows the full mapping of the reviews' outcomes to the automatically identified ones.

**Table 2.** Top 15 primary outcomes identified by the pipeline for chronic obstructive pulmonary disease.

| Primary outcome | Occurrences as primary, n |
|---|---|
| Mortality | 32 |
| FEV1[a] | 30 |
| Quality of life | 25 |
| Forced vital capacity | 14 |
| Exercise capacity | 14 |
| Adverse events | 13 |
| Dyspnea | 10 |
| Lung function | 10 |
| COPD[b] assessment test | 9 |
| Endurance time | 9 |
| Functional capacity | 9 |
| Safety | 9 |
| Oxygen saturation | 8 |
| Six-minute walk test | 7 |
| Maximum plasma concentration | 7 |

[a]FEV1: forced expiratory volume in 1 second.

[b]COPD: chronic obstructive pulmonary disease.

**Table 3.** Top 15 secondary outcomes identified by the pipeline, excluding outcomes occurring frequently as primary.

| Secondary outcomes | Occurrences as secondary, n |
| --- | --- |
| Heart rate | 30 |
| Length of hospital stay | 25 |
| St. George's Respiratory Questionnaire | 24 |
| Blood pressure | 21 |
| Physical activity | 20 |
| Inspiratory capacity | 18 |
| Body composition | 18 |
| Time to first COPD[a] exacerbation | 16 |
| Physician's global evaluation | 16 |
| Depression | 15 |
| Body mass index | 15 |
| Hospital anxiety and depression scale | 13 |
| Patient satisfaction | 13 |
| Use of rescue medication | 9 |
| Cost-effectiveness | 9 |

[a]COPD: chronic obstructive pulmonary disease.

**Table 4.** Top outcomes abstracted from published reviews.

| Outcome[a] | Source (references) |
| --- | --- |
| Baseline Dyspnea Index | [32-35] |
| Transition Dyspnea Index | [32-35] |
| Borg Dyspnea Scale | [32-35] |
| Medical Research Council Dyspnea Scale | [32-35] |
| Chronic Respiratory Disease Questionnaire | [32-35] |
| St. George's Respiratory Questionnaire | [32-35] |
| Body mass index, airflow obstruction, dyspnea, and exercise capacity | [32-34] |
| Six-minute walk test | [32-35] |
| Incremental shuttle walk test | [32-35] |
| SpO$_2$: peripheral oxygen saturation | [32-34] |
| Forced expiratory volume in 1 second (FEV1) | [32-35] |
| Forced vital capacity (FVC) | [32-35] |
| FEV1/FVC | [32-35] |
| Static lung volumes | [33-35] |
| Number of exacerbations | [33-35] |
| Mortality | [33-35] |

[a]Outcomes that appear in three or more reviews are shown. The full list of 80 outcomes and their equivalent from the pipeline can be seen in Multimedia Appendix 1.

While calculating the pipeline recall of the pipeline's output, we searched for the 80 outcomes abstracted from the expert reviews and found 74 of them among the automatically generated outcomes, thus yielding recall of 92%. For calculating the pipeline precision as described in the methods section, the entire pipeline output required manual review of all automatically generated outcomes since many of them represented the same concept but were phrased differently and used a different abbreviation or spelling. To streamline this part of the assessment, only outcomes used in four or more clinical trials were considered for grouping, which eventually yielded a total of 96 pipeline outcomes. We evaluated each of those to

see if they had an equivalent among the 80 outcomes abstracted from the expert reviews. Overall, 76 of the grouped outcomes had equivalent counterparts among the outcomes abstracted from the expert reviews, yielding a precision of 79%.

### Examining the Differences Between Pipeline Outcomes and Review Outcomes

To better understand the quality of the pipeline's output, we looked at the difference in results between what the pipeline generated and the outcomes from the literature. Textbox 1 lists the review outcomes that had no equivalent in the pipeline output, while Table 5 lists the top pipeline outcomes that had no equivalent in the abstracted reviews.

**Textbox 1.** Top false negatives (outcomes from the abstracted reviews with no match among pipeline-identified outcomes).

| Outcomes |
| --- |
| Nottingham Health Profile |
| Medical Outcomes Study 6-Item General Health Survey (MOS-6A) |
| Symptom Severity Index |
| Two-minute step-in-place test |
| Time spent in weight-bearing activities |
| Sputum visual analog scale |
| Manchester Respiratory Activities of Daily Living Questionnaire |

**Table 5.** Top false positives (outcomes generated by the pipeline but not appearing in any of the abstracted reviews).

| Outcome | Frequency |
| --- | --- |
| Sleep quality | 14 |
| Self-efficacy | 10 |
| Pharmacokinetics | 10 |
| Berg balance scale | 9 |
| Maximum plasma concentration | 9 |
| Duration of mechanical ventilation | 7 |
| Pulmonary vascular resistance | 7 |
| Diaphragmatic function | 6 |
| Cognitive function | 6 |
| Timed up and go | 5 |
| Patient activation | 5 |
| Neural respiratory drive | 5 |
| Handgrip strength | 4 |
| Short physical performance battery | 4 |
| Severe Respiratory Insufficiency Questionnaire | 4 |

## Discussion

### Principal Findings

We have introduced a general automated pipeline for evidence-based generation of clinical outcomes using data from ClinicalTrials.gov. We evaluated the quality of the generated outcomes for COPD by comparing to a list of outcomes collected from four comprehensive reviews. We found great overlap between the autoidentified outcomes and the manually abstracted ones. Treating the review outcomes as the gold standard, the pipeline results achieve 0.92 recall overall and 0.79 precision for the top outcomes (used in more than three studies).

In investigating the cause for lower precision relative to recall, we examined the FPs (those outcomes that are identified by the pipeline but are not part of the benchmark set). Table 5 lists the most frequent pipeline FPs. We find that most of these FPs appear relevant to the underlying condition (COPD) yet have not been covered in any of the four reviews we considered. This argues that the relatively low precision is not due to the pipeline generating irrelevant outcomes, but rather the pipeline identifying outcomes not included in the benchmark set. This points to the potential of this automated evidence-based

approach to highlight measures and domains that might be underused in the literature.

## Limitations

While the results are encouraging, there are two main limitations to the data-driven evidence-based approach. First, there is a great deal of fragmentation in how the same outcome could be described when the data are entered into the trial registry, which leads to a large number of overlapping outcomes being identified. While the text normalization module can handle surface-level variations, some of the variations will require specialized ontologies (eg, to recognize that "spirometry" and "FEV1" are related outcomes). Some variations will still require some human judgment (eg, should "number of readmissions" and "number of hospitalizations" be grouped together for more compactness of the results or is the semantic difference sufficient to warrant keeping them distinct?).

The second limitation is that while the pipeline can be very useful in giving a data-backed view of the most frequent outcomes, it cannot replace the traditional role of experts in providing guidelines for which outcomes are suitable to use in a given situation. Combining the proposed evidence-based pipeline with expert analysis has the potential to greatly facilitate traditional workflow for CDE development. Recent publication evaluating methodology for the development of clinical outcome sets expressed concerns that the currently accepted methodology relies entirely on agreement and lacks alternatives [36,37]. Methods used in the selection of instruments for outcomes included in trial outcome sets can be improved by including automated means for identifying common disease-specific outcomes used in registered clinical trials [38,39].

There are many studies that used ClinicalTrials.gov data for systematic analysis. However, most of those studies focused on analyzing the quality and compliance of the data on ClinicalTrials.gov. For example, Huser and Cimino linked records of interventional studies to PubMed publications and showed that a large segment of trial sponsors failed to meet their mandate in publishing trial results [20]. Compliance with result reporting obligations was also the focus of the work by Anderson et al [21]. Other studies have also utilized ClinicalTrials.gov data to cluster clinical trials with similar eligibility criteria features [40], to characterize semantic heterogeneity of data elements [41], and to analyze nonpublication rates of registered clinical trials [26].

With respect to CDEs, there have been very few efforts to make use of ClinicalTrials.gov's data. Kentgen et al [42] collected data from patient care forms related to acute coronary syndrome and then used the UMLS to semantically annotate and generate a list of the most common data elements. As in our study, the authors noted a lack of standardized and semantically enriched documentation for clinical outcomes. In another study, Holz et al [43] used UMLS to identify and harmonize a semantic core of CDEs for acute myeloid leukemia. However, neither of these studies used data from ClinicalTrials.gov. Among the few works that made use of trial registry data for CDE identification, Luo et al [29] proposed a semiautomatic approach for identifying disease-specific eligibility criteria. They used UMLS semantic types to parse CDEs from inclusion criteria free text. Their results showed that an automated approach can achieve very good performance compared with human annotators. The main difference between their work and ours is that they focused on eligibility criteria CDEs while we focused on clinical outcomes.

Vodicka et al [28] analyzed the proportion of trials that used patient-reported outcomes. While their work similarly includes parsing of ClinicalTrials.gov data, the focus of their analysis was characterizing the temporal trends of the usage of a predefined class of outcomes and the variation by sponsor type. To the best of our knowledge, this is the first report that focuses on the automated identification of clinical outcomes and evaluates the coverage of the identified outcomes by comparing to comprehensive and widely cited reviews.

## Future Directions

For future work, we plan to address the fragmentation issue by using the UMLS [44] in conjunction with the MetaMap API [45] and ontologies on BioPortal [46] to cluster related outcomes and allow the user to explore them by outcome domain or measure. According to Huser et al [47], optimal analyses of CDEs require engagement of multiple data sources and biomedical ontologies as well as real-world research use cases.

Furthermore, we believe that there is a lot of potential in the other data elements that ClinicalTrials.gov provides. This includes time frames of the outcomes as well as the Medical Subject Heading (MeSH) terms. For the time frames, the pipeline is currently parsing them along with the outlines, and the next step would be to fine tune the parsing and aggregation of time frames to include them in the output of the pipeline. MeSH terms are potentially very useful in aiding the classification and navigation of the variety of extracted outcomes. Since those MeSH terms typically include information about additional conditions, inclusion criteria, and intervention types, grouping the outcomes by the associated MeSH terms can offer the user of the data a way to zoom in and zoom out as needed.

## Conclusions

ClinicalTrials.gov offers a wealth of data that has not been fully utilized. An automated pipeline that leverages these data to identify relevant clinical outcomes for any given condition can greatly aid the traditional processes around clinical outcome selection and facilitate clinical trial fidelity and comparability.

XSL•FO

RenderX

**Conflicts of Interest**

None declared.

Multimedia Appendix 1

Outcomes abstracted from the literature reviews.

[DOCX File , 16 KB - medinform_v9i2e18298_app1.docx ]

**References**

1.  Armstrong K. Methods in Comparative Effectiveness Research. JCO 2012 Dec 01;30(34):4208-4214. [doi: 10.1200/jco.2012.42.2659]

2.  Sox HC, Greenfield S. Comparative effectiveness research: a report from the Institute of Medicine. Ann Intern Med 2009 Aug 04;151(3):203-205 [FREE Full text] [doi: 10.7326/0003-4819-151-3-200908040-00125] [Medline: 19567618]

3.  Williams CM, Skinner EH, James AM, Cook JL, McPhail SM, Haines TP. Comparative effectiveness research for the clinician researcher: a framework for making a methodological design choice. Trials 2016 Aug 17;17(1):406 [FREE Full text] [doi: 10.1186/s13063-016-1535-6] [Medline: 27530915]

4.  Clarke M. Standardising outcomes for clinical trials and systematic reviews. Trials 2007 Nov 26;8(1):39 [FREE Full text] [doi: 10.1186/1745-6215-8-39] [Medline: 18039365]

5.  Clarke M, Williamson PR. Core outcome sets and systematic reviews. Syst Rev 2016 Jan 20;5:11 [FREE Full text] [doi: 10.1186/s13643-016-0188-6] [Medline: 26792080]

6.  Williamson PR, Altman DG, Blazeby JM, Clarke M, Devane D, Gargon E, et al. Developing core outcome sets for clinical trials: issues to consider. Trials 2012 Aug 06;13(1):132 [FREE Full text] [doi: 10.1186/1745-6215-13-132] [Medline: 22867278]

7.  Tunis SR, Clarke M, Gorst SL, Gargon E, Blazeby JM, Altman DG, et al. Improving the relevance and consistency of outcomes in comparative effectiveness research. J Comp Eff Res 2016 Mar;5(2):193-205 [FREE Full text] [doi: 10.2217/cer-2015-0007] [Medline: 26930385]

8.  Hirsch BR, Califf RM, Cheng SK, Tasneem A, Horton J, Chiswell K, et al. Characteristics of oncology clinical trials: insights from a systematic analysis of ClinicalTrials.gov. JAMA Intern Med 2013 Jun 10;173(11):972-979. [doi: 10.1001/jamainternmed.2013.627] [Medline: 23699837]

9.  Gargon E, Gurung B, Medley N, Altman DG, Blazeby JM, Clarke M, et al. Choosing important health outcomes for comparative effectiveness research: a systematic review. PLoS One 2014 Jun 16;9(6):e99111 [FREE Full text] [doi: 10.1371/journal.pone.0099111] [Medline: 24932522]

10. Kirkham JJ, Gargon E, Clarke M, Williamson PR. Can a core outcome set improve the quality of systematic reviews?--a survey of the Co-ordinating Editors of Cochrane Review Groups. Trials 2013 Jan 22;14(1):21 [FREE Full text] [doi: 10.1186/1745-6215-14-21] [Medline: 23339751]

11. Christmas M, Song B, Bell R, Iliodromiti S, Mitchell C, Hickey M, International COMMA (Core OutcoMes in MenopAuse) Consortium. Variation in outcome reporting and measurement tools in clinical trials of treatments for genitourinary symptoms in peri- and postmenopausal women: a systematic review. Menopause 2020 Sep;27(9):1070-1080. [doi: 10.1097/GME.0000000000001570] [Medline: 32852462]

12. Clarke M, Williamson P. Core outcome sets and trial registries. Trials 2015 May 14;16(1):216 [FREE Full text] [doi: 10.1186/s13063-015-0738-6] [Medline: 25971905]

13. Thurmond VA, Hicks R, Gleason T, Miller AC, Szuflita N, Orman J, et al. Advancing integrated research in psychological health and traumatic brain injury: common data elements. Arch Phys Med Rehabil 2010 Nov;91(11):1633-1636. [doi: 10.1016/j.apmr.2010.06.034] [Medline: 21044705]

14. Duhaime A, Gean AD, Haacke EM, Hicks R, Wintermark M, Mukherjee P, Common Data Elements Neuroimaging Working Group Members, Pediatric Working Group Members. Common data elements in radiologic imaging of traumatic brain injury. Arch Phys Med Rehabil 2010 Nov;91(11):1661-1666. [doi: 10.1016/j.apmr.2010.07.238] [Medline: 21044709]

15. Sheehan J, Hirschfeld S, Foster E, Ghitza U, Goetz K, Karpinski J, et al. Improving the value of clinical research through the use of Common Data Elements. Clin Trials 2016 Dec 09;13(6):671-676 [FREE Full text] [doi: 10.1177/1740774516653238] [Medline: 27311638]

16. Redeker NS, Anderson R, Bakken S, Corwin E, Docherty S, Dorsey SG, et al. Advancing Symptom Science Through Use of Common Data Elements. J Nurs Scholarsh 2015 Sep 06;47(5):379-388 [FREE Full text] [doi: 10.1111/jnu.12155] [Medline: 26250061]

17. Ioannidis JPA, Horbar JD, Ovelman CM, Brosseau Y, Thorlund K, Buus-Frank ME, et al. Completeness of main outcomes across randomized trials in entire discipline: survey of chronic lung disease outcomes in preterm infants. BMJ 2015 Jan 26;350(jan26 7):h72-h72 [FREE Full text] [doi: 10.1136/bmj.h72] [Medline: 25623087]

18. Zarin DA, Tse T, Williams RJ, Califf RM, Ide NC. The ClinicalTrials.gov Results Database — Update and Key Issues. N Engl J Med 2011 Mar 03;364(9):852-860. [doi: 10.1056/nejmsa1012065]

19. Huser V, Cimino JJ. Linking ClinicalTrials.gov and PubMed to track results of interventional human clinical trials. PLoS One 2013 Jul 9;8(7):e68409 [FREE Full text] [doi: 10.1371/journal.pone.0068409] [Medline: 23874614]

20. Huser V, Cimino JJ. Precision and negative predictive value of links between ClinicalTrials.gov and PubMed. AMIA Annu Symp Proc 2012;2012:400-408 [FREE Full text] [Medline: 23304310]

21. Anderson ML, Chiswell K, Peterson ED, Tasneem A, Topping J, Califf RM. Compliance with Results Reporting at ClinicalTrials.gov. N Engl J Med 2015 Mar 12;372(11):1031-1039. [doi: 10.1056/nejmsa1409364]

22. Bourgeois FT, Murthy S, Mandl KD. Outcome reporting among drug trials registered in ClinicalTrials.gov. Ann Intern Med 2010 Aug 03;153(3):158-166 [FREE Full text] [doi: 10.7326/0003-4819-153-3-201008030-00006] [Medline: 20679560]

23. Hartung DM, Zarin DA, Guise J, McDonagh M, Paynter R, Helfand M. Reporting Discrepancies Between the ClinicalTrials.gov Results Database and Peer-Reviewed Publications. Ann Intern Med 2014 Apr 01;160(7):477. [doi: 10.7326/m13-0480]

24. Negoro T, Okura H, Maehata M, Hayashi S, Yoshida S, Takada N, et al. Trends in clinical trials for stroke by cell therapy: data mining ClinicalTrials.gov and the ICTRP portal site. NPJ Regen Med 2019 Nov 06;4(1):20 [FREE Full text] [doi: 10.1038/s41536-019-0082-7] [Medline: 31728206]

25. Parvanova I, Finkelstein J. Disparities in Racial and Ethnic Representation in Stem Cell Clinical Trials. Stud Health Technol Inform 2020 Jun 26;272:358-361. [doi: 10.3233/SHTI200569] [Medline: 32604676]

26. Al-Durra M, Nolan RP, Seto E, Cafazzo JA, Eysenbach G. Nonpublication Rates and Characteristics of Registered Randomized Clinical Trials in Digital Health: Cross-Sectional Analysis. J Med Internet Res 2018 Dec 18;20(12):e11924 [FREE Full text] [doi: 10.2196/11924] [Medline: 30485832]

27. Alag S. Unique insights from ClinicalTrials.gov by mining protein mutations and RSids in addition to applying the Human Phenotype Ontology. PLoS ONE 2020 May 27;15(5):e0233438. [doi: 10.1371/journal.pone.0233438]

28. Vodicka E, Kim K, Devine E, Gnanasakthy A, Scoggins J, Patrick D. Inclusion of patient-reported outcome measures in registered clinical trials: Evidence from ClinicalTrials.gov (2007-2013). Contemp Clin Trials 2015 Jul;43:1-9. [doi: 10.1016/j.cct.2015.04.004] [Medline: 25896116]

29. Luo Z, Miotto R, Weng C. A human-computer collaborative approach to identifying common data elements in clinical trial eligibility criteria. J Biomed Inform 2013 Feb;46(1):33-39 [FREE Full text] [doi: 10.1016/j.jbi.2012.07.006] [Medline: 22846169]

30. Mayer CS, Williams N, Gold S, Fung KW, Huser V. Sharing of Individual Participant Data from Clinical Trials: General Comparison and HIV Use Case. AMIA Annu Symp Proc 2019;2019:647-654 [FREE Full text] [Medline: 32308859]

31. Ekmekci B, McAnany CE, Mura C. An Introduction to Programming for Bioscientists: A Python-Based Primer. PLoS Comput Biol 2016 Jun;12(6):e1004867 [FREE Full text] [doi: 10.1371/journal.pcbi.1004867] [Medline: 27271528]

32. Oliveira A, Marques AS. Outcome Measures Used in Pulmonary Rehabilitation in Patients With Acute Exacerbation of Chronic Obstructive Pulmonary Disease: A Systematic Review. Phys Ther 2018 Mar 01;98(3):191-204. [doi: 10.1093/ptj/pzx122] [Medline: 29228288]

33. Gross NJ. Chronic obstructive pulmonary disease outcome measurements: what's important? What's useful? Proc Am Thorac Soc 2005;2(4):267-71; discussion 290. [doi: 10.1513/pats.200504-036SR] [Medline: 16267347]

34. Glaab T, Vogelmeier C, Buhl R. Outcome measures in chronic obstructive pulmonary disease (COPD): strengths and limitations. Respir Res 2010 Jun 17;11:79 [FREE Full text] [doi: 10.1186/1465-9921-11-79] [Medline: 20565728]

35. Gross NJ. Outcome Measures for COPD Treatments: A Critical Evaluation. COPD: Journal of Chronic Obstructive Pulmonary Disease 2004 Apr 20;1(1):41-57. [doi: 10.1081/copd-120030413]

36. Gargon E, Williamson PR, Blazeby JM, Kirkham JJ. Improvement was needed in the standards of development for cancer core outcome sets. J Clin Epidemiol 2019 Aug;112:36-44 [FREE Full text] [doi: 10.1016/j.jclinepi.2019.04.006] [Medline: 31009657]

37. Beune I, Ganzevoort W, Gordijn S. Core outcome sets are valuable, but methodological evidence can improve robustness. BJOG 2020 Nov 29;127(12):1527-1527 [FREE Full text] [doi: 10.1111/1471-0528.16419] [Medline: 32672876]

38. Gorst SL, Prinsen CA, Salcher-Konrad M, Matvienko-Sikar K, Williamson PR, Terwee CB. Methods used in the selection of instruments for outcomes included in core outcome sets have improved since the publication of the COSMIN/COMET guideline. J Clin Epidemiol 2020 Sep;125:64-75 [FREE Full text] [doi: 10.1016/j.jclinepi.2020.05.021] [Medline: 32470621]

39. Gargon E, Williamson PR, Young B. Improving core outcome set development: qualitative interviews with developers provided pointers to inform guidance. J Clin Epidemiol 2017 Jun;86:140-152 [FREE Full text] [doi: 10.1016/j.jclinepi.2017.04.024] [Medline: 28495644]

40. Hao T, Rusanov A, Boland MR, Weng C. Clustering clinical trials with similar eligibility criteria features. J Biomed Inform 2014 Dec;52:112-120 [FREE Full text] [doi: 10.1016/j.jbi.2014.01.009] [Medline: 24496068]

41. Brown W, Weng C, Vawdrey D, Carballo-Diéguez A, Bakken S. SMASH: A Data-driven Informatics Method to Assist Experts in Characterizing Semantic Heterogeneity among Data Elements. AMIA Annu Symp Proc 2016;2016:1717-1726 [FREE Full text] [Medline: 28269930]

42. Kentgen M, Varghese J, Samol A, Waltenberger J, Dugas M. Common Data Elements for Acute Coronary Syndrome: Analysis Based on the Unified Medical Language System. JMIR Med Inform 2019 Aug 23;7(3):e14107 [FREE Full text] [doi: 10.2196/14107] [Medline: 31444871]

XSL•FO

RenderX

43.  Holz C, Kessler T, Dugas M, Varghese J. Core Data Elements in Acute Myeloid Leukemia: A Unified Medical Language System-Based Semantic Analysis and Experts' Review. JMIR Med Inform 2019 Aug 12;7(3):e13554 [FREE Full text] [doi: 10.2196/13554] [Medline: 31407666]

44.  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]

45.  Aronson AR, Lang F. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010 May 01;17(3):229-236 [FREE Full text] [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]

46.  Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res 2011 Jul 14;39(Web Server issue):W541-W545 [FREE Full text] [doi: 10.1093/nar/gkr469] [Medline: 21672956]

47.  Huser V, Amos L. Analyzing Real-World Use of Research Common Data Elements. AMIA Annu Symp Proc 2018;2018:602-608 [FREE Full text] [Medline: 30815101]

## Abbreviations

**API:** application programming interface
**CDE:** common data element
**COPD:** chronic obstructive pulmonary disease
**CTRDB:** Clinical Trial Registry Data Bank
**FDAAA:** Food and Drug Administration Amendments Act
**FN:** false negative
**FP:** false positive
**MeSH:** Medical Subject Headings
**NCT:** National Clinical Trial
**RCT:** randomized controlled trial
**TP:** true positive
**UMLS:** Unified Medical Language System

XSL•FO
**RenderX**

# Identifying Myocardial Infarction Using Hierarchical Template Matching–Based Myocardial Strain: Algorithm Development and Usability Study

Jayendra Maganbhai Bhalodiya[1], BE, MPhil; Arnab Palit[2], BE, MTech, PhD; Gerard Giblin[3], MBBCh; Manoj Kumar Tiwari[4], BE, MTech, PhD; Sanjay K Prasad[3], MBChB, MD; Sunil K Bhudia[3], BSc, MBBS, FRCS, MD; Theodoros N Arvanitis[1], RT, DPhil; Mark A Williams[2], BEng, MSc, PhD

[1]Institute of Digital Healthcare, Warwick Manufacturing Group, University of Warwick, Coventry, United Kingdom

[2]Warwick Manufacturing Group, University of Warwick, Coventry, United Kingdom

[3]Royal Brompton and Harefield NHS Foundation Trust, London, United Kingdom

[4]National Institute of Industrial Engineering, Mumbai, India

**Corresponding Author:**
Jayendra Maganbhai Bhalodiya, BE, MPhil
Institute of Digital Healthcare
Warwick Manufacturing Group
University of Warwick
Gibbet Hill Rd
Coventry, CV4 7AL
United Kingdom
Phone: 44 7448404975
Email: jayendra.bhalodiya@warwick.ac.uk

## *Abstract*

**Background:**   Myocardial infarction (MI; location and extent of infarction) can be determined by late enhancement cardiac magnetic resonance (CMR) imaging, which requires the injection of a potentially harmful gadolinium-based contrast agent (GBCA). Alternatively, emerging research in the area of myocardial strain has shown potential to identify MI using strain values.

**Objective:**   This study aims to identify the location of MI by developing an applied algorithmic method of circumferential strain (CS) values, which are derived through a novel hierarchical template matching (HTM) method.

**Methods:**   HTM-based CS H-spread from end-diastole to end-systole was used to develop an applied method. Grid-tagging magnetic resonance imaging was used to calculate strain values in the left ventricular (LV) myocardium, followed by the 16-segment American Heart Association model. The data set was used with k-fold cross-validation to estimate the percentage reduction of H-spread among infarcted and noninfarcted LV segments. A total of 43 participants (38 MI and 5 healthy) who underwent CMR imaging were retrospectively selected. Infarcted segments detected by using this method were validated by comparison with late enhancement CMR, and the diagnostic performance of the applied algorithmic method was evaluated with a receiver operating characteristic curve test.

**Results:**   The H-spread of the CS was reduced in infarcted segments compared with noninfarcted segments of the LV. The reductions were 30% in basal segments, 30% in midventricular segments, and 20% in apical LV segments. The diagnostic accuracy of detection, using the reported method, was represented by area under the curve values, which were 0.85, 0.82, and 0.87 for basal, midventricular, and apical slices, respectively, demonstrating good agreement with the late-gadolinium enhancement–based detections.

**Conclusions:**   The proposed applied algorithmic method has the potential to accurately identify the location of infarcted LV segments without the administration of late-gadolinium enhancement. Such an approach adds the potential to safely identify MI, potentially reduce patient scanning time, and extend the utility of CMR in patients who are contraindicated for the use of GBCA.

XSL•FO
RenderX

## Introduction

### Background

Cardiovascular diseases (CVDs) account for 31% of global deaths [1]. Among CVDs, myocardial infarction (MI) can result from chronic progressive coronary atheromatous disease, with subsequent plaque rupture and thrombosis. Depending on the extent of infarction, there is potential for adverse myocardial remodeling, and subsequently, heart failure [2]. If a patient has MI, with either reduced left ventricular (LV) function or ongoing chest pain, clinicians need to evaluate ischemia and myocardial viability to determine whether there is a potential benefit to be derived from revascularization with either coronary artery bypass graft (CABG) or percutaneous coronary intervention (PCI) [2]. Ischemia can be demonstrated using cardiac magnetic resonance (CMR) imaging by assessing the first-pass perfusion, following the use of a vasodilator stress agent such as adenosine [3]. As a part of the myocardial viability test, it is crucial to show the location and extent of the infarcted myocardium [3].

### Clinical Practice and Literature Review

The current gold standard [3] for determining the location and extent of infarcted myocardium is late gadolinium enhancement (LGE), which acquires delayed CMR images following gadolinium-based contrast agent (GBCA) administration. A GBCA shows the infarcted myocardium with brighter image intensity compared with a noninfarcted myocardium [3]. However, GBCA-administered LGE has the following limitations: (1) the risk of nephrogenic systemic sclerosis in patients with advanced renal impairment [4], (2) concerns of gadolinium accumulation in tissues in normal renal patients [4], and (3) prolonged scan time resulting in adverse situations, such as panic attacks, especially in claustrophobic patients [5]. Therefore, a method that allows infarction detection, without the need for GBCA administration, has the potential to reduce patient scanning time and extend the use of CMR imaging in a wider patient population.

Strain-based techniques, such as speckle tracking and CMR-feature tracking (CMR-FT), have been previously reported to identify infarction [6-8]. Strain-based techniques do not require the use of GBCA. Hence, the aforementioned limitations of LGE could be avoided. Speckle tracking has limited accuracy [7] and CMR-FT is prone to endocardium and epicardium definition because it only uses vessel boundaries and does not include accurate details of structural deformation within the myocardium as grid-tagging magnetic resonance imaging (MRI) does [7]. The LV myocardium has a helical structure of muscles and different mechanics throughout the myocardium [9]. Therefore, the proposed method uses a novel algorithmic hierarchical template matching (HTM)–based diagnosis method [10,11], which calculates myocardium strain values by considering the details of structural deformation within the myocardium for infarction detection. The HTM method provides higher accuracy in calculating two-dimensional (2D) CMR tagging LV strain when compared with the benchmark nonrigid registration using the free-form deformation algorithm, which has been reported to be the most accurate in comparison with other state-of-the-art methods, including optical flow, harmonic phase, and B-snake grid [12,13].

### Aim of the Study

The aim of this study aims to develop an applied algorithmic method to identify MI using HTM-based circumferential strain (HTM-CS) values. The purpose of this work is to show the performance of HTM-CS-based infarction detection with respect to the gold-standard LGE-based detections.

## Methods

### Data Collection and Preparation

A data set of 38 patients with MI and 5 healthy volunteers was collected from the CMR unit of the Royal Brompton and Harefield National Health Service (NHS) Trust (RBHT) through SB. Ethical approval for retrospective data collection was obtained from the NHS (IRAS project ID: 211977). Additionally, Biomedical and Scientific Research Ethics Committee (BSREC) approval (REGO 2016–1865) was obtained from the University of Warwick to process the anonymized data.

All participants were selected retrospectively. Patients with an MI identified on CMR imaging were determined from the referral details and scan reports. The inclusion criteria were as follows: (1) a patient with a known history of infarcted myocardium, or (2) a patient referred for a clinically indicated CMR imaging scan, on the basis of symptoms suggestive of myocardial ischemia, with or without an elevation in serum troponin levels and with a confirmed myocardial infarct on the subsequent CMR imaging. Initially, 55 patients were screened for the study. Among them, 38 patients with MI were included. From the CMR referral details, these 38 patients with MI had clinical conditions such as heavy chest pain, high troponin findings, previous history of known infarction, or intervention history of CABG or PCI. Patients with nonischemic cardiomyopathy or normal findings without infarction were excluded. The characteristic details of these 38 patients with MI are summarized in Table 1. The included patients underwent a standard departmental CMR using either a vasodilator stress perfusion protocol or a viability protocol, both of which included comprehensive late gadolinium enhancement imaging. All participants with MI had anonymized images of LGE imaging and grid-tagging MRI. The images were acquired with three different 1.5T Siemens MRI scanners with ECG triggering. LGE images were acquired with sequences that allowed normal breathing of the patient and had infarcted myocardium with high-intensity values due to postgadolinium enhancement. Grid-tagging MRI was acquired with breath-holds, using a grid structure of myocardial tagging lines with a spacing of 6 mm. LGE images were not available for healthy participants.

**Table 1.** Characteristic details of the data set (N=43).

| Participant types and characteristics | Value |
| --- | --- |
| **Patients with MI[a] (n=38, 88%)** | |
| **Sex, n (%)** | |
| Male | 33 (87) |
| Female | 5 (13) |
| Age (years), mean (SD) | 63.65 (13.31) |
| Height (m), mean (SD) | 1.74 (0.10) |
| Weight (kg), mean (SD) | 84.42 (19.79) |
| Pixel size (mm$^2$) | 1.32×1.32 to 1.75×1.75 |
| **Healthy participants (n=5, 12%)** | |
| **Sex (male or female)** | |
| Male | 2 (40) |
| Female | 3 (60) |
| Age (years), mean (SD) | 41.20 (14.38) |
| Height (m), mean (SD) | 1.68 (0.08) |
| Weight (kg), mean (SD) | 65.2 (13.04) |
| Pixel size (mm) | 1.48×1.48 |

[a]MI: myocardial infarction.

In all the data subjects, images from 3 short-axis (SAX) planes of the LV were processed: basal, midventricular, and apical SAX planes of LV. Basal refers to the LV slice near the mitral valve and before the beginning of the papillary muscle, midventricular refers to the LV slice at the approximate middle of papillary muscle length, and apical refers to the LV slice towards the apex but above the apex. This definition is based on the literature [14,15]. The MRI SAX plane covers many anatomical details of the chest. Therefore, to efficiently process data, the LV area of each image was cropped using ImageJ [16] software. When cropping the images, we ensured that all images for a participant were coregistered well with each other and had the same image dimensions. The cropped images were processed to have normalized (zero-mean unit-variance) intensity values. The images were noise-free, that is, they did not contain imaging artifacts or motion blur artifacts.

In patients with MI, the findings of LGE imaging were used as a ground truth to differentiate between the infarcted and normal (or healthy) myocardium when an area of hyperintensity was present. LV segments were defined according to the American Heart Association (AHA) [14]. In healthy participants, all LV segments were considered healthy. The total number of infarcted segments in 38 patients with MI was 109 (basal: 38, midventricular: 44, apical: 27). The total number of healthy segments was 579 (basal: 220, midventricular: 214, apical: 145), which inlcuded 499 segments from patients with MI and 80 segments from healthy participants.

## Model of Applied HTM Method

The flowchart of the model is shown in Figure 1.

**Figure 1.** Flowchart of the proposed applied method to identify infarcted LV segments using circumferential strain values. Here, $H_{healthy\_same}$ could be $H_{healthy\_basal}$ or $H_{healthy\_mid}$ or $H_{healthy\_apical}$. HA: American Heart Association; CS: circumferential strain; $ES_{calculated}$: median of calculated circumferential strain values at the end-systolic frame; $ES_{literature}$: median of literature referred circumferential strain values at the end-systolic frame; LV: left ventricular or left ventricle.
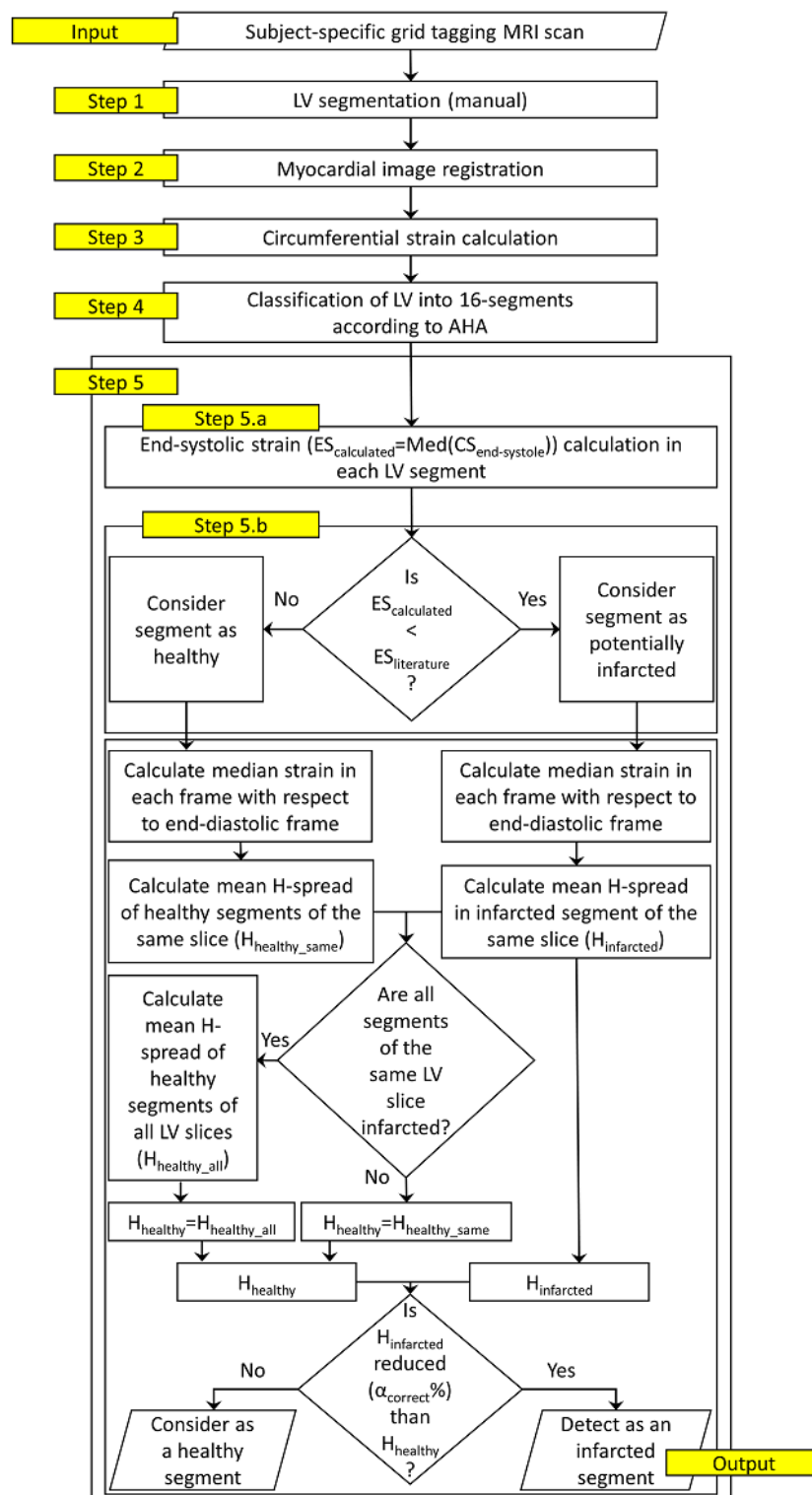


### Step 1—Segmentation of LV Myocardium

The LV myocardium was manually segmented using ImageJ [16] at the end-diastolic frames of the basal, midventricular, and apical SAX planes. The segmentation was verified by a clinical expert.

### Step 2—Myocardial Image Registration

Myocardial image registration was performed using the HTM [10] method to perform myocardial tracking and strain calculation. As shown in Figure 2, HTM takes a stack of MRI with segmented myocardium as inputs. The image textures in all the images from end-diastole to end-systole were tracked

using HTM. The tracking provided the position of each myocardial point during a cardiac cycle. The tracking was formulated by calculating a difference vector, $V(x,y)$, between the initial position, $P_{end-diastole}(x_1, y_1)$, and current position, $P_{current-frame}(x_2, y_2)$, of a myocardial point (Equation 1). The displacement gradient at the end-systolic frame was calculated using the spatial positions at end-diastole and end-systole (Equation 2). $\nabla V$ refers to the displacement gradient.

$$V(x, y) = P_{end-diastole}(x_1, y_1) - P_{current-frame}(x_2, y_2)$$
(1)



**Figure 2.** (a) MRI scans of a cardiac cycle at three LV levels: Basal, Mid-ventricular, and Apical. (b) HTM method to calculate strain values at each muscle point. V is mentioned in (Equation 2). (c) LV strain values, which are analyzed using the 16-segment AHA model. AHA: American heart association; HTM: hierarchical template matching; LV: left ventricle.
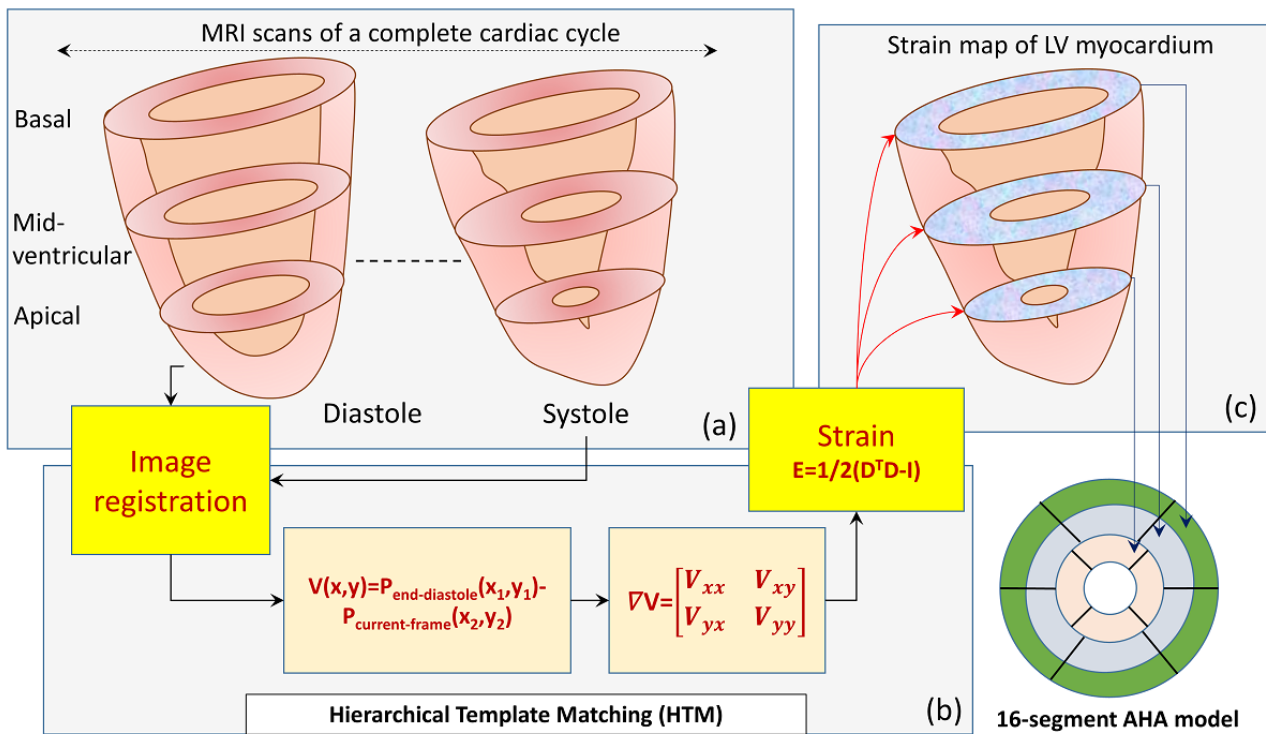


### Step 3—Myocardial Strain Calculation

Equation 3 and Equation 4 show the deformation gradient $D$ and myocardial strain (Green-Lagrange strain [17]) $E$, with respect to the end-diastolic frame, respectively, where I is the identity matrix.

$$D = (I - \nabla V)^{-1}$$ (3)

$$E = 1/2 (D^T D - I)$$ (4)

Strain values at each myocardial point were calculated using a local coordinate system [18]. Only CS was used for analysis, as discussed in the Discussion section. The positive direction of CS is in the short-axis plane and counterclockwise when observed from the base. It is parallel to the surface of the epicardium and perpendicular to the long axis of the LV.

### Step 4—Classification of LV into 16-Segments

The LV myocardium was divided into 16 segments according to the 16-segment AHA model [14]. The basal, midventricular, and apical slices were divided into six, six, and four segments, respectively. Thereafter, the CS in each segment was used during Step 5 to separate healthy and infarcted segments.

### Step 5—Detection of Infarcted Segments

**Initial Detection Using End-Systolic HTM-CS**

The median end-systolic strain in each of the AHA LV segments was calculated as $ES_{calculated} = Med(CS_{end-systole})$. Here, a median of strain values was used to avoid the influence of any large outliers. Subsequently, the calculated median strain was compared with the literature benchmark CS values ($ES_{literature}$) for healthy segments of LV, as mentioned in the literature [18] (Table 2). If the $Med(CS_{end-systole})$ value was less than the benchmark CS values of the healthy myocardium, the segment was considered as a "potentially infarcted" segment (Figure 1).

**Table 2.** End-systolic circumferential strain in healthy left ventricular myocardium from literature.

| Left ventricular slice | Anteroseptal strain, mean (SD) | Inferoseptal strain, mean (SD) | Anterior strain, mean (SD) | Anterolateral strain, mean (SD) | Inferolateral strain, mean (SD) | Inferior strain, mean (SD) |
|---|---|---|---|---|---|---|
| Basal | –0.17 (0.03) | –0.17 (0.03) | –0.20 (0.03) | –0.21 (0.03) | –0.21 (0.03) | –0.16 (0.03) |
| Midventricular | –0.16 (0.03) | –0.16 (0.03) | –0.23 (0.04) | –0.22 (0.03) | –0.22 (0.03) | –0.16 (0.05) |
| Apical | –0.18 (0.03) | –0.18 (0.03) | –0.24 (0.06) | –0.24 (0.04) | –0.24 (0.04) | –0.23 (0.04) |

### Final Detection Using H-spread of HTM-CS

Infarcted myocardium does not contract or lengthen like healthy muscles because of the presence of fibrosis [19]. This property was quantified by the H-spread of the HTM-CS distribution.

H-Spread of the HTM-CS values in each AHA segment was calculated by the union of median strain values in each frame from end-diastole to end-systole. For example, the H-spread of an infarcted segment (Figure 3; Equation 5) was calculated by the union of values p1,p2,…,p8 (Figure 3). Similarly, the H-spread of the healthy segment was calculated by the union of values q1,q2,…,q8 (Figure 3). The MATLAB function iqr() was used to calculate the H-spread [20].



where ES = end-systolic frame, ED = end-diastolic frame, and $S_{fi} = Med(CS_{end-systole})$ at the $i^{th}$ frame.

**Figure 3.** Representation of an infarcted and a healthy segment using circumferential strain. (a) End-systolic strain in a healthy segment (q8) and an infarcted segment (p8). (b) pi and qi represent median strain values at the ith frame of a cardiac cycle. (c) Strain HSpread is calculated using (Equation 5).
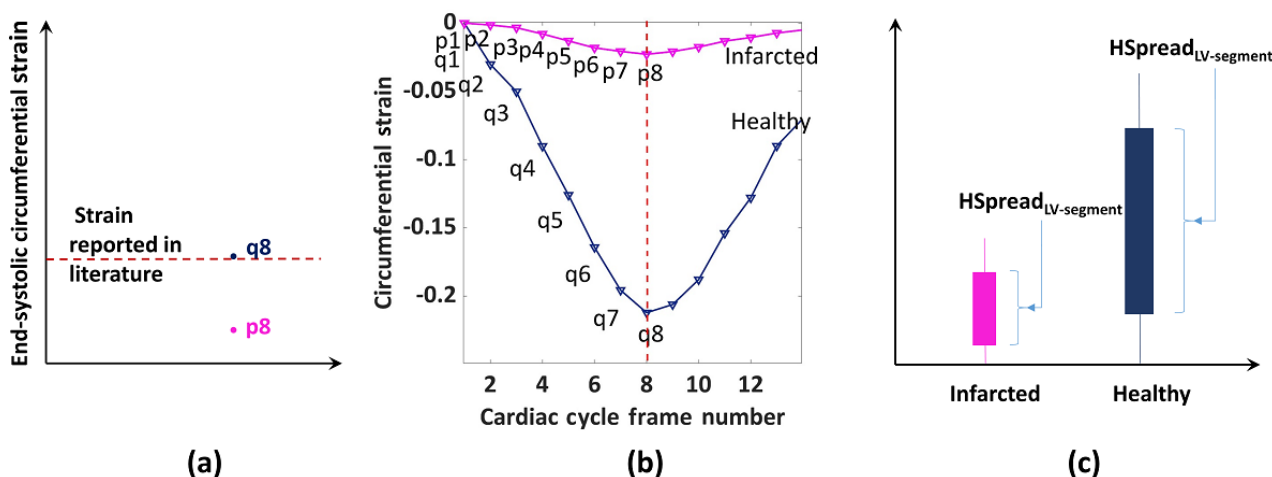


After calculating the H-spread of "potentially infarcted" segments, *"Infarction Condition"* was checked to decide whether the "potentially infarcted" segment was actually an infarcted segment; during this step, the H-spread values were also calculated, as described in Equation 6.

$$H_{infarcted} = HSpread_{LV-segment}$$

$$H_{healthy\_basal} = \mu_{basal}(HSpread_{LV-segment})$$

$$H_{healthy\_mid} = \mu_{mid}(HSpread_{LV-segment})$$

$$H_{healthy\_apical} = \mu_{apical}(HSpread_{LV-segment})$$

$$H_{healthy\_all} = \mu_{basal-mid-apical}(HSpread_{LV-segment})$$

$$H_{healthy} = H_{healthy\_basal} \ OR \ H_{healthy\_mid} \ OR \ H_{healthy\_apical}$$
$$OR \ H_{healthy\_all} \ (6)$$

### Infarction Condition

A potentially infarcted segment has at least $\alpha\%$ reduced strain H-spread compared to the mean H-spread of remaining LV segments of the same LV slice. If all segments of a slice are infarcted, then healthy segments of the whole LV are considered instead of only considering the same slice.

$\alpha$ could be any value greater than 0 and less than 100. For example, assume that segments 1 and 2 of a basal slice are "potentially infarcted" segments. These segments will be considered as actually infarcted if $HSpread_{LV-segment}$ is at least $\alpha\%$ reduced compared with the mean H-spread of other segments in the basal slice ($H_{healthy\_basal}$), which includes segments 3, 4, 5, and 6. When all segments of a slice were "potentially infarcted," healthy segments of basal, midventricular, and apical slices were used to calculate ($H_{healthy\_all}$). For example, if segments 13, 14, 15, and 16 of an apical slice were "potentially infarcted," the healthy segments of basal and midventricular slices were considered together during H-spread comparison of the "Infarction Condition."

In this work, the following values of $\alpha$ were considered (Equation 7) to determine the most appropriate $\alpha_{correct}$ value, and it was assumed that the values within this range of $\alpha$ values would not change the results considerably.

$$\alpha = \{10, 20, 30, \ldots, 100\} \ (7)$$

The detection accuracy using different α values is discussed in the Results section. The α value corresponding to the highest accuracy was called α<sub>correct</sub>, which was then used to detect the MI. Tests were performed using k-fold cross-validation [21,22].

## K-Fold Cross-Validation for Model Training and Testing

In each patient, infarcted segments (AHA segments) were identified using LGE CMR imaging by a cardiothoracic consultant and surgeon who had more than 10 years of experience. These infarcted LV segments were used as the ground truth to validate the proposed HTM-CS–based predictions. Therefore, the validation hypothesis for the statistical analysis was that the infarcted LV myocardial AHA segments which were identified using HTM-CS would be the same as the ground truth of LGE. This section explains k-fold cross-validation [22] using the receiver operating characteristics (ROC) curve test [23,24], which was used in this study to validate the proposed method.

The area under the curve (AUC) of the true negative rate (sensitivity) and the false-positive rate (1-specificity) were calculated from ROC curve tests [23,24], as prediction performance criteria, where AUC 1.0 is the highest accuracy and AUC 0.5 is the lowest accuracy. ROC tests were performed using the MATLAB function perfcurve() [25]. The ROC tests had a 95% CI. Data were prepared by dividing each LV into 16 AHA segments, and the segments were arranged as per basal, midventricular, and apical slices. The total number of basal segments was 258 ($258 = 43 \times 6$), the total number of midventricular segments was 258 ($258 = 43 \times 6$), and the total

number of apical segments was 172 ($172 = 43 \times 4$). Each segment was assigned a label, as infarcted or healthy, according to the LGE ground truth. Then, HTM-CS H-spread reduction was assigned to each "potentially infarcted" segment (healthy segments were considered with 0% H-spread reduction).

During the k-fold tests, *"Infarction Condition"* was evaluated using each α value of Equation 7. Each test assigned a score to each segment as infarcted or healthy. For example, an evaluation test with α=10 scored a potentially infarcted LV segment as infarcted if it satisfied the *"Infarction Condition;"* otherwise, it was scored as healthy. These scores and the ground truth labels of each segment were given as input to the ROC test, as mentioned previously. The α value was selected as α<sub>correct</sub> if the corresponding ROC test had the highest AUC. This α<sub>correct</sub> value was used with the test data set during the k-fold test to calculate the final accuracy.

The training data set and test data set were split as per k-fold cross-validation tests. Initially, k=5 and k=10 were considered, as suggested in the literature [22]. After that, k-fold tests were performed with k=10. When the experiments were performed with the test data set using k=10, we noted that the results were less realistic in the case of k=10. For example, for k=10 using test data, the ROC tests had an AUC of 1.0 (ie, 100% accuracy) in 3 tests (1 out of 10 in each of basal, midventricular, and apical LV slices) and AUC of 0.5 (ie, 0% accuracy) in 1 test (1 out of 10 in apical slices). However, an AUC of 1.0 and 0.5 were not found in the case of k=5. Therefore, we used k=5, and the k-fold tests were repeated 10 times in each basal, midventricular, and apical LV slice with a random selection of data. Figure 4 shows the analysis with k=5.

**Figure 4.** Overall results of 10 k-fold cross-validation tests. Overall AUC values using each α are shown in (a) basal, (b) mid-ventricular, and (c) apical slices. Red circles denote α<sub>correct</sub> corresponding to the maximum AUC for the respective slice. Examples of ROC test results for (d) basal, (e) mid-ventricular, and (f) apical slices, respectively. AUC: area under the curve; ROC: receiver operating characteristic.
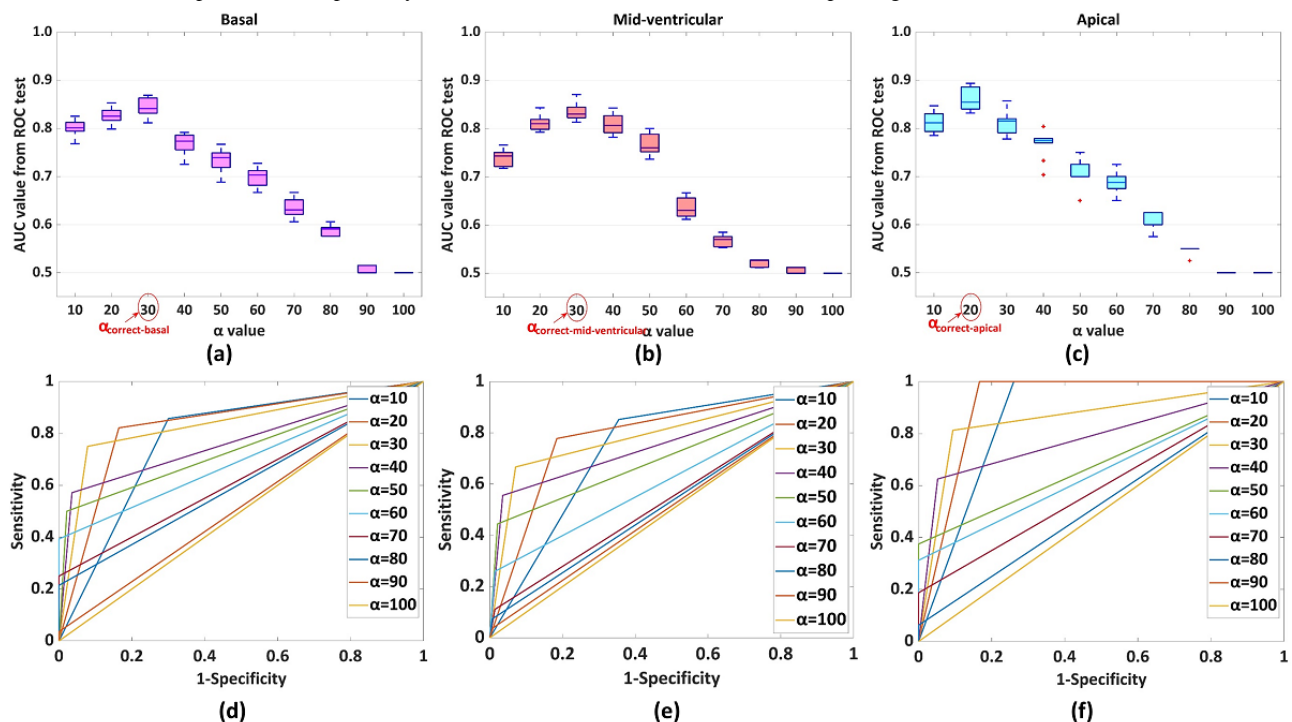
Using the test data set, true positives (infarcted segments detected as infarcted), true negatives (noninfarcted segments detected as noninfarcted), false positives (noninfarcted segments detected as infarcted), and false negatives (infarcted segments detected as noninfarcted) were calculated. Additionally, the sensitivity (true positive rate) and 1-specificity (false negative rate) of the detection of infarcted segments are provided.

## Results

### Primary Analysis of the Model Using Training Data

To find $\alpha_{correct}$ among $\alpha$ values, a training data set of k-fold cross-validation test was used. Figure 4 presents the AUC values in each ROC test for the basal, midventricular, and apical slices using the training data set. $\alpha_{correct}$ was selected from the ROC test, corresponding to the maximum AUC. Accordingly, $\alpha_{correct-basal}$, $\alpha_{correct-midventricular}$, and $\alpha_{correct-apical}$ were found. Figure 4 shows the maximum AUC at $\alpha=30$, midventricular at

$\alpha=30$, and apical at $\alpha=20$. Therefore, $\alpha_{correct-basal}=30$, $\alpha_{correct-midventricular}=30$, and $\alpha_{correct-apical}=20$. Further, k-fold cross-validation tests were performed with test data and $\alpha_{correct}$ values.

### Accuracy Analysis of HTM-CS-Based Model Using Test Data

To analyze the performance of the HTM-CS-based method, the test data set was used. Each -fold cross-validation test used a random 5% sample as test data. $\alpha_{correct-basal}=30$, $\alpha_{correct-midventricular}=30$, and $\alpha_{correct-apical}=20$ were considered with "Infarction Condition" to predict each segment as healthy or infarcted. The results are plotted in Figure 5. In Figure 5, the results of the ROC tests were derived using the MATLAB function perfcurve() [25]. The basal, midventricular, and apical areas had AUC values of 0.85, 0.82, and 0.87, respectively. Table 3 shows an example of a detection.

**Figure 5.** (a) Average of 10 k-fold cross-validation tests. Accuracy of detecting infarcted segments using the proposed method with $\alpha_{correct}$ in basal, mid-ventricular, and apical slices using test data from the k-fold cross-validation tests. (b) An example of an ROC test result using the test data. AUC: area under the curve; ROC: receiver operating characteristic.
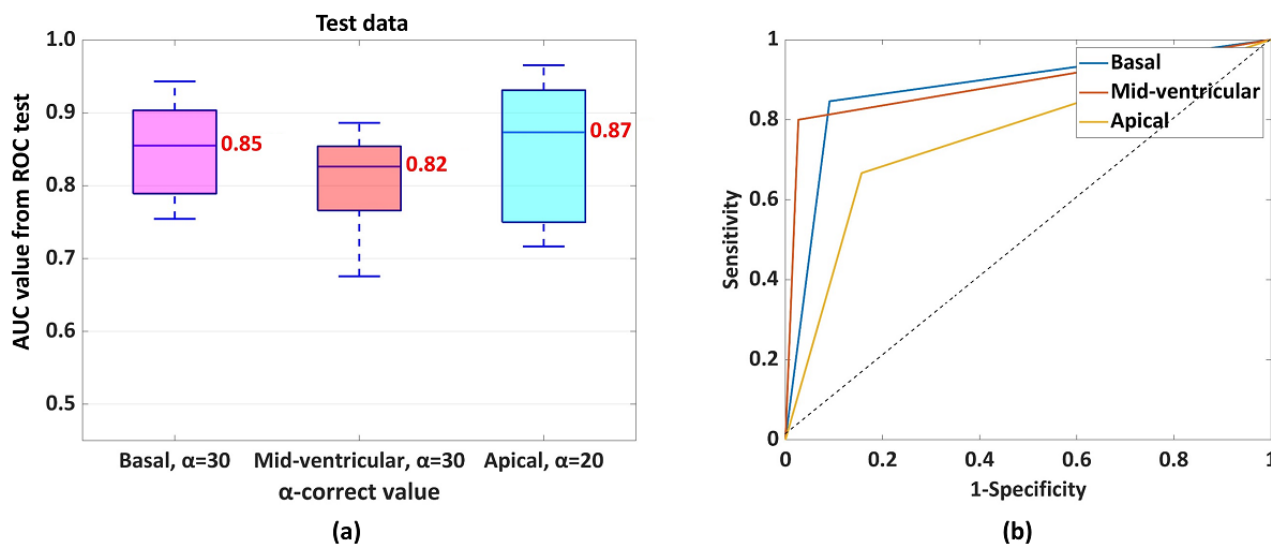


**Table 3.** Analysis of a myocardial infarction patient. Truth and Detected show the results of late gadolinium enhancement and the proposed method, respectively.

| Patient number | Truth | | Detected | |
|---|---|---|---|---|
| | Infarcted? | Which segment? | Infarcted? | Which segment? |
| **Patient number D9** | | | | |
| Basal | Yes | 1,2 | Yes | 1,2 |
| Midventricular | Yes | 7,8 | Yes | 7,8,12 |
| Apical | Yes | 13,14 | Yes | 13,14,15 |

"Truth" shows infarcted segments using LGE and "Detected" shows infarcted segments using the proposed method. The best case would be to have the same "Truth" and "Detected" segments in all three slices. Table 4 summarizes the detections in test patients.

The higher true-positive rate and the lower false-positive rate together determine the best result. For detections in a patient

with MI, the best case should have true-positive rate=1 and false-positive rate=0, and in a healthy participant, false-positive rate=0 should be the best case.

Figure 6 shows an example of detecting infarcted LV segments using the proposed algorithmic method compared with the LGE method. As shown in the LGE images, the white area of the LV myocardium (due to gadolinium deposition) highlights

infarction. In strain analysis, red color shows healthy LV segments and white color shows infarcted LV segments. Some of the segments are both red and white. All segments were characterized as healthy or infarcted by considering the H-spread as mentioned in the "Infarction Condition" of the Methods section.

**Table 4.** Results of detecting infarcted left ventricular segments.

| Patient number | Total LV[a] segments | Infarcted segments | TP[b] | TN[c] | FP[d] | FN[e] | Sensitivity (TPR)[f] = TP/(TP + FN) | 1-Specificity (FPR)[g] = FP/(TN + FP)[h] |
|---|---|---|---|---|---|---|---|---|
| D1[i] | 16 | 5 | 3 | 9 | 2 | 2 | 0.75 | 0.18 |
| D2 | 16 | 2 | 1 | 13 | 1 | 1 | 0.5 | 0.07 |
| D3 | 16 | 1 | 1 | 12 | 3 | 0 | 0.5 | 0.2 |
| D4 | 16 | 1 | 1 | 14 | 1 | 0 | 1 | 0.06 |
| D5 | 16 | 2 | 2 | 14 | 0 | 0 | 0.25 | 0 |
| D6 | 16 | 3 | 3 | 12 | 1 | 0 | 1 | 0.07 |
| D7 | 16 | 7 | 5 | 9 | 0 | 2 | 1 | 0 |
| D8 | 16 | 6 | 6 | 8 | 2 | 0 | 0.66 | 0.2 |
| D9 | 16 | 5 | 5 | 11 | 0 | 0 | 1 | 0 |
| H1[j] | 16 | 0 | 0 | 12 | 4 | 0 | NaN[k] | 0.25 |
| H2 | 16 | 0 | 0 | 16 | 0 | 0 | NaN | 0 |

[a]LV: left ventricular.

[b]TP: true positives.

[c]TN: true negatives.

[d]FP: false positives.

[e]FN: false negatives.

[f]TPR: true-positive rate (sensitivity).

[g]FPR: false-positive rate (1-specificity).

[h]The higher sensitivity and the lower (1-specificity) together determines the best result (eg, D7, D9, and H2).

[i]D: diseased.

[j]H: healthy.

[k]NaN: not a number.

**Figure 6.** (a) LGE showing infarcted segments in white; (b) proposed CS-based analysis showing infarcted and healthy LV segments in white and red, respectively. CS: circumferential strain; HTM: hierarchical template matching; LGE: late gadolinium enhancement.
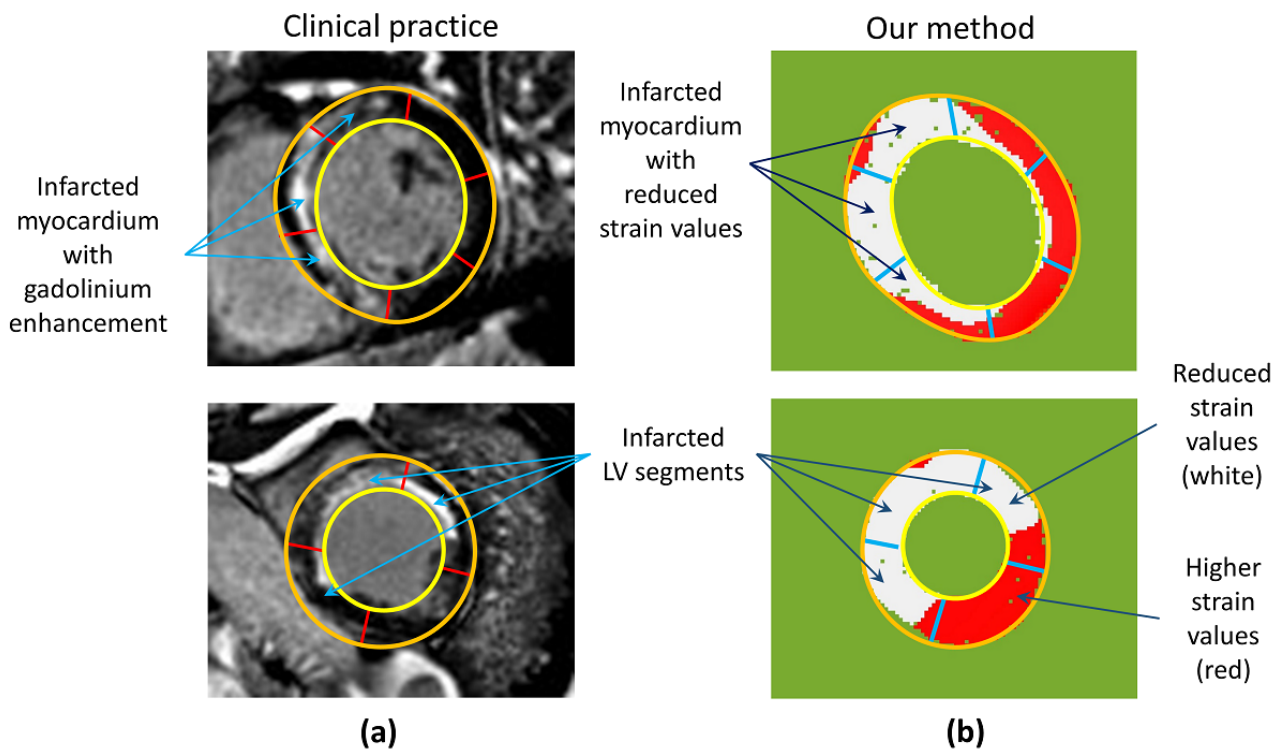


## Discussion

### Principal Findings

This paper elaborates an applied HTM-CS–based infarction detection method, which does not require GBCA administration for conventional late gadolinium CMR imaging. The results demonstrated the promising accuracy of the proposed method, which was compared with the gold-standard method of infarction detection, that is, LGE.

### Data Set

The data set used in this study is a grid-tagging short-axis MRI because the scope of the article is to show that the CS-based applied method can identify MI. This CS can be calculated using grid-tagging CMR images, as reported in the literature [26].

### Model of the Applied HTM Method

In the proposed algorithmic model, the HTM method is used for myocardial tracking. HTM uses normalized cross-correlation in a hierarchical manner to establish point correlations. In this hierarchical matching process, as defined in the article [10], the larger features of images (ie, templates) can correlate well with larger areas of images; however, they are not efficient in matching smaller areas of images. Therefore, HTM has also used smaller image features (ie, segments, chunks, and windows [10]) to correlate smaller areas of images and provide a dense set of correlated points among images. This hierarchical matching is ultimately helpful for myocardial tracking, which is also effective in smaller areas of LV images.

We used CS because HTM uses grid-tagged CMR images in an image texture tracking-based method, and a sequence of grid-tagged CMR images can show better circumferential

movement compared with radial movement. A similar observation is reported in the literature [26], with a potential solution to merge cine CMR images with grid-tagged CMR images to capture and use radial movement more efficiently. However, the scope of this paper is limited to showing the applicability of CS; therefore, we have used only CS.

In Step 5a, the benchmark strain values of Table 2 were obtained from the literature [18]. Table 2 reported strain values for LV septal and lateral areas without subdividing it into inferoseptal, anteroseptal, inferolateral, and anterolateral segments. Therefore, to be able to use literature-referred values, we used the same strain values for both anterolateral and inferolateral segments, and similarly, the same strain values for the anteroseptal and inferoseptal segments.

During Step 5b, the "Infarction Condition" is evaluated, which compares the strain H-spread among the segments of the same LV slice, and in specific conditions, when all segments of an LV slice are infarcted, the healthy segments of other slices are used for H-spread comparison. The clarification is that, in all 43 participants, the overall mean of only healthy basal segments ($\mu(H_{healthy\_basal})$), an overall mean of only midventricular healthy segments ($\mu(H_{healthy\_mid})$), and an overall mean of only apical healthy segments ($\mu(H_{healthy\_apical})$) were mean 0.0772 (SD 0.0372), 0.0862 (SD 0.0366), and 0.0992 (SD 0.0491), respectively. Moreover, the mean H-spread of all healthy segments from all slices together ($\mu(H_{healthy\_all})$) was 0.0851 (SD 0.0271). Here, we performed a paired-sample t-test, which hypothesized that the distribution of differences between each pair of $H_{healthy\_all}$ and $H_{healthy\_apical}$ (or $H_{healthy\_basal}$ or $H_{healthy\_mid}$) is a normal distribution with mean zero and unknown variance. The test does not reject our hypothesis for $P=.04$. Therefore,

we assumed that such differences among them will not change the accuracy of the method considerably, and therefore, for specific conditions when all segments of a slice are infarcted, we have used $H_{healthy\_all}$ during comparison, instead of using only $H_{healthy\_basal}$ or $H_{healthy\_mid}$ or $H_{healthy\_apical}$.

In k-fold cross-validation, the results are derived using k=5 because k=10 has reported a high variance in the prediction (some of the tests have 100% prediction accuracy and some of them have approximately 50%). Therefore, k=5 was used for consistent prediction, as suggested in the literature [22].

## Validation Method

Validation was performed using LGE CMR images, which is a clinical gold standard method for identifying infarction using CMR imaging [3].

## Accuracy of the Applied HTM Method

The infarcted myocardium does not shorten or lengthen the healthy myocardium due to replacement fibrosis, and the average shortening of healthy basal muscles in the circumferential direction is 20% [18,27]. Moreover, due to nonuniform cardiac LV mechanics, basal, midventricular, and apical slices have different average end-systolic shortening of 18.5%, 19.25%, and 22.25%, respectively [18]. Therefore, the proposed method has analyzed detections separately in basal, midventricular, and apical slices to find separate $\alpha_{correct-basal}$, $\alpha_{correct-midventricular}$, and $\alpha_{correct-apical}$. $\alpha_{correct-basal}$=30 and $\alpha_{correct-midventricular}$=30 show that the infarcted LV segments have at least 30% reduced strain H-spread compared with the healthy LV segments in basal and midventricular slices, respectively. Similarly, $\alpha_{correct-apical}$=20 shows that the infarcted LV segments in apical slices have at least 20% reduced strain H-spread compared with healthy LV segments. This difference is due to partially infarcted LV segments. Figure 5 shows different accuracies at different LV levels. Moreover, Table 4 shows that some of the infarcted segments were detected as healthy (false negatives). A possible reason is that the proposed method is fundamentally based on image texture tracking and is sensitive to image quality. LV slices suffer from texture fading due to breathing or blood flow, and motion artifacts due to patient movement. Consequently, the method could not track muscles, which cause an error in strain calculation and ultimately result in incorrect detection. Table 4 shows the results of randomly selected 9 patients with MI and 2 healthy participants. Healthy volunteers do not have infarcted segments. Therefore, true positive and false negative detections were zero, and the true-positive rate was not a number. However, there were false detections in healthy participants, which resulted in a false-positive rate. False positives were due to reduced CS values.

## Clinical Impact

The proposed method could detect infarcted LV segments without using GBCA, which can extend the utility of CMR in conditions such as chronic kidney disease stage 4 or 5 patients (glomerular filtration rate <30 ml/min/1.73 $m^2$ [28]). These patients have a contraindication to the use of GBCA due to the risk of nephrogenic systemic fibrosis. Moreover, the concerns of gadolinium accumulation in normal renal patients can be avoided with the proposed method. The method could potentially reduce scanning time, as it identifies infarction by postimage analysis, and a patient is not required to be inside the scanner for an additional LGE scan, which in most instances requires at least 5 to 10 min extra scanning time following GBCA administration. A study reported that patients who undergo MRI often have claustrophobia, anxiety, and panic attacks (approximately 13%) [5]. Therefore, an overall reduced scanning time may help improve patient care. Moreover, GBCA usage costs an additional €50 (US $61.39) to a patient [29]. Hence, HTM-CS–based analysis could be more economical.

## Future Work and Limitations

The proposed algorithmic method used three 2D slices and a 16-segment AHA model. However, the methodology could adapt to a different number of segments. A higher number of slices could be included after a rigorous literature review to obtain generalized strain values. Hypokinetic segments could be detected as infarcted using the proposed method. The proposed method can locate infarcted LV segments, and further investigations are required to determine the extent of infarction (transmurality). A possible reason is the lower resolution of grid-tagging MRI and faded endocardium and epicardium borders. However, the method could be improved by combining multiple CMR imaging modalities for higher accuracy. The method is not fully automatic; therefore, evaluation at the scanner is not possible at this stage. As the method is semiautomatic and requires image cropping, manual segmentation, and nonrigid image registration, the evaluation time is subjective, such as 5 to 6 hours for a patient. Moreover, images with artifacts due to breath-holding, blood flow, or motion could affect the accuracy of our method.

## Conclusions

In this paper, an applied method for detecting MI based on CS analysis is proposed. The results are compared with the clinical gold-standard (LGE) in detecting MI, and it is observed that the proposed HTM-CS–based approach can provide accurate detections. Moreover, the proposed method avoids the use of GBCA, leading to reduced material cost and scanning time, which may be of particular benefit in individuals with claustrophobia.

unit team members of the RBHT for data collection. The authors also acknowledge the support of the UK Biobank for providing grid-tagging CMR imaging data during the early stage of method development (application number: 32263).

Data Access: The dataset can be accessed through Dr Sunil Bhudia who is affiliated with the Royal Brompton and Harefield NHS Foundation Trust.

## Conflicts of Interest

None declared.

## References

1. Cardiovascular diseases (CVDs). World Health Organization. 2017. URL: http://www.who.int/mediacentre/factsheets/fs317/en/ [accessed 2020-12-28]
2. Hulten E, Carlson D. Invasive coronary angiography. In: Atherosclerosis: Clinical Perspectives Through Imaging. London: Springer London; 2012:77-89.
3. Pennell DJ, Sechtem UP, Higgins CB, Manning WJ, Pohost GM, Rademakers FE, Society for Cardiovascular Magnetic Resonance, Working Group on Cardiovascular Magnetic Resonance of the European Society of Cardiology. Clinical indications for cardiovascular magnetic resonance (CMR): Consensus Panel report. Eur Heart J 2004 Nov;25(21):1940-1965. [doi: 10.1016/j.ehj.2004.06.040] [Medline: 15522474]
4. McDonald RJ, McDonald JS, Kallmes DF, Jentoft ME, Murray DL, Thielen KR, et al. Intracranial gadolinium deposition after contrast-enhanced MR imaging. Radiology 2015 Jun;275(3):772-782. [doi: 10.1148/radiol.15150025] [Medline: 25742194]
5. Thorpe S, Salkovskis PM, Dittner A. Claustrophobia in MRI: the role of cognitions. Magn Reson Imaging 2008 Oct;26(8):1081-1088. [doi: 10.1016/j.mri.2008.01.022] [Medline: 18524527]
6. Crosby J, Amundsen BH, Hergum T, Remme EW, Langeland S, Torp H. 3-D speckle tracking for assessment of regional left ventricular function. Ultrasound Med Biol 2009 Mar;35(3):458-471. [doi: 10.1016/j.ultrasmedbio.2008.09.011] [Medline: 19056164]
7. Mangion K, McComb C, Auger DA, Epstein FH, Berry C. Magnetic resonance imaging of myocardial strain after acute ST-segment-elevation myocardial infarction: a systematic review. Circ Cardiovasc Imaging 2017 Aug;10(8). [doi: 10.1161/CIRCIMAGING.117.006498] [Medline: 28733364]
8. Tanabe Y, Kido T, Kurata A, Sawada S, Suekuni H, Kido T, et al. Three-dimensional maximum principal strain using cardiac computed tomography for identification of myocardial infarction. Eur Radiol 2017 Apr;27(4):1667-1675. [doi: 10.1007/s00330-016-4550-9] [Medline: 27541353]
9. Sengupta PP, Tajik AJ, Chandrasekaran K, Khandheria BK. Twist mechanics of the left ventricle: principles and application. JACC Cardiovasc Imaging 2008 May;1(3):366-376 [FREE Full text] [doi: 10.1016/j.jcmg.2008.02.006] [Medline: 19356451]
10. Bhalodiya JM, Palit A, Tiwari MK, Prasad SK, Bhudia SK, Arvanitis TN, et al. A novel hierarchical template matching model for cardiac motion estimation. Sci Rep 2018 Mar 14;8(1):4475 [FREE Full text] [doi: 10.1038/s41598-018-22543-y] [Medline: 29540762]
11. Bhalodiya JM, Palit A, Ferrante E, Tiwari MK, Bhudia SK, Arvanitis TN, et al. Hierarchical template matching for 3D myocardial tracking and cardiac strain estimation. Sci Rep 2019 Aug 28;9(1):12450 [FREE Full text] [doi: 10.1038/s41598-019-48927-2] [Medline: 31462651]
12. Smal I, Carranza-herrezuelo N, Klein S, Niessen W, Meijering E. Quantitative comparison of tracking methods for motion analysis in tagged MRI. In: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. 2011 Presented at: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro; March 30 to April 2, 2011; Chicago, IL, USA p. 348. [doi: 10.1109/isbi.2011.5872420]
13. Shi W, Zhuang X, Pizarro L, Bai W, Wang H, Tung K, et al. Registration using sparse free-form deformations. Med Image Comput Comput Assist Interv 2012;15(Pt 2):659-666. [doi: 10.1007/978-3-642-33418-4_81] [Medline: 23286105]
14. Cerqueira MD, Weissman NJ, Dilsizian V, Jacobs AK, Kaul S, Laskey WK, American Heart Association Writing Group on Myocardial SegmentationRegistration for Cardiac Imaging. Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart. A statement for healthcare professionals from the Cardiac Imaging Committee of the Council on Clinical Cardiology of the American Heart Association. Circulation 2002 Jan 29;105(4):539-542. [doi: 10.1161/hc0402.102975] [Medline: 11815441]
15. Selvadurai BS, Puntmann VO, Bluemke DA, Ferrari VA, Friedrich MG, Kramer CM, et al. Definition of left ventricular segments for cardiac magnetic resonance imaging. JACC Cardiovasc Imaging 2018 Jun;11(6):926-928 [FREE Full text] [doi: 10.1016/j.jcmg.2017.09.010] [Medline: 29248663]
16. Fiji. ImageJ. URL: https://imagej.net/Fiji [accessed 2020-12-28]
17. Abd-Elmoniem KZ, Stuber M, Prince JL. Direct three-dimensional myocardial strain tensor quantification and tracking using zHARP. Med Image Anal 2008 Dec;12(6):778-786 [FREE Full text] [doi: 10.1016/j.media.2008.03.008] [Medline: 18511332]

XSL•FO

RenderX

18. Moore CC, Lugo-Olivieri CH, McVeigh ER, Zerhouni EA. Three-dimensional systolic strain patterns in the normal human left ventricle: characterization with tagged MR imaging. Radiology 2000 Feb;214(2):453-466 [FREE Full text] [doi: 10.1148/radiology.214.2.r00fe17453] [Medline: 10671594]

19. Pfeffer MA, Braunwald E. Ventricular remodeling after myocardial infarction. Experimental observations and clinical implications. Circulation 1990 Apr;81(4):1161-1172. [doi: 10.1161/01.cir.81.4.1161] [Medline: 2138525]

20. Interquartile range - MATLAB iqr. MathWorks. URL: https://www.mathworks.com/help/stats/prob.normaldistribution.iqr.html [accessed 2020-12-28]

21. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics 2005 Aug 01;21(15):3301-3307. [doi: 10.1093/bioinformatics/bti499] [Medline: 15905277]

22. Kuhn M, Johnson K. Applied Predictive Modeling. Switzerland: Springer Nature; 2013:71.

23. Hajian-Tilaki K. Receiver Operating Characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian J Intern Med 2013;4(2):627-635 [FREE Full text] [Medline: 24009950]

24. Šimundić AM. Measures of diagnostic accuracy: basic definitions. EJIFCC 2009 Jan;19(4):203-211 [FREE Full text] [Medline: 27683318]

25. Receiver operating characteristic (ROC) curve or other performance curve for classifier output. MathWorks. URL: https://uk.mathworks.com/help/stats/perfcurve.html [accessed 2020-12-28]

26. Shi W, Zhuang X, Wang H, Duckett S, Luong DV, Tobon-Gomez C, et al. A comprehensive cardiac motion estimation framework using both untagged and 3-D tagged MR images based on nonrigid registration. IEEE Trans Med Imaging 2012 Jun;31(6):1263-1275. [doi: 10.1109/TMI.2012.2188104] [Medline: 22345530]

27. Tobon-Gomez C, De Craene M, McLeod K, Tautz L, Shi W, Hennemuth A, et al. Benchmarking framework for myocardial tracking and deformation algorithms: an open access database. Med Image Anal 2013 Aug;17(6):632-648. [doi: 10.1016/j.media.2013.03.008] [Medline: 23708255]

28. Thomsen HS, Morcos SK, Almén T, Bellin M, Bertolotto M, Bongartz G, ESUR Contrast Medium Safety Committee. Nephrogenic systemic fibrosis and gadolinium-based contrast media: updated ESUR Contrast Medium Safety Committee guidelines. Eur Radiol 2013 Feb;23(2):307-318. [doi: 10.1007/s00330-012-2597-9] [Medline: 22865271]

29. Boldt J, Leber AW, Bonaventura K, Sohns C, Stula M, Huppertz A, et al. Cost-effectiveness of cardiovascular magnetic resonance and single-photon emission computed tomography for diagnosis of coronary artery disease in Germany. J Cardiovasc Magn Reson 2013 Apr 10;15:30 [FREE Full text] [doi: 10.1186/1532-429X-15-30] [Medline: 23574690]

## Abbreviations

**2D:** two-dimensional
**AHA:** American Heart Association
**AUC:** area under the curve
**CABG:** coronary artery bypass graft
**CMR:** cardiac magnetic resonance
**CMR-FT:** CMR-feature tracking
**CS:** circumferential strain
**CVD:** cardiovascular diseases
**GBCA:** gadolinium-based contrast agent
**HDR:** Health Data Research
**HTM:** hierarchical template matching
**HTM-CS:** HTM-based CS
**LGE:** late gadolinium enhancement
**LV:** left ventricular or left ventricle
**MI:** myocardial infarction
**MRI:** magnetic resonance imaging
**NHS:** National Health Service
**PCI:** percutaneous coronary intervention
**RBHT:** Royal Brompton and Harefield NHS Foundation Trust
**ROC:** receiver operating characteristics
**SAX:** short-axis

Original Paper

# Similarity-Based Unsupervised Spelling Correction Using BioWordVec: Development and Usability Study of Bacterial Culture and Antimicrobial Susceptibility Reports

Taehyeong Kim[1*], BS; Sung Won Han[1*], PhD; Minji Kang[1], BS; Se Ha Lee[2], BS; Jong-Ho Kim[3], PhD; Hyung Joon Joo[3], MD, PhD; Jang Wook Sohn[4], MD, PhD

[1]Division of Industrial Management Engineering, Korea University, Seoul, Republic of Korea

[2]Information Computing Office, Korea University Anam Hospital, Seoul, Republic of Korea

[3]Department of Cardiology, Cardiovascular Center, Korea University College of Medicine, Seongbuk-gu, Republic of Korea

[4]Division of Infectious Disease, Department of Internal Medicine, Korea University College of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

**Corresponding Author:**
Hyung Joon Joo, MD, PhD
Department of Cardiology
Cardiovascular Center
Korea University College of Medicine
73, Inchon-ro
Seongbuk-gu, 02841
Republic of Korea
Phone: 82 2 920 6411
Email: drjoohj@gmail.com

## Abstract

**Background:**   Existing bacterial culture test results for infectious diseases are written in unrefined text, resulting in many problems, including typographical errors and stop words. Effective spelling correction processes are needed to ensure the accuracy and reliability of data for the study of infectious diseases, including medical terminology extraction. If a dictionary is established, spelling algorithms using edit distance are efficient. However, in the absence of a dictionary, traditional spelling correction algorithms that utilize only edit distances have limitations.

**Objective:**   In this research, we proposed a similarity-based spelling correction algorithm using pretrained word embedding with the BioWordVec technique. This method uses a character-level N-grams–based distributed representation through unsupervised learning rather than the existing rule-based method. In other words, we propose a framework that detects and corrects typographical errors when a dictionary is not in place.

**Methods:**   For detected typographical errors not mapped to Systematized Nomenclature of Medicine (SNOMED) clinical terms, a correction candidate group with high similarity considering the edit distance was generated using pretrained word embedding from the clinical database. From the embedding matrix in which the vocabulary is arranged in descending order according to frequency, a grid search was used to search for candidate groups of similar words. Thereafter, the correction candidate words were ranked in consideration of the frequency of the words, and the typographical errors were finally corrected according to the ranking.

**Results:**   Bacterial identification words were extracted from 27,544 bacterial culture and antimicrobial susceptibility reports, and 16 types of spelling errors and 914 misspelled words were found. The similarity-based spelling correction algorithm using BioWordVec proposed in this research corrected 12 types of typographical errors and showed very high performance in correcting 97.48% (based on F1 score) of all spelling errors.

**Conclusions:**   This tool corrected spelling errors effectively in the absence of a dictionary based on bacterial identification words in bacterial culture and antimicrobial susceptibility reports. This method will help build a high-quality refined database of vast text data for electronic health records.

XSL•FO
RenderX

# Introduction

## Background

Among various industries, the medical industry produces many unstructured forms of examination reports. It is very important to establish a structured form of accurate medical documentation to provide accurate diagnoses and treatments to patients [1]. False medical information because of spelling errors can lead to medical and/or treatment errors, resulting in serious risks for patients. For example, errors in the spelling of organism names or drugs with similar spelling in bacterial culture tests have negative effects on not only the diagnosis and treatment of patients, but also the management of infectious diseases and nosocomial infections in hospitals.

While many patient electronic health records are documented in a structured form, the bacterial culture report is still stored as images or as an unrefined text data form in most hospitals. Mapping terms for bacterial identification are necessary to proceed with medical data studies, such as detection and diffusion path studies of infectious diseases. However, since large-scale clinical text data are mostly written by doctors or semiautomatic systems, there can be problems with data consistency, typographical errors, and stop words [2].

In clinical text data, the extraction-transformation-load (ETL) process for medical terms is typically performed through exact string matching of words that appear in the dictionary. However, words not present in the dictionary or severely misspelled words have difficulty matching to terms. Because medical terms are complex and field specific, this problem makes it difficult to apply the same general data refining methods [3]. Rule-based spelling correction algorithms cannot ensure the accuracy and reliability of the data because of incorrect data preprocessing. This method also has to check all test results and find the errors directly, resulting in a considerable cost problem.

## Related Work

### Spelling Correction in the Medical Domain

It is very difficult to construct dictionaries for all medical terms and abbreviations. A related study of spelling correction algorithms specialized in medical record text data was conducted. Lai et al [4] proposed a noisy channel-based spelling check algorithm for medical text. Named entity recognition (NER) was used to achieve an error detection performance of up to 94.4% with a spelling correction accuracy of up to 88.2%, producing high performance spelling correction results in various clinical documents. Fivez et al [5,6] proposed a spelling check algorithm for clinical free text using fastText of the N-gram embedding technique. After generating misspelled words in MIMIC-III [7] to measure similarity with the candidate group that fits the context, the similarity was ranked using the Damerau-Levenshtein distance. This method suggested a way to solve the out-of-vocabulary (OOV) problem in clinical data.

### Subword-Level Word Vector Representation

Traditional spelling correction algorithms using edit distance or pronunciation algorithms have limitations in correcting word-level issues that fit the context. There are subword-level embedding methods for learning concurrent word information to consider context understanding. FastText [8] expresses a word by the sum of the N-gram vector of the character level. The embedding method at the subword level solves the disadvantages that involve difficulty in application to languages with varying morphological changes or low frequency. This method was strong at solving the OOV problem, and accuracy was high for rare words in the word set. BioWordVec [9] learns clinical record data from PubMed and MIMIC-III clinical databases using fastText. Based on 28,714,373 PubMed documents and 2,083,180 MIMIC-III clinical database documents, the entire corpus was built. The Medical Subject Headings (MeSH) term graph was organized to create a heading sequence and to carry out word embedding based on a sequence combining MeSH and PubMed. BioWordVec provided a 200-dimensional pretrained word embedding matrix.

### Limitations With Existing Approaches

The method proposed by Lai et al [4] has a limitation in that spelling corrections are not made in the absence of a dictionary. The method proposed by Fivez et al [5,6] solves the OOV problem, but has a similar limitation in that spelling corrections are not made in the absence of a dictionary.

### Our Approach

This paper proposes a similarity-based spelling correction algorithm through pretrained word embedding in medical field data. Using the BioWordVec model of the character level, which has pretrained clinical record data from the MIMIC-III clinical database, the model progresses learning on spelling corrections end-to-end. The proposed model has the advantage of being able to make spelling corrections in the absence of a dictionary. In addition, it is effective against new types of typographical errors that may occur in the future, and it is highly utilized in the field because it uses unsupervised learning that does not require direct label assignment. We aimed to use this model to develop a spelling correction system suitable for various types of medical text data.

# Methods

## Data Set

### Bacterial Culture and Antimicrobial Susceptibility Reports

In this study, the bacterial culture and antimicrobial susceptibility reports from Korea University Anam Hospital, Korea University Guro Hospital, and Korea University Ansan Hospital were used. The bacterial culture and antimicrobial susceptibility report data were collected for 17 years (from 2002 to 2018), and in each year, reports for 1 month were used for the experiment. In total, 180,000 items were retrieved, with

XSL•FO

**RenderX**

27,544 having meaningful test results. Using the self-developed rule-based ETL algorithm [10], unstructured bacterial culture and antimicrobial susceptibility reports were converted into structured text data. After preprocessing through lexical processing, such as sentence segmentation, tokenization, and stemming using regular expressions, there were 320 types of bacterial identification words in the report. Among the extracted bacterial identification words, 16 types of spelling errors and 914 misspelled words were found. Table 1 presents the typographical errors based on their occurrence.

**Table 1.** Misspelling frequency table.

| Misspelling | Occurrence, n |
| --- | --- |
| staphylococcus | 827 |
| sstreptococcus | 21 |
| adecarboxylate | 19 |
| parpinfluenzae | 18 |
| papatyphi | 7 |
| pseudodiphthericum | 6 |
| urealyticm | 5 |
| chromogens | 2 |
| flavbacterium | 2 |
| ferentum | 1 |
| koneensis | 1 |
| ochrobacterium | 1 |
| orytihabitans | 1 |
| shingobacterium | 1 |
| stacherbrandfii | 1 |
| perosis | 1 |

## Methodology

### Misspelling Detection

Systematized Nomenclature of Medicine (SNOMED) clinical terms (CT) [11] is a set of systematically structured medical terms used in medical clinical documents and reports. It is the world's largest multilingual clinical terminology system. In the corpus constructed by tokenizing the bacterial identification result reports, words that were not mapped to SNOMED CT were defined and detected as typographical errors [12].

### Candidate Generation

Using the fastText [8] technique, prelearned word embedding was used to generate a group of corrected word candidates with high similarity considering the edit distance. In this study, the BioWordVec [9] model that was prelearned from the clinical database was used.

The number of words that were most similar, cosine similarity, and edit distance were set as hyperparameters for generating a correction candidate group. In addition, constraints for candidate words were used based on the dictionary constructed for the existing general terms, the length of the word, and the frequency of the word. In this study, the number of most similar words was set to 30, cosine was set to 0.80, and edit distance was set to 3 as hyperparameters.

Character-based spell checking algorithms were used to determine edit distances to generate or rank candidate groups.

The Levenshtein edit distance [13] is the number of operations required to convert one word into another. It can find the minimum editing distance that considers the insertion, deletion, replacement, and transposition (replacement of two adjacent characters) for most spelling errors. The model proposed in this paper uses the Damerau-Levenshtein distance [14] as the edit distance. The formula is as follows:



### Candidate Ranking

The final correction word is suggested by ranking the correction candidate groups. The pretrained word embedding was learned by the fastText technique, and the vocabulary was sorted in descending order according to frequency. The methodology proposed in this study has two assumptions. First, in clinical databases, correctly spelled words may appear relatively more frequently than misspelled words [15]. Second, the larger the corpus used for learning, the greater the frequency of correctly spelled words [15]. The BioWordVec [9] model used in this research can sufficiently satisfy the above two assumptions.

The model proposed in this research limited the search for the range of the most similar words. Through a grid search, a similarity-based candidate group that considers the frequency of words was proposed [16]. After sorting the ranking of the generated correction candidate words based on similarity, typographical errors can be corrected.

## Overall Architecture

Figure 1 shows the architecture of the spelling correction algorithm proposed in this paper.

**Figure 1.** Similarity-based unsupervised spelling correction architecture. SNOMED: Systematized Nomenclature of Medicine.
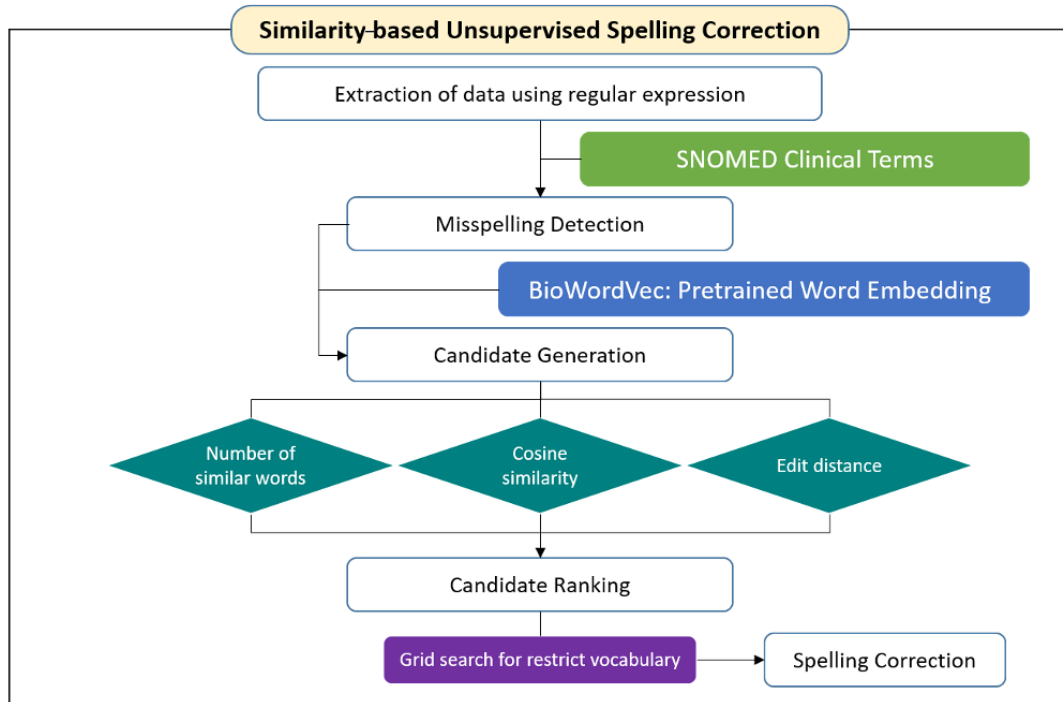


## Results

### Experiments

A typographical error that appears in bacterial culture and antimicrobial susceptibility reports is a word that can be corrected within three edit distances, as shown in Table 2. Most typographical errors have a correctly spelled word within one edit distance. Therefore, in the model proposed in this study, the critical value of the editing distance for generating the correction candidate group was set to 3 or less.

**Table 2.** Correction table using edit distance.

| Correction | Edit distance |
|---|---|
| stapylococcus to staphylococcus | 1 |
| sstreptococcus to streptococcus | 1 |
| adecarboxylate to adecarboxylata | 1 |
| parpinfluenzae to parainfluenzae | 1 |
| papatyphi to paratyphi | 1 |
| pseudodiphtericum to pseudodiphtheriticum | 2 |
| urealyticm to urealyticum | 1 |
| chromogens to chromogenes | 1 |
| flavbacterium to flavobacterium | 1 |
| ferentum to fermentum | 1 |
| koneensis to koreensis | 1 |
| ochrobacterium to ochrobactrum | 2 |
| orytihabitans to oryzihabitans | 1 |
| shingobacterium to sphingobacterium | 1 |
| stacherbrandfii to stackebrandtii | 3 |
| perosis to peroris | 1 |

XSL•FO

**RenderX**

### Comparison of Pretrained Embeddings

All of the pretrained word embeddings used in this study were learned based on the fastText methodology, and the corpus was constructed without distinction between spelling errors and correct spelling during learning. To compare the performance of the BioWordVec model introduced in the previous study, four pretrained embeddings provided by Facebook were used.

The following are the five pretrained embeddings: (1) BioWordVec, 200-dimensional embedding vectors learned using fastText for PubMed and MIMIC-III; (2) English word vectors, 300-dimensional embedding vectors learned using fastText for general text and from Wikipedia; (3) Crawled English subword vectors, 300-dimensional embedding vectors learned using fastText for the 2,000,000 lower words that appear in English word documents; (4) Wiki word vectors, 300-dimensional embedding vectors learned using fastText in

Wikipedia; (5) Simple Wiki word vectors, 300-dimensional embedding vectors learned using fastText in Simple Wikipedia.

The cosine similarity of all models was set to 0.80 or higher, the editing distance threshold was set to 3 or less, and the most similar words were tested under the same conditions with 30 words. The evaluation index is the exact spelling of the total 16 typographical errors that appear in the bacterial assimilation report with correction rate. Table 3 shows the rate of correction for typographical errors according to pretrained embeddings.

The spelling correction algorithm using BioWordVec showed very high performance compared to the performance of the other pretrained word embedding models. The methodology proposed in this study has the advantage of being used even in the absence of a dictionary. However, it was confirmed that pretrained word embedding based on the clinical database is necessary to correct errors in the bacterial identification report.

**Table 3.** Comparison of pretrained embedding.

| Pretrained embedding model | Correction rate |
| --- | --- |
| BioWordVec | 0.75 |
| English word vectors | 0.00 |
| Crawled English subword vectors | 0.00 |
| Wiki word vectors | 0.31 |
| Simple Wiki word vectors | 0.19 |

### Evaluation

Through a comparative experiment as shown in Table 3, it is possible to correct typographical errors using pretrained word embedding without building a dictionary. To evaluate the performance of the model proposed in this study, its performance was compared with a rule-based spelling correction algorithm [17] using a dictionary and a situation without spelling correction. SymSpell [18] was used as a spelling correction algorithm based on the edit distance rule.

SymSpell [18] can correct typographical errors 1 million times faster than rule-based spelling correction [17] and can use existing dictionaries through a symmetric deletion spelling correction algorithm. SymSpell uses the Damerau-Levenshtein edit distance [14], which was set to 3 for the experiment under the same conditions as the model proposed in this study. SCOWL [19] and Dorland medical dictionary [20,21] were used as dictionaries for SymSpell, and a total of 100,000 correct word dictionaries were constructed.

Table 4 shows the evaluation results through the NER task that extracts the bacterial identification words. In the table, accuracy

is the number of words corrected for all misspellings. Precision is the proportion of corrected words that the actual corrections match exactly. Recall is the proportion of correct corrected words among actual typographical errors. F1 score is the harmonic mean of precision and recall. SUSC (similarity-based unsupervised spelling correction) in Table 4 is the model proposed in this study.

In this study, the similarity-based spell checking algorithm SUSC using BioWordVec corrected 12 types of typographical errors and showed very high performance in correcting 97.48% (based on F1 score) of all spelling errors. Both models were able to correct frequent typographical errors, so the overall correction rate was high. However, since SymSpell only corrects certain words, the F1 score showed little difference compared with the nonspelling situation. The Dorland medical dictionary was not able to fully understand bacterial identification names for infectious diseases, and the rule-based spell checking algorithms using edit distance did not work well according to the established dictionaries. Constructing an accurate dictionary that can be used in a rule-based spell checking algorithm is very expensive and time consuming.

**Table 4.** Model performance using BioWordVec.

| Model | Accuracy | Precision | Recall | F1 score |
| --- | --- | --- | --- | --- |
| No spelling correction | 0.98 | 0.94 | 0.93 | 0.94 |
| SymSpell | 1.00 | 0.94 | 0.94 | 0.94 |
| SUSC[a] (BioWordVec) | 1.00 | 0.97 | 0.97 | 0.97 |

[a]SUSC: similarity-based unsupervised spelling correction.

XSL•FO

RenderX

### Comparison of Similarity

Using the SUSC model proposed in this study, the degree of similarity of words depending on correction was examined. Table 5 shows the similarity of words according to whether they are corrected.

As shown in Table 5, typographical errors that were not corrected with the correct spelling have low cosine similarity with the correctly spelled word as a whole. In the case of nonword errors, which involve words that do not actually exist, most of the words were corrected accurately. Miscorrected

typographical errors included real-word errors where the word actually exists but is not appropriate for grammar or context. Since real-word errors are determined to be similar in meaning to words that do not fit the situation, the cosine similarity is relatively low for the word vector to be corrected. The model proposed in this study has the advantage of quantitatively expressing the relative distance between typographical errors and correctly spelled words by utilizing the similarity between words. Through the proposed model, it is possible to compare and determine whether the error detected with the framework is actually a typographical error that can occur often.

**Table 5.** Comparison of similarity according to correction.

| Change | Correction | Similarity |
| --- | --- | --- |
| adecarboxylate to adecarboxylata | Corrected | 0.90 |
| flavbacterium to flavobacterium | Corrected | 0.83 |
| koneensis to koreensis | Corrected | 0.87 |
| ochrobacterium to ochrobactrum | Corrected | 0.93 |
| orytihabitans to oryzihabitans | Corrected | 0.90 |
| papatyphi to paratyphi | Corrected | 0.89 |
| parpinfluenzae to parainfluenzae | Corrected | 0.86 |
| pseudodiphtericum to pseudodiphtheriticum | Corrected | 0.93 |
| shingobacterium to sphingobacterium | Corrected | 0.93 |
| sstreptococcus to streptococcus | Corrected | 0.95 |
| stapylococcus to staphylococcus | Corrected | 0.88 |
| urealyticm to urealyticum | Corrected | 0.84 |
| chromogens to chromogenes | Not corrected | 0.71 |
| ferentum to fermentum | Not corrected | 0.47 |
| perosis to peroris | Not corrected | 0.42 |
| stacherbrandfii to stackebrandtii | Not corrected | 0.59 |

## Discussion

It is difficult to compare our results with previous results because the system implementation and data set used in the related work are not publicly available. The model proposed in this research was capable of spelling correction through unsupervised learning, but it lacked the performance required for infrequent typographical errors and real-word errors. In addition, there was a problem of randomly setting the reference values for cosine similarity and edit distance when creating a correction candidate group. Methods should be devised to establish appropriate thresholds for hyperparameters through experiments.

This research proposes a similarity-based spelling correction algorithm using pretrained word embedding to extract correct

medical terminology from unstructured text data related to infectious diseases. The suggested algorithm has the advantage of being able to check spelling and make corrections in the absence of a correct spelling dictionary. In addition, it solves the OOV problem and can modify words based on context.

As a result of the experiments conducted in this research, we were able to detect and correct spelling errors in the absence of a dictionary for bacterial terms appearing in bacterial culture and antimicrobial susceptibility reports. Our model efficiently refined and processed large medical text data. It has been proven experimentally that it is a method suitable for processing natural language involving high expertise and complexity, such as medical terminology. Ideally, the results of this research will serve as a foundation to build vast amounts of text data in electronic health records into high-quality databases.

XSL•FO

**RenderX**

## Conflicts of Interest

None declared.

## References

1. Wong W, Glance D. Statistical semantic and clinician confidence analysis for correcting abbreviations and spelling errors in clinical progress notes. Artif Intell Med 2011 Nov;53(3):171-180. [doi: 10.1016/j.artmed.2011.08.003] [Medline: 21924593]
2. Hersh WR, Campbell EM, Malveau SE. Assessing the feasibility of large-scale natural language processing in a corpus of ordinary medical records: a lexical analysis. Proc AMIA Annu Fall Symp 1997:580-584 [FREE Full text] [Medline: 9357692]
3. Patrick J, Sabbagh M, Jain S, Zheng H. Spelling correction in clinical notes with emphasis on first suggestion accuracy. In: Proceedings of 2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2010). 2010 Presented at: 2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining; 2010; Malta.
4. Lai KH, Topaz M, Goss FR, Zhou L. Automated misspelling detection and correction in clinical free-text records. J Biomed Inform 2015 Jun;55:188-195 [FREE Full text] [doi: 10.1016/j.jbi.2015.04.008] [Medline: 25917057]
5. Fivez P, Šuster S, Daelemans W. Unsupervised Context-Sensitive Spelling Correction of English and Dutch Clinical Free-Text with Word and Character N-Gram Embeddings. Computational Linguistics in the Netherlands Journal 2017;7:39-52 [FREE Full text]
6. Fivez P, Šuster S, Daelemans W. Unsupervised Context-Sensitive Spelling Correction of Clinical Free-Text with Word and Character N-Gram Embeddings. In: Proceedings of the BioNLP 2017 Workshop. 2017 Presented at: BioNLP 2017 Workshop; August 4, 2017; Vancouver, Canada p. 143-148 URL: https://www.aclweb.org/anthology/W17-2317.pdf [doi: 10.18653/v1/w17-2317]
7. Johnson AEW, Pollard TJ, Shen L, Lehman LH, 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]
8. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. TACL 2017 Dec;5:135-146 [FREE Full text] [doi: 10.1162/tacl_a_00051]
9. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci Data 2019 May 10;6(1):52 [FREE Full text] [doi: 10.1038/s41597-019-0055-0] [Medline: 31076572]
10. CDM_ETL-bacteria. GitHub. URL: https://github.com/KU-RIAS/CDM_ETL-bacteria [accessed 2021-01-08]
11. Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. Proc AMIA Symp 2001:662-666 [FREE Full text] [Medline: 11825268]
12. Zhou X, Zheng A, Yin J, Chen R, Zhao X, Xu W, et al. Context-Sensitive Spelling Correction of Consumer-Generated Content on Health Care. JMIR Med Inform 2015 Jul 31;3(3):e27 [FREE Full text] [doi: 10.2196/medinform.4211] [Medline: 26232246]
13. Levenshtein V. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady 1966 Feb;10:707.
14. Damerau FJ. A technique for computer detection and correction of spelling errors. Commun. ACM 1964 Mar;7(3):171-176. [doi: 10.1145/363958.363994]
15. Crowell J, Zeng Q, Ngo L, Lacroix E. A frequency-based technique to improve the spelling suggestion rank in medical queries. J Am Med Inform Assoc 2004 Jun;11(3):179-185 [FREE Full text] [doi: 10.1197/jamia.M1474] [Medline: 14764616]
16. Pande H. Effective search space reduction for spell correction using character neural embeddings. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. 2017 Presented at: 15th Conference of the European Chapter of the Association for Computational Linguistics; April 2017; Valencia, Spain p. 170-174 URL: https://www.aclweb.org/anthology/E17-2027.pdf [doi: 10.18653/v1/e17-2027]
17. How to Write a Spelling Corrector. Norvig. URL: http://norvig.com/spell-correct.html [accessed 2021-02-16]
18. SymSpell. GitHub. URL: https://github.com/wolfgarbe/SymSpell [accessed 2021-02-16]
19. SCOWL (And Friends). Wordlist. URL: http://wordlist.aspell.net/ [accessed 2021-02-16]
20. Nazir A, Qamar U, Zafar I, Shaheen A, Fatima A, Maqbool B. Automated misspelling detection and correction in clinical free-text records. 2018 Presented at: 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD); May 26-28, 2018; Chengdu p. 277-280. [doi: 10.1109/icaibd.2018.8396209]
21. Dorland. Dorland's Dictionary of Medical Acronyms and Abbreviations - 7th Edition. Amsterdam: Elsevier; 2015.

## Abbreviations

**CT:** clinical terms
**ETL:** extract-transformation-load
**MeSH:** Medical Subject Headings
**NER:** named entity recognition
**OOV:** out-of-vocabulary

XSL•FO

RenderX

**SNOMED:** Systematized Nomenclature of Medicine
**SUSC:** similarity-based unsupervised spelling correction

Original Paper

# Electronic Health Record–Based Prediction of 1-Year Risk of Incident Cardiac Dysrhythmia: Prospective Case-Finding Algorithm Development and Validation Study

Yaqi Zhang[1,2*], PhD; Yongxia Han[1,2*], PhD; Peng Gao[2,3*], PhD; Yifu Mo[1,4*], MS; Shiying Hao[5,6], PhD; Jia Huang[2,7], MD; Fangfan Ye[2,8], MD; Zhen Li[2], PhD; Le Zheng[5,6], PhD; Xiaoming Yao[9], MD, PhD; Zhen Li[2,10], PhD; Xiaodong Li[2,11], PhD; Xiaofang Wang[12], PhD; Chao-Jung Huang[13], PhD; Bo Jin[14], MS; Yani Zhang[15], MD; Gabriel Yang[16], HS; Shaun T Alfreds[17], MBA; Laura Kanov[14], MBA; Karl G Sylvester[2], MD; Eric Widen[14], MHA; Licheng Li[1], PhD; Xuefeng Ling[2,6], PhD

[1]School of Electrical Power Engineering, South China University of Technology, Guangzhou, China

[2]Department of Surgery, Stanford University, Stanford, CA, United States

[3]College of Pharmacy, Shandong University of Traditional Chinese Medicine, Shandong, China

[4]China Southern Power Grid Company Limited, Guangzhou, China

[5]Department of Cardiothoracic Surgery, Stanford University, Stanford, CA, United States

[6]Clinical and Translational Research Program, Betty Irene Moore Children's Heart Center, Lucile Packard Children's Hospital, Palo Alto, CA, United States

[7]Department of Critical Care Medicine, The Third People's Hospital of Shenzhen, The Second Affiliated Hospital of Southern University of Science and Technology, Shenzhen, China

[8]Department of Anesthesiology, The Third People's Hospital of Shenzhen, The Second Affiliated Hospital of Southern University of Science and Technology, Shenzhen, China

[9]Translational Medicine Laboratory, Queen Mary Hospital, Hong Kong University, Hong Kong, China

[10]School of Electrical Engineering, Southeast University, Nanjing, China

[11]School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China

[12]Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan, Jinan, China

[13]National Taiwan University–Stanford Joint Program Office of Artificial Intelligence in Biotechnology, Ministry of Science and Technology Joint Research Center for Artificial Intelligence Technology and All Vista Healthcare, Taipei, China

[14]HBI Solutions Inc, Palo Alto, CA, United States

[15]Tianjin Yunjian Medical Laboratory Institute Co Ltd, Tianjing, China

[16]The Harker School, San Jose, CA, United States

[17]HealthInfoNet, Portland, ME, United States

*these authors contributed equally

**Corresponding Author:**
Xuefeng Ling, PhD
Department of Surgery
Stanford University
S370 Grant Building
300 Pasteur Drive
Stanford, CA, 94305
United States
Phone: 1 6504279198
Email: bxling@stanford.edu

## *Abstract*

**Background:** Cardiac dysrhythmia is currently an extremely common disease. Severe arrhythmias often cause a series of complications, including congestive heart failure, fainting or syncope, stroke, and sudden death.

**Objective:** The aim of this study was to predict incident arrhythmia prospectively within a 1-year period to provide early warning of impending arrhythmia.

**Methods:**   Retrospective (1,033,856 individuals enrolled between October 1, 2016, and October 1, 2017) and prospective (1,040,767 individuals enrolled between October 1, 2017, and October 1, 2018) cohorts were constructed from integrated electronic health records in Maine, United States. An ensemble learning workflow was built through multiple machine learning algorithms. Differentiating features, including acute and chronic diseases, procedures, health status, laboratory tests, prescriptions, clinical utilization indicators, and socioeconomic determinants, were compiled for incident arrhythmia assessment. The predictive model was retrospectively trained and calibrated using an isotonic regression method and was prospectively validated. Model performance was evaluated using the area under the receiver operating characteristic curve (AUROC).

**Results:**   The cardiac dysrhythmia case-finding algorithm (retrospective: AUROC 0.854; prospective: AUROC 0.827) stratified the population into 5 risk groups: 53.35% (555,233/1,040,767), 44.83% (466,594/1,040,767), 1.76% (18,290/1,040,767), 0.06% (623/1,040,767), and 0.003% (27/1,040,767) were in the very low-risk, low-risk, medium-risk, high-risk, and very high-risk groups, respectively; 51.85% (14/27) patients in the very high-risk subgroup were confirmed to have incident cardiac dysrhythmia within the subsequent 1 year.

**Conclusions:**   Our case-finding algorithm is promising for prospectively predicting 1-year incident cardiac dysrhythmias in a general population, and we believe that our case-finding algorithm can serve as an early warning system to allow statewide population-level screening and surveillance to improve cardiac dysrhythmia care.

**KEYWORDS**

## Introduction

Cardiac dysrhythmia is a series of conditions in which the heartbeat is irregular, too fast, or too slow. There are many types of dysrhythmias, and most are mild; however, some severe arrhythmias increase the risk of serious or even life-threatening complications, such as congestive heart failure, syncope, stroke, and sudden death. More than 850,000 people are hospitalized for arrhythmias each year in the United States [1]. Sudden cardiac death is the cause of approximately half of the deaths due to cardiovascular disease and approximately 15% of all deaths globally [2]. Approximately 80% of sudden cardiac deaths are caused by ventricular arrhythmias [3]. If the risk and severity of cardiac arrhythmias can be accurately predicted, actionable medical treatments can be applied to proactively reduce incidence and prevent disease deterioration.

A few arrhythmia risk prediction tools have been applied in programs for screening, prevention of life-threatening arrhythmias [4], and selection of therapy and intervention [5]. Most models were developed for specific populations or special conditions: a prediction model was developed in a consecutive cohort of 1138 patients who underwent carotid endarterectomy [6] with a C statistic of 0.69 (0.64-0.73); a $^{123}$I-metaiodobenzylguanidine single photon emission computed tomography model had an area under the receiver operating characteristic curve (AUROC) of 0.76 [7]; and a cardiac magnetic resonance imaging with late gadolinium enhancement model, with AUROC values ranging from 0.721 to 0.812 [8] for different scar characteristics, relied on specially captured information that was not available for risk prediction in a large-population cohort. The main limitations of these models included small sample sizes from a single source of data, lack of consideration of the interactions among multiple risk factors, and insufficient real-time monitoring of predictor changes [6,9].

To date, no well-performing and widely recognized risk assessment model has been implemented for clinical application in a large general population.

With the widespread use of electronic health records (EHRs) in hospitals and clinics, an individual's physical and mental condition can be assessed to potentially improve the effectiveness of health management [10]. Individuals' comprehensive clinical histories have been used to build risk models with various risk factors [11]. The multidimensional clinical data elements and the generality of the EHR-based data sets are promising to extract more comprehensive risk patterns [12]. Empowered by statewide health information exchange platforms, we applied advanced machine learning including deep learning analytics to deliver actionable information that can help health care organizations identify high-risk individuals, which could improve patients' health and lower costs [13,14].

The purpose of this study was to retrospectively develop and prospectively validate our case-finding algorithm for patients at risk of 1-year incident cardiac dysrhythmia in Maine, United States.

## Methods

### Ethics Statement

Protected personal health information was deidentified for model development. Due to the nature of the development with deidentified data, this study was exempted from ethics review by the Stanford University Institutional Review Board (October 16, 2017).

### Experimental Design and Workflow

A complete workflow (data collection, exclusion, and application) is presented in Figure 1. Figure 2 illustrates the detailed modeling process with an ensemble learning method.

**Figure 1.** A workflow diagram that describes data collection, data classification, model building, and model evaluation. AUC: area under the curve; EHR: electronic health record; HIE: health information exchange; PPV: positive predictive value.
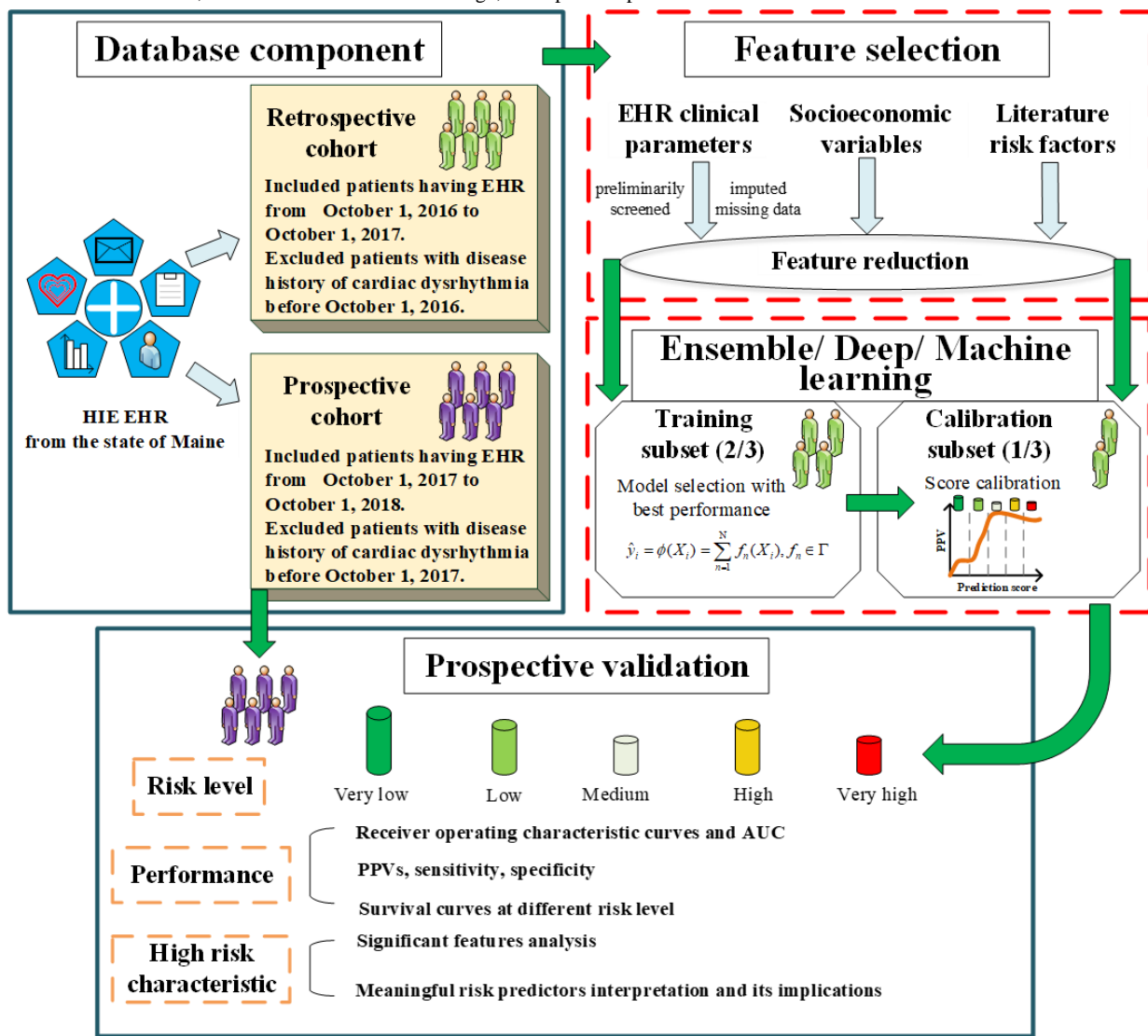
**Figure 2.** The detailed modeling process with the ensemble learning method.
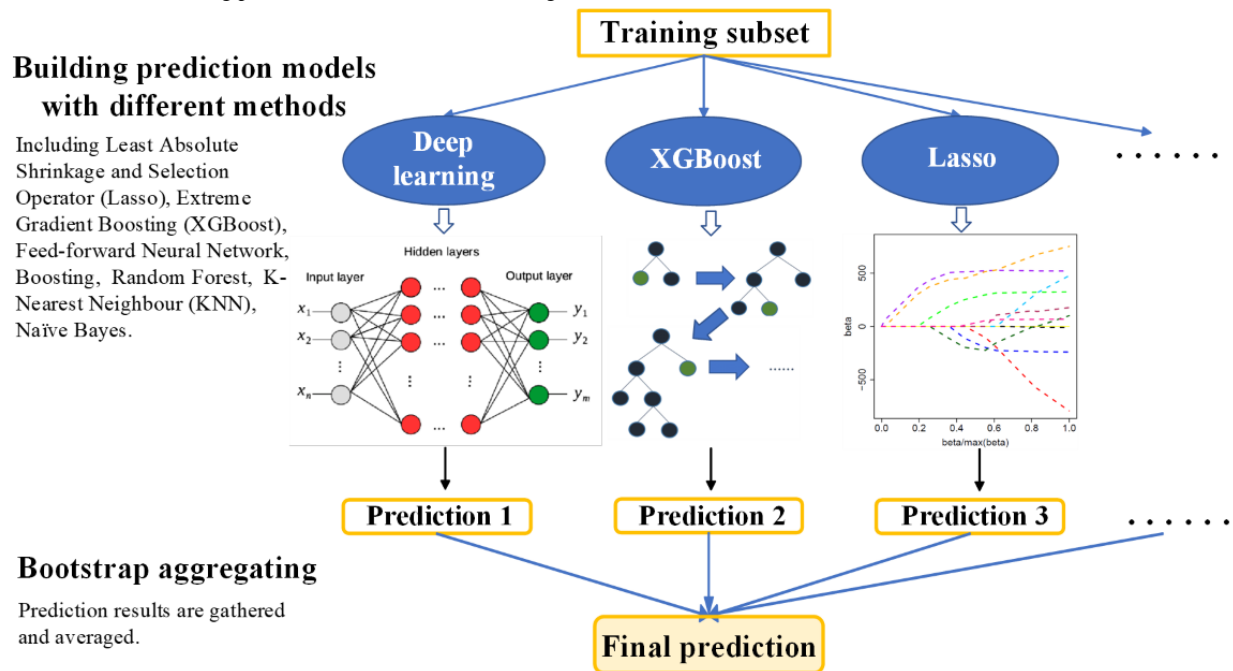


## Data Sources

Nearly 95% of the population in Maine was included in the study. Clinical variables were collected, including demographic information, socioeconomic status, laboratory, and radiographic tests coded according to Logical Observation Identifier Names and Codes, outpatient medication prescriptions coded according to the National Drug Code, and primary and secondary diagnoses and procedures, which were coded using International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM).

## Definition of Cardiac Dysrhythmia

Cardiac dysrhythmia was defined according to ICD-10-CM, including paroxysmal supraventricular tachycardia, paroxysmal ventricular tachycardia, atrial fibrillation, atrial flutter, premature beats, sinoatrial node dysfunction, and other cardiac dysrhythmias (ICD-10-CM diagnosis codes from I47 to I49).

## Study Population

The individuals included in this study were patients who visited any medical institutions in the Maine health information exchange network from October 1, 2015, to October 1, 2018. The retrospective timeframe was from October 1, 2016, to October 1, 2017. The prospective timeframe was from October 1, 2017, to October 1, 2018. Individuals were excluded if they died during the study period or were diagnosed with cardiac dysrhythmia before October 1, 2016, for the retrospective analysis and before October 1, 2017, for the prospective analysis.

## Features

Information regarding the medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and laboratory test results were extracted from EHRs. Relevant socioeconomic variables were extracted from the US Census and US Department of Agriculture websites

[15,16] (see Multimedia Appendix 1). Individuals were aggregated into several age categories. Socioeconomic age-related features such as work or retirement status and conditions about insurance and pension status were structured and standardized as socioeconomic features.

Missing values in the data matrix for machine learning most likely arise due to the lack of the order of the tests or lack of coding for the absence of relevant comorbidity. The data matrix was constructed with all the entries to document the binary outcome (0 or 1) or the counts of the utilization. Therefore, missing entries caused by the data matrix consolidation were filled as 0 outcome or 0 count. Given the confounding effects of stratification factors among various features in the large number of samples, the Cochran-Mantel-Haenszel test was applied to analyze the relationship between the features and corresponding outcome under age-group strata [17]. A total of 658 features were screened out of the original 17,865 features for the subsequent modeling analytics.

Multihypothesis test correction was conducted to ensure the false discovery rate of the remaining features was in an acceptable range [18].

## Correlation Networks

To investigate associations between feature categories, we built correlation networks among the features based on Spearman correlations. In these networks, vertices correspond to features, and an edge existed between 2 vertices if and only if a correlation (absolute value of the Spearman coefficient) >0.1 between 2 features was observed. In real clinical settings, these features are most likely not independent of each other, and more complicated causative or associative relationships may exist among these significant feature categories.

## Model Building

The retrospective cohort was divided into 2 parts: two-thirds of the data in the retrospective cohort were used for training, and

the remaining one-third was used for model calibration. For training, multiple algorithms were applied, including least absolute shrinkage and selection operator (LASSO) [19], feed-forward neural network [20], random forest [21], boosting [22], extreme gradient boosting (XGBoost) [23,24], naïve Bayes [25], and k-nearest neighbor [26]. Bayesian probabilistic ensemble setting was adopted to integrate various model results for better performance [27]. The hypothesis defines a conditional probability distribution of

$$h(x) = P(f(x) = y \mid x, h) \ (\mathbf{1})$$

where $x$ is a given data point. The ensemble method in which the ensemble consists of all of the hypotheses in $H$ each weighted by its posterior probability $P(h \mid x)$, which we used as the positive predictive value (PPV) for each individual model, can be expressed as:



in which each hypothesis is also multiplied by the prior probability of that hypothesis and where $Y$ is the integrated predicted class, $C$ is the set of all possible classes in the training space predicted label, and $c$ indicates a specific class in each classifier $f(x) = y \mid x$.

The isotonic regression method was used to calibrate the model [28], producing a calibrated value $y'$. The $y'$ estimates of the calibration subset were calculated and mapped to PPVs. The PPV for a certain $y'$ was the corresponding proportion of incident arrhythmia events in the cohort having predictive estimates equal to or larger than this $y'$. Therefore, our risk scores, quantifying the probability of an incident arrhythmia event within the subsequent 1 year, can be interpreted by the PPVs.

## Model Evaluations

To find individuals at different risk levels, 5 risk groups were created and assigned to bins: very low-risk, low-risk, medium-risk, high-risk, and very high-risk bins. The model performance was evaluated through sensitivity, specificity, and PPV. The AUROC values were utilized to illustrate the relationship between sensitivity and specificity by composition.

Survival analysis was applied to track the timing of arrhythmia diagnosis in different risk bins. Kaplan-Meier curves were plotted for different risk levels to stratify the time to events of new incidences. The Cox proportional hazards regression method was used for multivariate analysis.

## *Results*

### Baseline Characteristics

The cohort baseline characteristics are shown in Table 1. There was no obvious difference in demographic and clinical patterns between the retrospective and prospective cohorts.

**Table 1.** Baseline characteristics of the retrospective and prospective cohorts.

| Characteristic | Retrospective (n=1,033,856), n (%) | Prospective (n=1,040,767), n (%) |
| --- | --- | --- |
| **Age (years)** | | |
| <35 | 391,613 (37.90) | 399,545 (38.40) |
| 35-50 | 178,348 (17.20) | 176,995 (17.00) |
| 50-65 | 245,580 (23.80) | 243,161 (23.40) |
| 65-75 | 132,444 (12.80) | 135,600 (13.00) |
| >75 | 85,871 (8.30) | 85,466 (8.20) |
| **Gender** | | |
| Male | 464,796 (45.00) | 571,821 (54.90) |
| Female | 569,060 (55.00) | 468,946 (45.10) |
| **Chronic disease** | | |
| Cardiovascular disease[a] | 215,059 (20.80) | 228,757 (22.00) |
| Chronic obstructive pulmonary disease | 38,029 (3.70) | 42,778 (4.10) |
| Chronic kidney disease | 21,277 (2.70) | 23,653 (2.30) |
| Type 2 diabetes | 83,387 (8.10) | 84,649 (8.10) |
| Disorder of metabolism | 214,606 (20.80) | 221,168 (21.30) |
| Hypothyroidism | 73,319 (7.10) | 74,495 (7.20) |
| **Acute disease** | | |
| Pain in throat and chest | 48,507 (4.69) | 55,368 (5.30) |
| Anemia | 10,921 (1.06) | 11,099 (1.10) |
| Edema | 17,096 (1.65) | 17,768 (1.70) |
| Syncope and collapse | 10,662 (1.03) | 11,855 (1.10) |
| Malaise and fatigue | 47,793 (4.62) | 46,834 (4.50) |
| **Health status** | | |
| Long term (current) drug therapy | 86,501 (8.40) | 107,834 (10.40) |
| Personal history of other diseases and conditions | 104,282 (10.10) | 130,955 (12.60) |
| BMI>33.0 | 3055 (0.30) | 3842 (0.40) |
| **Laboratory test** | | |
| Glomerular filtration rate | 391,613 (37.90) | 16,726 (1.60) |
| Coagulation assay | 12,197 (1.20) | 9616 (0.90) |
| Carboxyhemoglobin in blood | 35,773 (3.50) | 20,850 (2.00) |
| **Medication** | | |
| Beta-adrenergic blocker | 66,750 (6.50) | 69,572 (6.70) |
| Proton pump inhibitor | 68,123 (6.60) | 69,278 (6.70) |
| Vitamin K antagonist | 6858 (0.70) | 5541 (0.50) |

[a]Cardiovascular diseases included heart failure, rheumatic mitral valve diseases, atrioventricular and left bundle-branch block, cardiomyopathy, nonrheumatic aortic, tricuspid and mitral valve disorders, atherosclerosis, and other disorders of arteries and arterioles.

## Feature Community Structure and Correlation Networks

The original features (n=17,865) were extracted from the EHRs and socioeconomic databases. The model building process identified 307 features with contributing weights, including 2 demographic features, 18 socioeconomic characteristics, 101 chronic disease diagnostics, 147 confirmed acute disease and disease events, 7 procedures, 5 utilization variables, 9 factors of the health status, 9 medication prescriptions, and 9 laboratory tests. The top 60 important features and their odds ratios in the model are tabulated in Multimedia Appendix 2.

We built correlation networks among these 307 features based on Spearman correlations. The integral correlation networks contain 307 vertices and 325 edges. The majority of edges

involved diagnostic diseases (n=206) and demographic features (n=34), with an additional 28 edges involving clinical medications, 27 involving laboratory tests, 18 involving health status, 15 involving procedures, and 10 involving socioeconomic characteristics, as well as all the utilization variables. The community structure of the 153 impactful features in different types and their correlation networks containing 127 edges, as an example, is shown in Figure 3. The important network structure of the predictive diagnostic features is shown in Figure 4.

**Figure 3.** The community structure of the 153 impactful features and their correlation networks (absolute value of the correlation coefficient >0.1).



Arrhythmia is an important group of cardiovascular diseases and is associated with other cardiovascular diseases. Heart disease–related features were revealed in our community structures and correlational networks, including acute myocardial infarction, aortic dissection, chronic ischemic heart disease, cardiomyopathy, and atherosclerosis caused by chronic obstructive pulmonary disease. These features imply a potential causative relationship with arrhythmia. Electrolyte imbalance, heart enlargement, heart failure, and myocardial ischemia may be related to the pathogenesis of arrhythmia and may be complications of chronic kidney disease, metabolic syndrome

(type 2 diabetes), and hypertension. The associative relationship between arrhythmia and these chronic diseases is shown in Figure 4. Chronic kidney disease patients in later stages can have comorbidity with oliguria, anuria, or uremic cardiomyopathy, leading to uremic toxin accumulation in the body, and imbalance of homeostasis, leading to arrhythmia attack. Diabetes can cause microvascular and macrovascular complications with different pathological mechanisms, such as diabetic cardiomyopathy and diabetic kidney disease. Hypertension and chronic kidney disease are closely interlinked pathophysiologic states. Anemia is common in patients with

heart disease. It was present in approximately one-third of patients with congestive heart failure and 10% to 20% of patients with coronary heart disease. These diseases can lead to cardiac dysfunction, inducing malignant arrhythmia.

Figure 4. The important network structure of the predictive diagnostic features.



## Model Performance

### AUROC

The AUROC results from our predictive methods, including LASSO, feed-forward neural network, random forest, boosting, XGBoost, naïve Bayes, k-nearest neighbor, and ensemble learning, were compared to demonstrate the effectiveness of the cardiac dysrhythmia risk prediction (Table 2 and Multimedia Appendix 3). An ensemble learning method was applied to harmonize and vote (the results of multiple algorithms) for the best modeling method in this study (Multimedia Appendix 4; AUROC 0.827). In addition, the prediction AUROC (0.819) calculated by LASSO was comparable to that calculated by the ensemble learning method. The following analysis was based on LASSO model predictions.

Table 2. Model performance.

| Model | AUROC[a] (95% CI) |
| --- | --- |
| Ensemble learning | 0.827 (0.824-0.830) |
| Least absolute shrinkage and selection operator | 0.819 (0.816-0.822) |
| Extreme gradient boosting | 0.808 (0.805-0.811) |
| Feed-forward neural network | 0.807 (0.804-0.810) |
| Boosting | 0.775 (0.771-0.778) |
| Random forest | 0.695 (0.691-0.699) |
| k-Nearest neighbor | 0.631 (0.627-0.635) |
| Naïve Bayes | 0.611 (0.607-0.614) |

[a]AUROC: area under the receiver operating characteristic curve.

### The Risk Score Metric

Patients in the prospective cohort were divided into 5 risk categories (very low, low, medium, high, and very high; Figure 5) based on the predictive scores. Over 90% of the prospective patients (1,021,827/1,040,767) were categorized into the very low-risk or low-risk categories, while 0.063% (650/1,040,767) were classified as high-risk or very high-risk; 51.85% of patients (14/27) in the very high-risk group had confirmative diagnosis of arrhythmia in the subsequent year (Multimedia Appendix 5).

Survival curve analysis was applied to quantify the effectiveness of the 5-risk bin stratification for future cardiac dysrhythmia events within 1 year (Figure 6). Survival curves with different risk levels were distinguished with hazard ratios varying from 8.04 to 202.13 at different risk bins. Sensitivities, specificities, and PPVs in the 5 risk groups were documented (Multimedia Appendix 5); 0.34% (1,873/555,233) and 51.85% (14/27) of the patients were stratified as very low-risk and very high-risk, respectively, for arrhythmia in the subsequent 1 year.

**Figure 5.** The proportion of patients and their positive predictive values with different prediction scores on the prospective cohort.



**Figure 6.** Survival curves of the 5 risk groups. HR: hazard ratio.

XSL•FO
RenderX

## Characterization of the Predictive Clinical Parameters

Age distributions at 5 different risk levels are shown in Multimedia Appendix 6. Young individuals (<35 years) were enriched in the low-risk or very low-risk bins, while older individuals were enriched in the high-risk or very high-risk bins; 68.46% of patients (380,112/555,233) in the very low-risk group were young individuals. The majority of the patients (397,164/466,594, 85.12%) in the low-risk group were older than 50 years. In the medium-, high-, and very high-risk groups, patients aged 65 years or older were the majority (medium: 16,194/18,290, 88.54%; high: 576/623, 92.46%; very high: 27/27, 100%; see Multimedia Appendix 7). There were more male patients than female patients in the medium-risk (10,308/18,290, 56.36%), high-risk (381/623, 61.16%), and very high-risk groups (14/27, 51.85%).

Many patients in the high-risk and the very high-risk groups had comorbidity with chronic (cardiovascular disease: 436/650, 67.07%; metabolism disorder: 304/650, 46.77%; type 2 diabetes: 183/650, 28.15%; chronic obstructive pulmonary disease: 159/650, 24.46%; chronic kidney disease: 146/650, 22.46%) or acute diseases (cardiovascular disease: 220/650, 33.85%; syncope and collapse: 28/650, 4.31%; dizziness and giddiness: 47/650, 7.23%; pain in throat and chest: 81/650, 12.46%; breathing abnormalities: 151/650, 23.23%; Multimedia Appendix 7). Time-to-event analysis was applied to these disease groups (Multimedia Appendix 8). There were apparent differences in time-to-event patterns between patients with low risk and those with high risk. Furthermore, in the high-risk and very high-risk bins, different disease groups demonstrated different patterns. Patients with construction disorders had less probability of experiencing arrhythmia than those with other chronic diseases. Patients with dizziness and giddiness had higher chance of experiencing arrhythmia in a certain time period than those with other acute diseases.

Multimedia Appendix 9 shows the 9 health status features used in the model: (1) long-term (current) drug therapy, (2) personal history of certain other diseases, (3) personal history of other diseases and conditions, (4) BMI >33 kg/m$^2$ (overweight), (5) encounter for supervision of normal pregnancy, (6) presence of cardiac and vascular implant and graft, (7) acquired absence of organs, (8) encounter for follow-up examination after completed treatment for conditions other than malignant neoplasm, and (9) presence of other functional implants; 57.85% of patients (376/650) in the high-risk and very high-risk bins had at least 1 abnormal health status feature, in contrast to 23.35% (238,612/1,021,827) in the low-risk and very low-risk bins.

Multimedia Appendix 9 summarizes the 9 laboratory test predictors used in the model: (1) coagulation assay, (2) glomerular filtration rate, (3) carboxyhemoglobin in blood, (4) cardiac troponin T antibodies in blood, (5) blood glucose, (6) creatine kinase, (7) reticulocytes in blood, (8) n-terminal prohormone B-type natriuretic peptide in serum or plasma, and (9) estimated average glucose level. Having abnormal results of the coagulation assay had the highest weight in the model, and 34.62% of patients (225/650) in the high-risk and very high-risk bins had at least 1 abnormal result among the 9 laboratory tests, in contrast to only 5.81% (59,367/1,021,827) in the low-risk and very low-risk bins.

Multimedia Appendix 9 also shows the 9 prescriptions used as predictors in the model: (1) beta-adrenergic blocker, (2) 3-hydroxy-3-methylglutaryl coenzyme A reductase inhibitor, (3) loop diuretic, (4) calcium channel blocker, (5) proton pump inhibitor, (6) vitamin K antagonist, (7) dihydropyridine calcium channel blocker, (8) angiotensin-converting enzyme inhibitor, and (9) factor Xa inhibitor. All drugs, except proton pump inhibitors, were mainly used to treat cardiovascular diseases such as hypertension, congestive heart failure, coronary artery disease, hypertrophic cardiomyopathy, deep vein thrombosis, and acute pulmonary embolism; 90% of individuals (585/650) in the high-risk and very high-risk bins had medication histories with at least 1 drug, in contrast to 16.67% (170,317/1,021,827) of individuals in the low-risk and very low-risk bins.

Socioeconomic features reflected the social disparities of individuals' living environments and living conditions. The Spearman rank method was used to study the correlation between socioeconomic factors and arrhythmia (Multimedia Appendix 10). Our analysis revealed a high enrichment of higher education ($\rho=-0.0036$), high-income families ($\rho=-0.0218$), and people with employer-based insurance ($\rho=-0.0114$) in the low-risk bins. Individuals with high income, high education, and private insurance accounted for 26.62% (173/650), 14.00% (91/650), and 41.38% (269/650), respectively, in the high-risk and very high-risk categories.

## *Discussion*

### Summary of Main Findings

In this study, we developed a case-finding tool to identify general population individuals at risk of future cardiac dysrhythmia events within 1 year using Maine statewide health information exchange aggregated EHR data sets. The predictive model was trained retrospectively (AUROC 0.854) and validated prospectively (AUROC 0.827). Our model was capable of prospectively stratifying the general population into 5 risk bins (very low, low, medium, high, and very high) of incident arrhythmia. It will support targeted care plans to manage patients in different risk categories.

### Comparison With Prior Work

Our case-finding tool is different from previous efforts in terms of the targeted population and predictive timeframe. Other models [29,30] applied the logistic regression method, focusing on individuals with symptoms of syncope; the AUROC values of these models ranged from 0.44-0.81. Given that our method is be applicable to the general population, with a 1-year prospective timeframe, our case-finding tool has additional translational advantages.

### Model Risk Predictors and Their Implications for Preventive Care and Early Intervention

The high-risk or very high-risk bin individuals in a prospective cohort are likely to have higher disease burdens, given the confirmative diagnosis of multiple chronic diseases or acute disease events as well as other major medical histories (Figure

7 and Multimedia Appendix 11). Cardiovascular disease, one of the top predictors of our model, was found to be associated with heart failure, cardiomyopathy, and some valve problems [7]. Another important predictor of the model—chronic kidney disease—was also related to a few acute or chronic diseases that caused severe outcomes. A bidirectional causative relationship may exist between atrial fibrillation and chronic kidney disease [31], and coexistence of atrial fibrillation and chronic kidney disease greatly increased morbidity and mortality. Patients with chronic kidney disease may have a

higher risk of death when implantable cardioverter defibrillators are used to treat ventricular arrhythmias [32]. Studies have shown that conditions with asymptomatic and persistent hypoglycemia increased the risk of arrhythmia [33]. Therefore, patients with chronic diabetes (including Type 1 and 2) often have varying degrees of arrhythmia risk. The direct effect of low glucose levels, hypokalemia, and catecholamine release can prolong cardiac repolarization, increasing the risk of early afterdepolarization and ventricular arrhythmias [33].

**Figure 7.** Patients' average clinical costs in the past 12 months with respect to the average number of chronic diseases; 17 common diseases are presented in the low- and very low–risk group (green) and the high- and very high–risk group (red). CKD: chronic kidney disease; COPD: chronic obstructive pulmonary disease; CVD: cardiovascular disease.
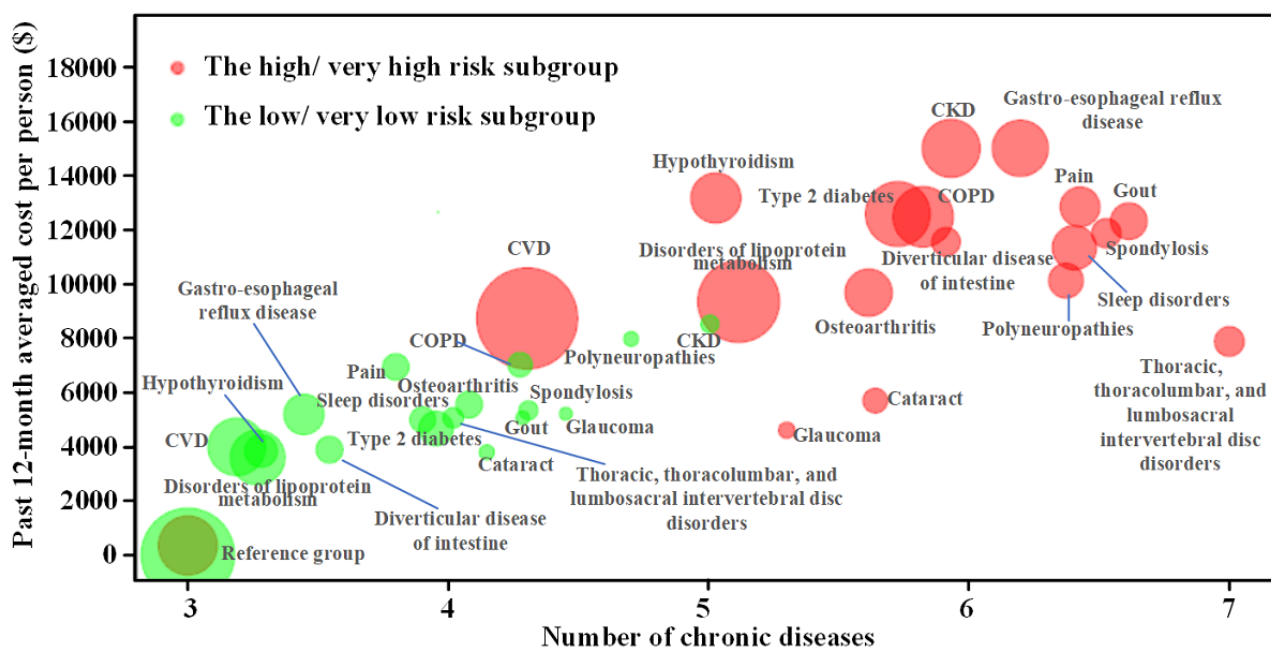


Studies have shown that certain social factors were indirect causes of arrhythmia. The predisposing factors of arrhythmia may involve various aspects of the psychosocial environment related to social status [34]. Social and psychosocial factors may influence the risk of arrhythmia through behavioral risk factors for smoking, exercise, and diet [35]. According to a survey study [36], the level of education and social status (occupational and occupational income) were inversely related to the incidence of cardiovascular disease. These observations are in line with the findings for socioeconomic factors included in our model: education, income, insurance type, regional features (towns, villages), and ethnic groups. We found that individuals with low incomes, low education levels, and nonprivate insurance groups had a higher probability of having arrhythmias (Multimedia Appendix 10).

### Practical Application of the Risk Model

Our study established a model for predicting the probability of arrhythmia disease within 1 subsequent year. By continuously tracking the influencing factors, the accuracy and applicability of the prediction results could be further improved. The predictive weight of different kinds of factors can provide insight into the formation mechanism of arrhythmia, the analysis of predisposing factors, and the research of preventive measures.

Our model will benefit physicians and health care organizations as well as patients. Model prediction and risk score results can be used as an auxiliary tool for physicians to diagnose and provide a reference for treatment planning. The stratification of the risk of the patient population also contributes to medical budget planning and target intervention. In the very high-risk category, 29.63% patients (8/27) would have been diagnosed with arrhythmia within the first 4 months of the subsequent year, which would then have gradually increased to 51.85% (14/27) within 1 year. Therefore, for those in the very high-risk group, it is necessary to formulate appropriate personalized intervention programs according to medical history, health status, living environment, and other conditions to prevent or delay the development of arrhythmia.

Given that our study showed that individuals at high risk of developing arrhythmias often have multiple chronic conditions, such as cardiovascular disease, metabolism disorders, type 2 diabetes, chronic obstructive pulmonary disease, and chronic kidney disease, aggressive interventions, including early routine testing and treatment of related chronic diseases, are needed for those patients. Arrhythmias after surgery are common and can lead to serious complications [37]. Therefore, for patients undergoing surgeries, especially chest and heart surgeries, it is necessary to take some preventive measures and conduct

XSL•FO
**RenderX**

continuous electrocardiogram monitoring. This measure helps identify this high-risk population and avoids an increased risk of cardiovascular events and death [37]. Our model found that people with employer insurance, higher education, or higher incomes have a lower probability of experiencing arrhythmia. We speculate that these people tend to pay more attention to their physical condition and exercise more often to maintain health. Targeted early intervention reduces the number of arrhythmia patients and arrhythmia conditions, which is a rational and effective allocation of health care resources.

## Limitations

Our research has some limitations that could be further improved. First, some information was missing from our data set. Lifestyle information (such as eating habits and amount of daily exercise) was not fully documented in the EHR data. Second, when there are too many missing variables, there may be some bias in data preprocessing with the k-nearest neighbor method, resulting in inaccurate estimation results. Third, arrhythmia is a very common symptom that can be triggered in many cases, and some occurrences are not dangerous. Although arrhythmia was defined in detail in our model, it was only

stratified by the probability of its occurrence, not by its severity. If the patient population based on the severity of the arrhythmia can be further subdivided, more accurate reference information for arrhythmia diagnosis and intervention will be provided.

## Conclusions

A risk prediction model of 1-year incidence of cardiac dysrhythmia was developed and prospectively validated using EHR data from 1.5 million people in Maine. The model was able to classify patients according to the predicted scores. The model had a good predictive performance (AUROC 0.827) in a prospective test. Age, gender, cardiovascular disease, chronic kidney disease, chest pain, pleural effusion, and socioeconomic factors were found to be related to new arrhythmia. For patients at high risk, early intervention should be carried out in a timely manner. Patients with low and moderate risk should maintain good living and eating habits, pay more attention to their physical condition, and exercise more often to avoid the occurrence of serious arrhythmias. This prediction model and analysis will ultimately benefit patient families, clinicians, and social health care institutions.

## Conflicts of Interest

None declared.

Multimedia Appendix 1
List of social determinant variables detailed in the data source and mapping method.
[DOCX File , 19 KB - medinform_v9i2e23606_app1.docx ]

Multimedia Appendix 2
List of the top 60 important features and their odds ratios in the model.
[DOCX File , 31 KB - medinform_v9i2e23606_app2.docx ]

Multimedia Appendix 3
Receiver operating characteristic curve comparisons of the model performance.
[DOCX File , 64 KB - medinform_v9i2e23606_app3.docx ]

Multimedia Appendix 4
Receiver operating characteristic curves derived from the prospective cohorts.
[DOCX File , 42 KB - medinform_v9i2e23606_app4.docx ]

Multimedia Appendix 5
Performance of the 1-year arrhythmia risk prediction model in the prospective cohort.
[DOCX File , 18 KB - medinform_v9i2e23606_app5.docx ]

Multimedia Appendix 6
Distribution of age and gender in the 5 risk categories.
[DOCX File , 28 KB - medinform_v9i2e23606_app6.docx ]

Multimedia Appendix 7

Distribution of top risk predictors across the 5 risk categories.
[DOCX File , 26 KB - medinform_v9i2e23606_app7.docx ]

Multimedia Appendix 8
Time-to-arrhythmia diagnosis curves of the chronic disease group in the low-risk/very low-risk and high-risk/very high-risk populations of the prospective cohort.
[DOCX File , 204 KB - medinform_v9i2e23606_app8.docx ]

Multimedia Appendix 9
(a) The feature importance score bar chart of health status, laboratory test, and medication. (b) Distributions of health status, laboratory test, and medication in the low-risk/very low-risk and high-risk/very high-risk categories.
[DOCX File , 98 KB - medinform_v9i2e23606_app9.docx ]

Multimedia Appendix 10
Spearman rank correlation between socioeconomic features and prospective arrhythmia risk score.
[DOCX File , 32 KB - medinform_v9i2e23606_app10.docx ]

Multimedia Appendix 11
Patients' average clinical costs and average number of chronic diseases in the past 12 months in the low-risk/very low-risk and the high-risk/very high-risk subgroups. The points presented 17 common diseases.
[DOCX File , 164 KB - medinform_v9i2e23606_app11.docx ]

## References

1. Heart disease and abnormal heart rhythm (arrhythmia). MedicineNet. URL: www.medicinenet.com/arrhythmia_irregular_heartbeat/article.htm [accessed 2021-01-29]

2. Mehra R. Global public health problem of sudden cardiac death. J Electrocardiol 2007 Nov;40(6 Suppl):S118-S122. [doi: 10.1016/j.jelectrocard.2007.06.023] [Medline: 17993308]

3. Winkel TA, Voute MT, de Melis M, Hoeks SE, Schouten O, Kessels R, et al. Sudden death during follow-up after new-onset ventricular tachycardia in vascular surgery patients. J Vasc Surg 2011 Mar;53(3):732-7; discussion 737 [FREE Full text] [doi: 10.1016/j.jvs.2010.09.013] [Medline: 21106322]

4. Peacock J, Whang W. Psychological distress and arrhythmia: risk prediction and potential modifiers. Prog Cardiovasc Dis 2013 May;55(6):582-589 [FREE Full text] [doi: 10.1016/j.pcad.2013.03.001] [Medline: 23621968]

5. Li DL, Quispe R, Madan N, Zhang L, Taub CC. A risk score for predicting atrial fibrillation in individuals with preclinical diastolic dysfunction: a retrospective study in a single large urban center in the United States. BMC Cardiovasc Disord 2019 Feb 27;19(1):47-47 [FREE Full text] [doi: 10.1186/s12872-019-1024-4] [Medline: 30813895]

6. van Lammeren GW, Catanzariti LM, Peelen LM, de Vries JP, de Kleijn DP, Moll FL, et al. Clinical prediction rule to estimate the absolute 3-year risk of major cardiovascular events after carotid endarterectomy. Stroke 2012 May;43(5):1273-1278. [doi: 10.1161/strokeaha.111.647958]

7. Bax JJ, Kraft O, Buxton AE, Fjeld JG, Parˇiˊzek P, Agostini D, et al. 123I-mIBG scintigraphy to predict inducibility of ventricular arrhythmias on cardiac electrophysiology testing. Circ Cardiovasc Imaging 2008 Sep;1(2):131-140. [doi: 10.1161/circimaging.108.782433]

8. de Haan S, Meijers TA, Knaapen P, Beek AM, van Rossum AC, Allaart CP. Scar size and characteristics assessed by CMR predict ventricular arrhythmias in ischaemic cardiomyopathy: comparison of previously validated models. Heart 2011 Dec 13;97(23):1951-1956. [doi: 10.1136/heartjnl-2011-300060] [Medline: 21917670]

9. Kosich F, Schumacher K, Potpara T, Lip GY, Hindricks G, Kornej J. Clinical scores used for the prediction of negative events in patients undergoing catheter ablation for atrial fibrillation. Clin Cardiol 2019 Mar 14;42(2):320-329 [FREE Full text] [doi: 10.1002/clc.23139] [Medline: 30578568]

10. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. Annu Rev Public Health 2016;37:61-81 [FREE Full text] [doi: 10.1146/annurev-publhealth-032315-021353] [Medline: 26667605]

11. Goldstein B, Navar A, Pencina M, Ioannidis J. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017 Jan;24(1):198-208 [FREE Full text] [doi: 10.1093/jamia/ocw042] [Medline: 27189013]

12. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One 2017 Apr 4;12(4):e0174944 [FREE Full text] [doi: 10.1371/journal.pone.0174944] [Medline: 28376093]

13. Ye C, Wang O, Liu M, Zheng L, Xia M, Hao S, et al. A real-time early warning system for monitoring inpatient mortality risk: prospective study using electronic medical record data. J Med Internet Res 2019 Jul 05;21(7):e13719 [FREE Full text] [doi: 10.2196/13719] [Medline: 31278734]

14. Tani LY. Echocardiographic screening for rheumatic heart disease. Circulation 2014 May 13;129(19):1912-1913. [doi: 10.1161/circulationaha.114.009406]

15. Census. United States Census Bureau. 2017. URL: https://www.census.gov/ [accessed 2017-10-26] [WebCite Cache ID 6uTzRZU0g]

16. Food environment atlas. United States Department of Agriculture. URL: https://www.ers.usda.gov/data-products/food-environment-atlas/ [accessed 2017-10-24]

17. Horwitz R. Statistical aspects of the analysis of data from retrospective studies of disease. J Chronic Dis 1979 Jan;32(1-2):ii. [doi: 10.1016/0021-9681(79)90031-6]

18. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Statist 2001 Aug;29(4):1165-1188. [doi: 10.1214/aos/1013699998]

19. Park T, Casella G. The Bayesian Lasso. J Am Stat Assoc 2012 Jan 01;103(482):681-686. [doi: 10.1198/016214508000000337]

20. Ilonen J, Kamarainen JK, Lampinen J. Differential evolution training algorithm for feed-forward neural networks. Neural Process Lett 2003;17(1):1370-4621. [doi: 10.1023/A:1022995128597]

21. RandomForest: Breiman and Cutler's random forests for classification and regression. CRAN R Project. URL: https://cran.r-project.org/web/packages/randomForest/index.html [accessed 2019-05-01] [WebCite Cache ID 781QCayMU]

22. Generalized boosted regression models. CRAN R Project. URL: https://cran.r-project.org/web/packages/gbm [accessed 2019-05-01] [WebCite Cache ID 781QSUtmm]

23. Batista G, Monard M. K-nearest neighbour as an imputation method: experimental results. ICMC-USP Technical Reports 2002;87:251-260.

24. Tianqi C, Carlos G. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016. 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17; San Francisco, California p. 785-794.

25. El KM, Bensaid A, Rachidi T. Automatic Arabic document categorization based on the Naïve Bayes algorithm. In: Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages. 2004 Presented at: Workshop on Computational Approaches to Arabic Script-based Languages; August 28; Geneva, Switzerland. [doi: 10.3115/1621804.1621819]

26. Yong Z, Youwen L, Shixiong X. An improved KNN text classification algorithm based on clustering. JCP 2009 Mar 01;4(3):230-237 [FREE Full text]

27. Fersini E, Messina E, Pozzi F. Sentiment analysis: Bayesian ensemble learning. Decis Support Syst 2014 Dec;68:26-38. [doi: 10.1016/j.dss.2014.10.004]

28. Niculescumizil A, Caruana R. Predicting good probabilities with supervised learning. 2005 Presented at: 22nd International Conference on Machine Learning; August 7-11; Bonn, Germany. [doi: 10.1145/1102351.1102430]

29. Sarasin FP. A risk score to predict arrhythmias in patients with unexplained syncope. Acad Emerg Med 2003 Dec 01;10(12):1312-1317. [doi: 10.1197/s1069-6563(03)00535-9]

30. Thiruganasambandamoorthy V, Stiell IG, Sivilotti MLA, Rowe BH, Mukarram M, Arcot K, et al. Predicting short-term risk of arrhythmia among patients with syncope: the Canadian Syncope Arrhythmia Risk score. Acad Emerg Med 2017 Nov 12;24(11):1315-1326 [FREE Full text] [doi: 10.1111/acem.13275] [Medline: 28791782]

31. Lau YC, Proietti M, Guiducci E, Blann AD, Lip GY. Atrial fibrillation and thromboembolism in patients with chronic kidney disease. J Am Coll Cardiol 2016 Sep 27;68(13):1452-1464. [doi: 10.1016/j.jacc.2016.06.057] [Medline: 27659468]

32. Cuculich PS, Sánchez JM, Kerzner R, Greenberg SL, Sengupta J, Chen J, et al. Poor prognosis for patients with chronic kidney disease despite ICD therapy for the primary prevention of sudden death. Pacing Clin Electrophysiol 2007 Mar;30(2):207-213. [doi: 10.1111/j.1540-8159.2007.00651.x] [Medline: 17338717]

33. Chow E, Bernjak A, Williams S, Fawdry RA, Hibbert S, Freeman J, et al. Risk of cardiac arrhythmias during hypoglycemia in patients with type 2 diabetes and cardiovascular risk. Diabetes 2014 May 22;63(5):1738-1747 [FREE Full text] [doi: 10.2337/db13-0468] [Medline: 24757202]

34. Marmot M, Bosma H, Hemingway H, Brunner E, Stansfeld S. Contribution of job control and other risk factors to social variations in coronary heart disease incidence. The Lancet 1997 Jul;350(9073):235-239. [doi: 10.1016/s0140-6736(97)04244-x]

35. Hemingway H, Malik M, Marmot M. Social and psychosocial influences on sudden cardiac death, ventricular arrhythmia and cardiac autonomic function. Eur Heart J 2001 Jul 01;22(13):1082-1101. [doi: 10.1053/euhj.2000.2534] [Medline: 11428849]

36. Albert MA, Glynn RJ, Buring J, Ridker PM. Impact of traditional and novel risk factors on the relationship between socioeconomic status and incident cardiovascular events. Circulation 2006 Dec 12;114(24):2619-2626. [doi: 10.1161/circulationaha.106.660043]

XSL•FO
RenderX

37.  Winkel TA, Schouten O, Hoeks SE, Flu W, Hampton D, Kirchhof P, et al. Risk factors and outcome of new-onset cardiac arrhythmias in vascular surgery patients. Am Heart J 2010 Jun;159(6):1108-1115. [doi: 10.1016/j.ahj.2010.03.035] [Medline: 20569727]

## Abbreviations

**AUROC:** area under the receiver operating characteristic curve
**EHR:** electronic health record
**ICD-10-CM:** International Classification of Diseases, Tenth Revision, Clinical Modification
**LASSO:** least absolute shrinkage and selection operator
**PPV:** positive predictive value
**XGBoost:** extreme gradient boosting

XSL•FO
**RenderX**

Review

# Culturally Competent Gender, Sex, and Sexual Orientation Information Practices and Electronic Health Records: Rapid Review

Kelly Davison[1,2], MN, MSc; Roz Queen[1]; Francis Lau[3], PhD; Marcy Antonio[1], BSc, MPH

[1]University of Victoria, Victoria, BC, Canada

[2]Canada Health Infoway, Toronto, ON, Canada

[3]School of Health Information Science, University of Victoria, Victoria, BC, Canada

**Corresponding Author:**
Kelly Davison, MN, MSc
University of Victoria
PO Box 1700 STN CSC
Victoria, BC, V8W 2Y2
Canada
Phone: 1 250 721 7211
Email: kdaviso@uvic.ca

## *Abstract*

**Background:**   Outdated gender, sex, and sexual orientation (GSSO) information practices in health care contribute to health inequities for sexual and gender minorities (SGMs). Governments, statistics agencies, and health care organizations are developing and implementing modernized practices that support health equity for SGMs. Extending our work, we conducted a rapid review of grey literature to explore information practices that support quality health care for SGMs.

**Objective:**   The aim of this rapid review of grey literature was to elucidate modern GSSO information practices from leading agencies for adaptation, adoption, and application by health care providers and organizations seeking to modernize outdated GSSO information practices that contribute to health inequities among SGMs.

**Methods:**   We searched MEDLINE and Google from 2015 to 2020 with terms related to gender, sex, sexual orientation, and electronic health/medical records for English-language grey literature resources including government and nongovernment organization publications, whitepapers, data standards, toolkits, health care organization and health quality practice and policy guides, conference proceedings, unpublished academic work, and statistical papers. Peer-reviewed journal articles were excluded, as were resources irrelevant to information practices. We also screened the reference sections of included articles for additional resources, and canvassed a working group of international topic experts for additional relevant resources. Duplicates were eliminated. ATLAS.ti was used to support analysis. Themes and codes were developed through an iterative process of writing and discussion with the research team.

**Results:**   Twenty-six grey literature resources met the inclusion criteria. The overarching themes that emerged from the literature were the interrelated behaviors, attitudes, and policies that constitute SGM cultural competence as follows: shared language with unambiguous definitions of GSSO concepts; welcoming and inclusive care environments and affirming practices to reduce barriers to access; health care policy that supports competent health care; and adoption of modernized GSSO information practices and electronic health record design requirements that address invisibility in health data.

**Conclusions:**   Health equity for SGMs requires systemic change. Binary representation of sex and gender in electronic health records (EHRs) obfuscates natural and cultural diversity and, in the context of health care, places SGM patients at risk of clinical harm because it leads to clinical assumptions. Agencies and agents in health care need to be equipped with the knowledge and tools needed to cultivate modern attitudes, policies, and practices that enable health equity for SGMs. Adopting small but important changes in the language and terminology used in technical and social health care systems is essential for institutionalizing SGM competency. Modern GSSO information practices depend on and reinforce SGM competency in health care.

XSL•FO
RenderX

## *Introduction*

### Background

Sexual and gender minorities (SGMs) face health inequities related to social and institutionalized prejudice [1-5]. Discrimination of SGMs should not happen in Canadian health care, where the ability of individuals to freely express their gender identity is a human right [4-6]. Yet, in its current state, each encounter with the health care system is fraught with the risks of emotional, psychological, and even physical harm to SGMs [1,3-5,7]. Recognition of the impact of outdated gender, sex, and sexual orientation (GSSO) information practices is growing [1-3,8,9], but information practices have changed little in recent decades [2], and are built on outdated, poorly defined, and highly constrained conceptualizations of sex, gender, and sexual orientation [1,2]. Current GSSO information practices are largely outdated and are not reflective of the progress that Canadian society has made in terms of being inclusive of SGMs and defending the human rights of SGMs [4,6].

The aim of this rapid review of grey literature was to elucidate GSSO information practices for SGMs from leading agencies for adaptation, adoption, and application to health care providers and organizations seeking to modernize outdated GSSO information practices that contribute to SGM health inequities. This research extends recent work done by our research team that explored peer-reviewed academic literature [1,2]. What is clear from our previous research is that throughout the Canadian eHealth landscape, GSSO data elements lack useful definitions; fundamentally conflate administrative, clinical, biological, and social concepts; and are structured around cisheteronormative binary constructs of sex and/or gender terms and codes [2-5] that are neither affirmative nor inclusive of SGMs and do not include Indigenous GSSO concepts [2].

### Definitions

GSSO information practices include the definition, collection, use, and sharing of GSSO information. The term *definition* refers to the meaning of the concepts used by humans to communicate and the corresponding concepts and codes used to represent this meaning in electronic health records (EHRs). *Collection* and *use* include clinical documentation, coding, and administrative input of health information into EHR systems to support direct clinical care and communication (primary uses) and organizational needs such as analytics and research (secondary uses). *Sharing* includes standardized protocols for the exchange of health information between health care providers and information systems for primary and secondary uses along a connected care continuum such as messaging standards. The term *providers* refers to regulated and unregulated health care staff and professionals (people employed to provide care). Unless specifically mentioned, we use the term EHRs in its broadest sense—health information systems that contain digital collections of a person's health history and care records.

## *Methods*

### Review Questions

Our central question for this rapid review of grey literature was "What are modern and emerging GSSO information practices that reduce barriers to access and to delivery of safe and quality health care for SGMs?" The objectives of this review include the discovery, review, synthesis, and description of SGM equity-oriented GSSO information practices in EHRs from leading experts and organizations.

### Study Selection and Synthesis

The rapid review method was chosen because rapid reviews have gained recognition in health care research as a suitable method for fast exploration of rapidly evolving topics [10] and because this work is a natural extension of and complement to our research team's recent rapid exploration of peer-reviewed academic literature [1] and EHR systems in Canada [2]. This work expands the domain of inquiry to the body of grey literature that houses key institutional, professional, and technical knowledge of safe, quality, and inclusive health care for SGMs.

The search strategy (Multimedia Appendix 1) for this review was developed with the input of a subject matter university librarian and specifies the search terms, including Medical Subject Heading terms, repositories, inclusion and exclusion criteria, and the codebook (Multimedia Appendix 2) developed to guide our analysis. The initial phase of this work involved a search using MEDLINE for English language government reports and documents, white papers, theses, care guidelines, and literature published in a 6-year period between January 2015 and July 2020. The second phase involved a Google search. Search terms for the first two phases included "sex," "gender," "sexual orientation," and "electronic health/medical records." MEDLINE search results were screened by title, abstract, or description and full-text review for relevance to the topic; peer-reviewed journal papers were excluded. Results from the Google search were screened by hand by reviewing the name of the page and the title of the linked PDF file for relevance. Close screening and selection of GSSO information practice-related resources from the references of each of the included resources were also conducted. Finally, we canvassed an international working group of SGM health experts and organizations for grey literature resources. Duplicates were eliminated. ATLAS.ti was used to support coding and analysis. Researchers met weekly to compare data extraction, ensure consistency of coding, and resolve disagreements through consensus building. Themes were developed through an iterative process of writing and research team discussions. Figure 1 shows a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [11] flow diagram of the article selection process.

XSL•FO

**RenderX**

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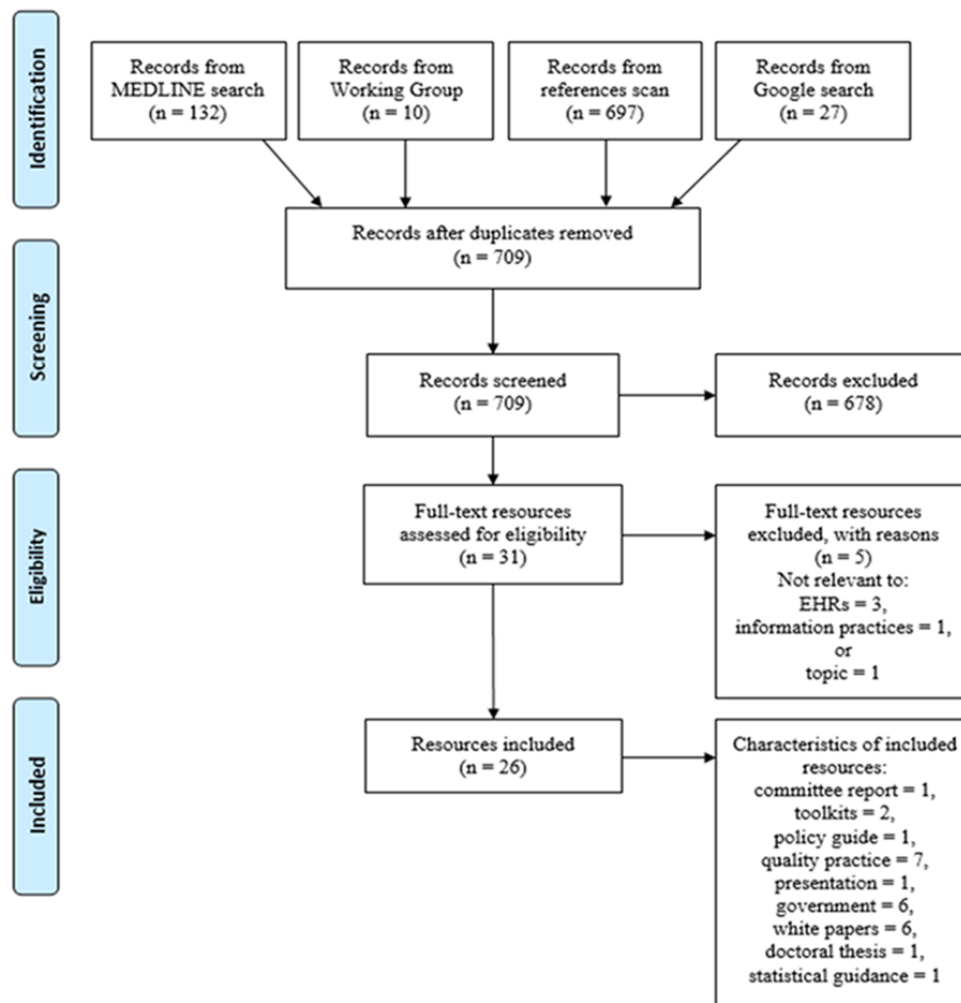


## Results

### Characteristics of Included Articles

The following 26 documents were identified for inclusion: one Canadian parliamentary standing committee on health report [3], two toolkits [4,5], one policy guide [12], seven quality practice documents [7-9,13-16], one presentation on GSSO information best practices [17], six government documents related to changing vital statistics information [18-23], six white papers or data collection standards papers [6,24-28], one doctoral thesis [29], and one statistical guidance paper [30]. Grey literature resources produced by the strategy are summarized in Multimedia Appendix 3.

### Summary of Themes

The primary threads that emerged from the review of grey literature included institutional policies, care practices, social attitudes, and GSSO information practices required to provide safe and quality health care for SGMs. These threads define SGM cultural competence; they naturally and logically constitute the narrative themes that emerged from the literature produced by the organizations addressing SGM barriers to health care through modern GSSO information practices. Cultural competence "…is a set of congruent behaviors, attitudes, and policies that come together in a system, agency, or amongst professionals and enables that system, agency, or those professionals to work effectively in cross-cultural situations" [31]. SGM-competent GSSO information practices are therefore modern GSSO information practices. Modern GSSO information practices emerge from and reinforce the delivery of culturally competent care for SGM. Themes that emerged included (1) essential competency through shared language, terms, and clear definitions of GSSO concepts; (2) the central importance of welcoming and inclusive care environments and inclusive practices for reducing barriers to accessing health care; (3) health care policy that supports competent SGM health care; and (4) modernized GSSO information practices and EHR design requirements that support SGM competence.

### Developing Competency Through Shared Language and Definitions

Gender, sex, and sexual orientation are very different and unambiguous concepts that relate to one another in complex ways for all people [1,4,6,9,16,29]. SGM-competent GSSO information practices are rooted in the language and terminology used to communicate with, for, and about SGMs in health care [4,9]. Standardized definitions and routine use of GSSO data in EHRs support SGM-competent practice by reducing ambiguity, normalizing diversity and inclusion, and affirming

SGM human rights. In this section, we provide working definitions of core GSSO concepts.

## Sex

Sex is a complex biological concept that includes anatomy, physiology, genes, and hormones [4,6,15,26]. Biological sex is based on the physical presence of sex organs (physical or phenotypic sex) [15], sex chromosomes (chromosomal or genotypic sex) [15], and hormone levels (hormonal sex) [6].

### Clinical Sex or Sex for Clinical Use

Clinical sex or sex for clinical use may include some or all the specific concepts of chromosomal sex, hormonal sex, or phenotypic sex, and is applicable to all people. Chromosomal sex does not change, but physical and hormonal sex characteristics can and do change for a wide variety of possible reasons besides hormone therapy (ie, aging and disease) [6]. Sex characteristics may change for a variety of reasons not related to gender transitions, and changes in physical sex characteristics for any patient may therefore warrant procedure-specific adaptations in clinical practice. The screening protocol for imaging [2] may change following a radical hysterectomy related to cancer treatment, as one possible example.

### Intersex People

Intersex people are born with a range of sex organ characteristics that do not fully conform to traditional assumptions of male or female anatomy [4,15,16]. A binary characterization of sex does not fully represent the natural spectrum of chromosomal, hormonal, and anatomical variation [16]. Labeling intersex people (those people who identify as intersex) with terms, words, or concepts that are not intersex, such as *transsexual*, *undifferentiated*, *unknown*, *ambiguous*, *male*, *female*, and *other*, in their health information may be stigmatizing.

### Sex Assigned at Birth

When a person is born, the current practice is that they are assigned a sex based on their external anatomy [4,6,15]. The newborn's sex information is documented in care records and is also submitted to vital statistics (ie, birth certificates). The external genitalia of some people are altered at or around birth via "normalization" procedures [3]. We put the term normalization in quotes because naming such surgeries in this manner may be considered a tangible example of cisgenderism, where binary male and female anatomies are considered normal and anatomies that do not fit within the traditional binary assumptions for anatomy are not normal, with these procedures being undertaken to correct the abnormality. There are growing calls to end these practices unless they are absolutely necessary, and practitioners would ideally wait until the person has appropriately matured and informed consent can be fully supported [3].

## Gender

Gender is a complex human psychological and sociocultural concept that can impact sex characteristics [15,17,26]. Like sex, gender is often erroneously conceptualized and represented as a binary construct (i.e. man/woman), but there is considerably more variation than this in reality [4,6,17]. Gender concepts

should not be conflated with sex concepts, but very frequently they are. Gender is both expressed and felt as a part of a person's identity. As such, the concept of gender and gender-related labels are frequently inappropriately applied to patients by external agents and then used inappropriately in clinical practice.

### Gender Expression

Gender expression is the way people outwardly display their gender [1,4,17,26,30]. There is no single or correct way to express gender; gender expression is dynamic, situational, and individual. People may express different gender identities at different times and in different situations. Only about half of all transgender people live in their felt gender full-time [6], which means that lived gender (or gender expression) may change across different encounters with health care. Lived gender is the gender they express outwardly the majority of the time [30]; lived gender is a person's typical gender presentation. It should therefore not be assumed that gender presentation is congruent with lived gender or gender identity.

### Gender Identity

Gender identity is a person's inner sense or felt gender [1,4,6,8,17,26,30]. Gender identity is different from lived gender because a person may not feel comfortable living/presenting outwardly in their felt gender.

### Names, Pronouns, and Honorifics

Names, pronouns, and honorifics are central to a person's gender identity and are expressions of gender identity for all people [4]. Utilizing chosen names and pronouns is particularly important in the competent care of transgender people because they are affirming of one's gender [4,13,16,29]. *Legal name* is the name that appears on official government identification documents (ID) and is typically the name that was recorded into government and health records at birth. *Chosen name* or name used is the name that people choose to use in interactions. Chosen name may or may not be reflected in government records [4]. *Pronouns* can be gendered (he/him/his and she/her/hers) or gender neutral (them/they/theirs) [4,5]. *Honorifics* are often gendered and include Mr, Mrs, Ms, Ma'am, and Sir [4,5].

### Cisgender, Transgender, and Nonbinary Gender Identities

A cisgender person's felt gender identity matches their sex assigned at birth; a transgender person's gender identity does not [4,15,26,30]. Some people whose felt gender is not a match with their sex assigned at birth may identify with a binary option as their felt gender [4-6]. People with binary gender identities identify as either a man or a woman. People with nonbinary gender identities do not identify within the man/woman binary [15] and may identify as nonbinary, enby, genderfluid, genderqueer, agender, and bigender [4,5,13,25].

### Transition

Transition is the term used by transgender people to encompass the complex dynamic social, public, and sometimes medical process of changing one's gender expression to match their felt gender [4]. Transition can involve changes to chosen names and pronouns used [4,5], clothing and jewelry worn, voice,

vocabulary, mannerisms, anatomy, and physiology [4,16]. Transition may also involve changing legal records to reflect lived gender or gender identity, such as changing one's name [4] or legal gender [5,22]. People who transition may or may not identify using a binary gender term during or following transition.

## Sexual Orientation

Sexual orientation is a term that broadly describes how a person characterizes their emotional and sexual attraction to others [4,12,13,26] and is a complex facet of human sexuality.

Human sexuality is dynamic and multidimensional, and includes a constellation of interrelated concepts such as sexual identity, sexual attraction, and sexual behavior [8,26]. For some people, these concepts align with monogamous heterosexualism. For others, there are important differences. Sexual identity depicts how a person self identifies (eg, lesbian, gay, bisexual, pansexual, and asexual) [4,7,26], and includes both how a person views themselves and what groups they consider themselves to be part of. Sexual attraction describes what gender group or groups one is or is not attracted to [4,7,26]. Sexual behavior refers to the sex acts that one engages in and may be periodic (eg, 1 year) [4,7,26]. Sexual behaviors may occur within one's sexual and/or romantic partners, but also may occur outside of relationships, in groups, or with oneself [26]. Sexual orientation does not strictly determine sexual behaviors [4], and assumptions about sexual behavior should not be made based on sexual orientation.

Health assessments should include information about sexual behaviors and context for all people, regardless of sexual orientation [8], because behavior and context are important factors in the characterization of risk and in the subsequent formulation of relevant preventative teaching and interventions.

## Two-Spirit

The Indigenous concept of Two-Spirit pre-dates Western colonial frameworks for sexual and gender diversity [8]. Ongoing colonial forces include the structural imposition of belief systems that condemn sexual and gender diversity and enshrine cisheteronormative binary constructs. This fuels transphobia and homophobia, resulting in Indigenous SGMs sometimes being forced from their home communities [5]. The English language term Two-Spirit was adopted by Indigenous SGMs for Indigenous SGMs in 1990 at the third gathering of Gay and Lesbian First Nation People in Winnipeg, Manitoba, Canada; the term Two-Spirit can encompass gender, sex, culture, sexuality, and spiritual identities [3-5,15,32]. Not all Indigenous people who are SGMs are Two-Spirit [5]. Adopting a stance of cultural humility and including Indigenous GSSO concepts is essential for culturally safe and SGM-competent health care for Indigenous people [9]. Work in this space must center around, be led by, and involve earnest partnership with Indigenous Two-Spirit people and engage with ongoing processes of decolonization and anti-Indigenous racism.

## Addressing Barriers With Welcoming Health Care Settings and Affirming Practices

In addition to a shared understanding of language and terminology, enabling safe, quality, and inclusive care for all people requires welcoming care environments [4,5,7]. EHRs that support expanded GSSO concepts, and psychologically safe and meaningful care interactions between patients and health care staff are essential structures that drive and support modern practices [7]. Modern GSSO information practices are free of assumption and stereotype, and meaningfully address psychological safety of care environments, administrative interactions, care interactions, and other barriers to access. Administrative databases, anatomic inventories, and special considerations for the challenges faced by transgender, Indigenous, older, and newcomer SGM people are core foci for health equity.

## Welcoming Care Environments

Health care ought to be affirming and welcoming; a place where patients can feel accepted, safe, and cared for [4,5]. All health care staff are ambassadors of the health care system and should therefore receive SGM-competency training [15]. Affirming health care environments address a key barrier to access for SGMs (pervasive cisheteronormativity) [4,5]. Symbols of diversity and inclusion, such as rainbow flags, affirming images, posters that signal respect and inclusion, and nongendered bathrooms [4,5,8,12,13,15,16,29], send the message that health care is safe and welcoming of diversity. The use of culturally competent language and terms [9,13,15]; careful respect of patient privacy and confidentiality [4,8,30]; and avoidance of assumptions about patient names, pronouns, honorifics, gender identity, sex at birth, and sexuality are essential because assumptions can lead to negative experiences for both patients and staff [16]. Every interaction with health care should not be discounted in terms of its overall importance to ensuring equity. Competencies of services and providers within them are assessed by each patient at each interaction, and decisions about accessing health care or recommending care to others are made. As a logical consequence of experiencing stigma and discrimination, SGMs may not feel comfortable disclosing details about their sexual orientation, sexual behaviors, or gender identity [4,12,29]. Patients should not be made to feel forced to disclose GSSO information.

## Modernized GSSO Information Practices

SGM-competent practices avoid assumptions, misgendering, outing, and deadnaming, protect patient privacy and confidentiality [4], and ensure that current GSSO information is collected and documented into EHR systems to maximize the safety and quality of primary and secondary uses [17]. Modern GSSO information practices are expanded to be inclusive and support affirming care by ensuring that name, gender identity, and pronouns used are meaningfully integrated into technical, social, and organizational layers of health care systems [24]. Routine collection of GSSO information is important for measuring, monitoring, and improving population health [8].

## Patient Registration

Patient registration involves gathering key administrative and demographic information from patients and entering that information into an electronic or paper patient record. Registration should be done in a private nonintrusive manner for all people [7,15,16], and in addition to being staff facilitated, it can be done through transcription of a paper form [7], direct patient entry self-interviews via a patient portal [7,16], a self-serve kiosk [16], or a mobile device [7]. When summoning patients to a registration desk is necessary, calling them by their last name without the use of gendered honorifics, pronouns, or legal names is a useful practice for avoiding negative experiences until preferred names, pronouns, and honorifics are confirmed by the patients and their consent to share that information with the care team is obtained. Patients may be more comfortable sharing GSSO information using self-interviews [12]. The dynamic nature of gender identity should be considered in patient intake and registration workflows, and GSSO information should be confirmed regularly [4,7,8,16]. Consistent and appropriate gender identification and use of appropriate names and correct pronouns for patients are simple yet important actions for providing affirming health care [4,5,8,13] at and beyond registration, and they demonstrate organizational and professional commitment to an inclusive and affirmative health care culture.

## Patient Identification

Governments are beginning to modernize vital statistics to be more SGM inclusive. At the federal level in Canada, citizens who do not identify with binary sex or gender designations can have an "X" printed on their passport, travel document, citizenship certificate, or permanent resident card [4,19,20]. Provinces, including Ontario, British Columbia, and Manitoba, are also aligning their systems to enable more diverse gender or sex designations on government-issued IDs through a similar mechanism [21-23].

Government-issued IDs are considered trusted sources for identifying patients for health care purposes [15]; however, there is a high degree of conflation of the two concepts within jurisdictional information systems [2] and subsequent variation in how administrative sex and gender data are labeled and presented on IDs from jurisdiction to jurisdiction. Administrative sex and/or gender on government IDs may not match with the gender identity (or felt gender), gender expression (or lived gender), or sex assigned at birth; administrative gender may even be labeled as *sex* on government IDs [22]. Use of administrative sex or gender for clinical care can lead to substantial confusion in information systems and clinical care and can increase the risk to patient safety.

A two-step approach that includes sex assigned at birth and gender identity is considered the best practice for collecting gender and sex information by leading organizations [6-8,13,15,24] because it eliminates conflation between administrative and clinical sex and/or gender concepts and reduces conflation-related risks. However, health organizations must respect the patient's choice to not provide this information, and staff are strongly cautioned to avoid making assumptions about patient identity or applying labels to patients that patients themselves may not identify with (ie, transgender) based on this information. Administrative sex and/or gender should not be labeled or documented as anything but administrative sex and/or gender to avoid conflation with clinical concepts. Nonbinary people may not wish to have gender markers on their documents [21]. Personal health numbers can, and should, be used for positive identification in the case of mismatches in administrative sex and/or gender information.

## Clinical Encounters

Thorough assessment and documentation of medical and surgical history in postregistration encounters can help ensure that opportunities for appropriate screening are not missed [4,13]. Preferred names and pronouns should be collected in the first visit [8]. Staff can positively identify their own names and pronouns when first meeting with a patient to build trust and a sense of safety [4,16]. Providers should ask about sexual orientation, sexual behavior, and gender identity using open-ended questions [8,12], and should confirm names and pronouns used with the patient so that they can correct any incorrect GSSO information that has been erroneously documented in the patient chart and minimize the possibility of it being used in subsequent clinical encounters. If affirming approaches are unsuccessful, GSSO information may be collected as part of a more formal and more focused sexual health assessment [12]. Clinicians who ensure that documentation is complete, accurate, and up to date with current GSSO information reduce the chances for negative experiences for both patients and providers in subsequent care interactions [24].

GSSO information should be collected without the presence of parents for youth aged 13 years or above, but parents should be involved in the conversation for children aged 12 years or below [7].

## Addressing Health Inequities

Gaps in preventative screening for SGMs exist because outdated GSSO information practices and binary administrative sex and/or gender data constructs interfere with appropriate matching of SGM patients' preventive screening and health care needs [4]. Anatomic inventories enable transparent and assumption-free therapeutic relationships between patients and providers. Safe therapeutic relationships lead to increased opportunities for prudent screening and prevention [1,2,4,5,9].

### Anatomic Inventories

Anatomic inventories are objective catalogues of anatomical parts that allow assumption-free unambiguous representation of diverse bodies to support appropriate screening, treatment, and referral decisions matched to biological needs [1]. If a person has a particular body part or organ and otherwise meets criteria for screening based on risk factors or symptoms, screening can and should proceed regardless of hormone use [13] or other sex and/or gender characteristics or identities. Anatomic inventories ought to be evaluated as the best practice for all people given that they eliminate unconscious profiling based on sex assigned at birth or administrative gender. Providers are encouraged to maintain a current anatomic/organ

inventory in their patients' charts to guide quality care planning [13].

### Screening Considerations

SGMs are diverse [16] and face specific barriers to health care and associated health risks [2,4,13] that necessitate differential screening [1,8]. SGMs are at higher risk of depression, suicidal ideation and attempts, substance use, chronic disease, sexually transmitted and blood-borne infections/illnesses, homelessness (especially youth), and food insecurity [3], and should be routinely and competently screened [13]. Long waitlists and overly complex assessments and referral processes, often from clinicians in private practice (eg, psychologists, psychiatrists, and social workers), for gender-affirming surgeries are possible barriers to access [3], and sex designation on government IDs and health documentation may affect billing and eligibility for sex-specific diagnostics or procedures [29].

### Transgender SGMs

Transgender people are made particularly vulnerable by health care that is not SGM competent. An estimated 30% of transgender people do not engage with emergency services when needed because of the perceived risk of being stigmatized, and half of those who do engage with health care report negative experiences [3]. It is important to note here that gender dysphoria is felt differently by different people regarding different aspects of their bodies, gender role, or assigned sex [5]. Not all transgender people seek intervention or support with transition beyond those necessary to affirm their own gender identity [4,13]. However, limited access to gender-affirming surgeries is associated with high-risk behaviors seeking to address intense gender dysphoria such as self-injection of silicone obtained from illicit sources without medical guidance, supervision, or follow-up [13]. Trans Care BC and Alberta Health Services are health organizations that have both released toolkits to support gender-affirming care practice in Canada [4,5].

### Indigenous SGMs

Two-Spirit and Indigenous SGMs face discrimination and stigma because of their sexual orientation, their gender identity, and, in some cases, their HIV or hepatitis C virus (HCV) status [3]. Culturally safe and SGM-competent care practice for Two-Spirit is critical for reducing barriers to access and for improving health outcomes for Indigenous SGMs [3-5].

### Older SGM Adults

Older SGM adults are at risk of social isolation and limited social support due to historic criminalization and ongoing stigma. In the absence of SGM-competent programs along the continuum of care (ie, SGM retirement homes), older SGM adults may face the decision to recloset in order to avoid the trauma of being stigmatized by other patients and staff in unfriendly and unwelcoming care environments [3].

### Newcomers

SGMs who immigrate face additional barriers that should be considered by clinicians providing care [3,9]. Cultural ways of expressing gender, sex, and sexual orientation in language, behavior, and appearance may vastly differ from practices in mainstream Canada. Furthermore, the intensity of stigma, trauma, and persecution that people may have faced before immigrating is an essential consideration in the care of newcomer SGMs. Cultural competency (ethnic) is central to cultivating health equity for culturally diverse people [14]. Providers should consider the following when providing care for newcomer SGMs:

- The immigration process in Canada can be stressful and costly and can involve psychiatric and other assessments [3]; newcomers may become aware of their HIV or HCV status through the process and should be offered appropriate counselling and support.
- Gender-positive messaging in languages other than official languages can help reduce language as a barrier to accessing health care, and anonymous and confidential access to sexual health services can increase opportunities for screening, prevention, referral, and treatment [3].
- Expanded gender identifiers can enable less stigmatizing immigration processes [19,20].

## SGM-Competent Health Care Policy

In 2017, the *Canadian Human Rights Act* was amended to include gender identity and gender expression as prohibited grounds for discrimination [4,6]. All Canadians have the fundamental right to identify and express gender in a manner that is consistent with their identity [4]. The requirement for mandatory ongoing training for all staff in health care (physicians, nurses, social workers, clerks, administrative staff, and others) is aligned with the objective of improving knowledge and changing attitudes toward SGMs in support of these rights [3,4,27,29]. Health care policies and procedures should be evaluated for diversity, equity, and inclusion.

## Training and Education

Health care providers have the responsibility to ensure respectful, compassionate, and equitable interactions with patients [16]. Since cultural competency requires changes in the attitudes and beliefs of the people who provide health care, both clinical and nonclinical staff require regular SGM-competency training [13,15] and evaluation. Educating staff about modern GSSO information practices is an essential step toward SGM-competent care [4,14,17]. SGM competency training reduces the knowledge gap and increases opportunities for psychologically safe therapeutic conversations between patients and providers. In turn, this can improve access to care [27] by reducing the barrier of stigma and the burden of education that many SGMs feel [3-5,29]. It is not uncommon for SGMs, especially transgender people, to have to educate their providers about their own care [29]. Providers should be receptive to this as a form of patient self-advocacy and engage in their own professional development [14]. Learning about the factors that impact SGM health may mean having to reflect on personal bias [4] and obtaining a nuanced understanding of GSSO.

Staff should be trained in small groups to support psychological safety. Health information technology staff should begin training early [8] because inclusive data structures support affirmative competent clinical practice and organizational culture [7]. Staff

from EHR vendors and insurance companies also need to be educated on the health needs of SGMs [7]. Enthusiastic champions should be used to lead culture change in these organizations [8]. Digital solutions for training can be offered and completed remotely and are one option for addressing disparities in urban-rural service offerings [5,8].

## Policies for Discrepancies in GSSO Information

EHRs facilitate instantaneous communication of accurate and inaccurate information among the care team. Policies and procedures for correcting inaccurate GSSO information at registration are required to avoid downstream confusion of terms [15] and risks to patient safety. Inclusion of expanded unambiguous GSSO concepts in EHRs provides appropriate space for GSSO information and is necessary to eliminate health data invisibility and to support SGM-competent care [7]. Where discrepancies exist between labels, codes, and data elements, labels and data dictionaries should be corrected quickly to avoid harm to patients resulting from their use in clinical care.

## Laboratory Policies

Clear organizational support and guidance for GSSO laboratory information practices is required to ensure patient safety [15] because laboratory results frequently inform clinical interventions such as the appropriate dosage of medications. Changes to administrative gender codes by jurisdictions may impact downstream laboratory and EHR systems [15]; implementation requires planning, particularly as it relates to the gender or sex value "X" [4,8]. Incongruencies in sex, gender, and name listed on samples should be confirmed, rather than discarded [16]. Clear labeling on screening samples is required to ensure that samples are processed correctly, despite the possibility of gender marker incongruence between EHRs and laboratory information systems [13,16]. Sex assigned at birth, legal sex, and gender identity data elements provide necessary information for interpretation [15], and incongruence between gender elements and sex at birth can guide safe laboratory interpretation and reporting practices [16] provided that staff have the required competency. Making both male and female reference ranges available to clinicians can improve safety for SGMs [8]. Hormone levels for nonbinary people may intentionally be in between typical male and female levels [13].

## Privacy and Confidentiality

Policies governing privacy and confidentiality are a cornerstone of SGM-competent health care. It is quite logical that SGM patients may be reluctant to share information about themselves, particularly in the context of open waiting rooms or areas where privacy cannot be assured [12] given the risk of stigma-related trauma. Privacy is very important to SGMs, especially transgender people [30] and people who have experienced harm via public acts of being misgendered, outed, or deadnamed in previous health care experiences. Organizations that collect GSSO information must ensure that the privacy and confidentiality of patients is maintained [4,7]. Technical safeguards must be in place to ensure privacy and confidentiality of the EHRs and data within them [12], and patients can be informed about these safeguards. SGM-competent health care,

with modern GSSO information practices, addresses risks related to breaches in privacy and confidentiality for SGM patients [4].

### Information Sharing

EHRs enable the instantaneous sharing of health information. While sharing of limited patient information between providers is sanctioned so long as it is limited to authorized purposes (ie, continuity of care) and follows specific security protocols [9], providers should still obtain consent to document GSSO information in patient charts. Patients can be assured that their health information will be kept confidential [12] according to health privacy legislation [4] and will not be shared unnecessarily.

## Informed Consent

Informed consent is an important patient right and, where applicable, should include reproductive and fertility considerations [13] for the wide range of gender-affirming hormonal therapies and surgical procedures that are available [4]. Transgender people, for example, have the same range of reproductive desires as nontransgender people; reproductive goals for gender-affirming procedures should be considered before they are undertaken. Gonadectomies, for instance, will have clear impacts on fertility that should be discussed [13].

## Insurance Coverage

Costs for medically necessary procedures are a barrier to health care and should not be borne by patients alone. Reimbursement policies should recognize the unique health needs of SGMs [3,7], and insurance companies should support gender-affirming surgeries that are deemed medically necessary [3]. Binary administrative sex and/or gender structures that govern coverage eligibility for certain procedures are commonly cited as problematic barriers to access and to health equity for transgender people [29].

## Conversion Therapy

Conversion therapy is a broad term that encompasses health care practices intended to change an individual's sexual orientation or gender identity. Conversion therapy is considered to have negative effects on the mental health of SGM patients, and there have been calls to ban such practices in progressive jurisdictions [3].

## Modern GSSO Information Practices and SGM Competency Within EHRs

EHRs support clinical practice and are an extension of clinical culture. Adoption of modern GSSO information practices in EHRs addresses the invisibility of SGMs on user interfaces, in data repositories, in health terminology, in classification and coding schemes, and in messaging and exchange standards. EHRs designed to support modern GSSO information practices improve outcomes by supporting SGM-competent health care.

## Standardized GSSO Data Collection

Data invisibility and institutional erasure of SGMs are increasingly recognized as structural problems contributing to poor outcomes for SGMs [3,6,7,16]. The passive erasure of gender and sexual diversity in EHR code systems has led to inaccurate documentation, unsafe care, invisibility of SGMs in

health data [16], and inaccurate information for health care providers, health information professionals, and researchers [7,33]. Addressing the erasure of SGMs in the data means developing and implementing unambiguous data standards as well as methods for analytics and statistics that are inclusive of SGMs. Policies and practices that support routine collection of expanded GSSO information are essential.

## Routine Collection and Use of Standardized Expanded GSSO Concepts

EHRs must include expanded GSSO concepts [7,16,17] or else they risk acting as barriers to quality care [29]. The use of expanded GSSO data elements (beyond the binary) enables culturally competent and clinically safe care and provides a richness of discrete variables for clinical care and analytic uses that does not currently exist in most EHRs [1,6,15-17,29]. Routine standardized collection of GSSO information that includes expanded GSSO concepts is recommended by health care quality organizations as best practice and as a foundational step to understanding and addressing inequities for SGMs [7,14].

## Not Using "Other" as a Code

Use of *other* as a sex or gender code is a common and outdated information practice in EHRs [2] and statistical surveys [18,26]. It is often done with the intent of capturing sex and gender data that are not binary male/female or man/woman, but has the effect of normalizing binary elements and obfuscating the range of possible nonbinary elements. This practice creates ambiguity in code sets and data sets and can be stigmatizing for SGMs, particularly in the context of patient-user interfaces where a person may be forced to "other" themself via a health care artefact. *Please specify* is one option [18] that can provide space for entering correct information where correct options do not exist. However, a full complement of discrete expanded GSSO terms is preferable because discrete elements can support logic-based rules, decisions, algorithms, and artificial intelligence.

## Health Analytics and Statistics

SGM health information facilitates improved knowledge about their health and liberates evidence that can be applied for quality improvement [7,26]. Gender-based analysis is important for statistical comparisons [30]. National [3] and international [30] efforts to address data invisibility for SGMs are now underway by national statistics organizations [6,26]. Due to the small proportion of people with diverse gender identities, special statistical considerations are required to prevent breaches of confidentiality on result maps and to ensure inclusiveness in mathematical and statistical modeling [2,30]. In order to harmonize and strengthen the measurement of health inequalities in Canada, the Canadian Institute for Health Information has proposed Canadian equity stratifiers (standard data elements) that support quality health information for improving outcomes [6].

## Administrative and Legal GSSO Concepts

Administrative or legal GSSO concepts are found in jurisdictional repositories and are used by governments for administrative reasons as minimum data elements collected and used for vital statistics, by insurance agencies for billing, and by health care organizations for patient registration [1,2,15,19-21,23,28]. Administrative GSSO concepts include legal name, administrative gender, administrative sex, and sex assigned at birth, amongst others. Many jurisdictional repositories represent sex and gender concepts as a single data element using a male/female binary supplemented by a category of *other*, *unknown*, or *undifferentiated* [1,2]. The integration of administrative and health care information systems must be done carefully such that administrative and clinical GSSO data elements are appropriately sequestered. Clinical data elements must be clearly labeled and easily differentiated from administrative sex and/or gender data elements in order to support safe and SGM-competent clinical care.

## *Discussion*

### SGM-Competent GSSO Information Practices as an Equity-Oriented Intervention

Health equity is the absence of unfair differences in access to quality health care or health outcomes [6,16], and is supported by modern GSSO information practices that aim to reduce barriers to access and provision of safe and quality health care to SGMs. Equitable access to quality health care is a right for all Canadians and should be a right for all people including SGMs.

### Beyond the Binary

Binary representation of sex and gender in EHRs obfuscates natural and cultural diversity and, in the context of health care, places SGM patients at risk of clinical harm because it leads to clinical assumptions. Outdated GSSO information practices render SGMs invisible in therapeutic interactions, data sets, and bias analytics, and ensures that researchers have little quality SGM data with which to improve health care for SGMs.

### Clinical Care That Affirms Gender Identity

Knowledge and use of inclusive and gender-affirming language in clinical interactions in the context of welcoming care environments with SGM-competent staff and EHRs that include and display unambiguous GSSO information are all key elements of modern GSSO information practices. Modern GSSO practices are assumption free, unambiguous, and antistigma, and are necessary to support the psychological safety of SGM patients in health care spaces. Modern GSSO information practices support quality therapeutic relationships that, in turn, enable thorough and appropriate screening, care, and referral of SGM patients.

### Moving Forward

Health care system performance may continue to improve as new and innovative technologies are implemented, but if GSSO information practices are not modernized, the health inequities of SGMs are likely to persist or worsen [6]. Further research is needed to improve primary care for transgender people [29], to development specific laboratory guidelines and reference ranges for the gender diverse community [15], and to improve the accuracy of hormone therapy [16]. SGM-inclusive analytics and outcome measures need to be developed, and future research projects should consider using an intersectional lens to gain

further insight into structural factors such as institutionalized prejudice faced by SGMs. Anatomic inventories are a core tool that should be evaluated as the best practice for modern GSSO information practices in health care for all people. Both private and public sector health technology regulations ought to include specifications for interoperable SGM-competent system design, including the use of expanded GSSO concepts and definitions in EHRs.

## Institutionalizing SGM Competency

Health equity for SGMs requires systemic change. Agencies and agents in health care need to be equipped with the knowledge and tools needed to cultivate modern attitudes, policies, and practices that enable health equity for SGMs. Adopting small but important changes in the language and terminology used in technical and social health care systems is essential for institutionalizing SGM competency. Modern GSSO information practices depend on and reinforce SGM competency in health care.

## Limitations

The primary limitations of this review include language and timeframe. Only English-language documents published in the past 6 years and indexed within a single database (MEDLINE) were included. The authors are not bilingual. Publication year range was limited for expediency and for recency. Relevant documents from outside of the time frame or published in other languages or databases may have been missed. All included articles were from Western countries and cultures, limiting generalizability to other settings and creating a bias for Western ideals and strategies. Efforts to mitigate these limitations included seeking resources from an international group of experts and stakeholders that represented non-Western countries and non–English-speaking persons.

## Conflicts of Interest

None declared.

Multimedia Appendix 1
Grey literature search strategy.
[DOCX File , 15 KB - medinform_v9i2e25467_app1.docx ]

Multimedia Appendix 2
Coding scheme.
[DOCX File , 14 KB - medinform_v9i2e25467_app2.docx ]

Multimedia Appendix 3
Results summary table.
[DOCX File , 17 KB - medinform_v9i2e25467_app3.docx ]

## References

1. Lau F, Antonio M, Davison K, Queen R, Devor A. A rapid review of gender, sex, and sexual orientation documentation in electronic health records. J Am Med Inform Assoc 2020 Nov 01;27(11):1774-1783 [FREE Full text] [doi: 10.1093/jamia/ocaa158] [Medline: 32935124]
2. Lau F, Antonio M, Davison K, Queen R, Bryski K. An Environmental Scan of Sex and Gender in Electronic Health Records: Analysis of Public Information Sources. J Med Internet Res 2020 Nov 11;22(11):e20050 [FREE Full text] [doi: 10.2196/20050] [Medline: 33174858]
3. Committee report: the health of LGBTQIA2 communities in Canada. Canadian House of Commons. 2019. URL: https://www.ourcommons.ca/DocumentViewer/en/42-1/HESA/report-28/ [accessed 2020-04-26]
4. Sexual Orientation, Gender Identity & Gender Expression (SOGIE) Safer Places Toolkit. Alberta Health Services. 2019. URL: https://www.albertahealthservices.ca/assets/info/pf/div/if-pf-div-sogie-safer-places-toolkit.pdf
5. Gender-affirming Care for Trans, Two-Spirit, and Gender Diverse Patients in BC: A Primary Care Toolkit. Provincial Health Services Authority. 2018. URL: http://www.phsa.ca/transcarebc/Documents/HealthProf/Primary-Care-Toolkit.pdf [accessed 2020-05-10]

6. In pursuit of equity: defining stratifiers for measuring health inequality-a focus on age, sex, gender, income, education and geographic location. Canadian Institute for Health Information. 2018 Apr. URL: https://www.cihi.ca/sites/default/files/document/defining-stratifiers-measuring-health-inequalities-2018-en-web.pdf [accessed 2020-05-12]

7. Ready, Set, Go! Guidelines and Tips For Collecting Patient Data on Sexual Orientation and Gender Identity (SOGI) – 2020 Update. National LGBT Health Education Centre (US). 2020. URL: https://www.lgbtqiahealtheducation.org/publication/ready-set-go-guidelines-tips-collecting-patient-data-sexual-orientation-gender-identity/ [accessed 2020-05-11]

8. Collecting Sexual Orientation and Gender Identity Data in Electronic Health Records. National LGBT Health Education Centre (US). 2018. URL: https://www.lgbtqiahealtheducation.org/wp-content/uploads/Collecting-Sexual-Orientation-and-Gender-Identity-Data-in-EHRs-2016.pdf [accessed 2020-05-11]

9. BC Centre for Disease Control. BCCDC COVID-19 language guide: guidelines for inclusive language for written and digital content. BC Centre for Disease Control. 2020 Jul. URL: http://www.bccdc.ca/Health-Info-Site/Documents/Language-guide.pdf [accessed 2020-07-31]

10. Haby M, Chapman E, Clark R, Barreto J, Reveiz L, Lavis J. What are the best methodologies for rapid reviews of the research evidence for evidence-informed decision making in health policy and practice: a rapid review. Health Res Policy Syst 2016 Nov 25;14(1):83 [FREE Full text] [doi: 10.1186/s12961-016-0155-7] [Medline: 27884208]

11. Moher D, Liberati A, Tetzlaff J, Altman D, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009 Jul 21;339:b2535 [FREE Full text] [doi: 10.1136/bmj.b2535] [Medline: 19622551]

12. Policy focus: how to gather data on sexual orientation and gender identity in clinical settings. National LGBT Health Education Centre (US). 2019. URL: https://www.lgbthealtheducation.org/wp-content/uploads/policy_brief_how_to_gather.pdf [accessed 2020-05-06]

13. Guidelines for the Primary and Gender-Affirming Care of Transgender and Gender Nonbinary People. Center of Excellence for Transgender Health. 2016. URL: https://transcare.ucsf.edu/sites/transcare.ucsf.edu/files/Transgender-PGACG-6-17-16.pdf [accessed 2020-05-10]

14. Environmental Scan Report: Guide to Improving Patient Safety in Primary Care Settings by Engaging Patients and Families. Agency for Healthcare Research and Quality. 2017 Mar. URL: https://www.ahrq.gov/sites/default/files/wysiwyg/professionals/quality-patient-safety/patient-family-engagement/pfeprimarycare/envscan-report-final.pdf [accessed 2020-05-10]

15. Laboratory considerations for care of gender diverse patients. British Columbia Provincial Laboratory Medicine Services. 2019. URL: http://www.bccss.org/bcaplm-site/Documents/Updates/Lab%20considerations%20for%20care%20of%20gender%20diverse%20patients.pdf [accessed 2020-05-11]

16. Care Considerations for Inclusion of Gender Diversity within Medical Laboratory Services White Paper. Institute for Quality Management in Healthcare (IQMH). 2019. URL: tinyurl.com/3cxl48np [accessed 2020-05-11]

17. Care Considerations for Inclusion of Gender Diversity within Medical Laboratory Services Presentation. Institute for Quality Management in Healthcare (IQMH). 2019. URL: https://med-fom-polqm.sites.olt.ubc.ca/files/2019/11/For-website-Gender-Presentation-For-Sharing-Post-Conference.pdf [accessed 2020-05-11]

18. Standard for sex and gender variables. Australian Bureau of Statistics. 2016. URL: https://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/1200.0.55.012Main%20Features242016?opendocument&tabname=Summary&prodno=1200.0.55.012&issue=2016&num=&view= [accessed 2020-05-12]

19. Canadians can now identify as gender "X" on their passports. Government of Canada. 2019 Jun 04. URL: https://www.canada.ca/en/immigration-refugees-citizenship/news/notices/gender-x-documents.html [accessed 2020-05-12]

20. Choose or update the gender identifier on your passport or travel document. Government of Canada. URL: https://www.canada.ca/en/immigration-refugees-citizenship/services/canadian-passports/change-sex.html [accessed 2020-05-12]

21. Gender on Health Cards and Driver's Licences. Ontario Ministry of Government and Community Services. 2016 Jun 29. URL: https://news.ontario.ca/en/backgrounder/40957/gender-on-health-cards-and-drivers-licences [accessed 2020-05-12]

22. Change your personal information. Government of British Columbia. URL: https://www2.gov.bc.ca/gov/content/governments/government-id/bc-services-card/your-card/change-personal-information [accessed 2020-05-12]

23. Manitoba Vital Statistics Agency. Change of sex designation. Published online. Winnipeg: Government of Manitoba Vital Statistics; 2020 Apr 2020. URL: https://vitalstats.gov.mb.ca/change_of_sex_designation.html [accessed 2020-05-12]

24. Little C, Cheevers C. Best practices: electronic medical record data collection on sexual orientation and gender identity (SOGI). One Colorado. 2019 Jun. URL: https://one-colorado.org/wp-content/uploads/2019/06/Best-Practices_-Electronic-Medical-Record-Data-Collection-on-Sexual-Orientation-and-Gender-Identity-SOGI.pdf [accessed 2020-05-12]

25. Morgan R, Dragon C, Daus G, Holzberg J, Kaplan R, Menne H, et al. Updates on terminology of sexual orientation and gender identity survey measures. Federal Committee on Statistical Methodology. 2020 Aug. URL: https://nces.ed.gov/FCSM/pdf/FCSM_SOGI_Terminology_FY20_Report_FINAL.pdf [accessed 2020-08-29]

26. Current measures of sexual orientation and gender identity in federal surveys working paper. United States Federal Interagency Working Group on Improving Measurement of Sexual Orientation and Gender Identity in Federal Surveys. 2016 Aug. URL: https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/04/WorkingGroupPaper1_CurrentMeasures_08-16.pdf [accessed 2020-05-12]

27.    Measuring health equity: demographic data collection and use in Toronto central LHIN hospitals and community health centres. Sinai Health System. 2017 Jul. URL: http://torontohealthequity.ca/wp-content/uploads/2013/02/ Measuring-Health-Equity-Demographic-Data-Collection-Use-in-TC-LHIN-Hospitals-and-CHCs-2017.pdf [accessed 2020-05-12]

28.    An overview of Albertas electronic health record information system. Alberta Health and Alberta Health Services. 2015. URL: https://www.albertanetcare.ca/documents/An_Overview_of_Albertas_ERHIS.pdf [accessed 2020-05-12]

29.    Zeigler E. Exploring the implementation and delivery of primary care services for transgender individuals: an Ontario case study. McMaster University. 2019. URL: https://macsphere.mcmaster.ca/handle/11375/24479 [accessed 2020-05-10]

30.    United Nations Economic Social Council. Economic commission for Europe conference of economic statisticians, 67th plenary session report: in-depth review of measuring gender identity. 2019 Feb Presented at: Conference of European Statisticians; June 26-28, 2019; Paris URL: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2019/ ECE_CES_2019_19-G1910227E.pdf

31.    Cross T, Bazron B, Dennis K, Isaacs M. Towards a culturally competent system of care: a monograph on effective services for minority children who are severely emotionally disturbed. Seattle Pacific University. 1989 Mar. URL: https://spu.edu/ ~/media/academics/school-of-education/Cultural%20Diversity/ Towards%20a%20Culturally%20Competent%20System%20of%20Care%20Abridged.ashx [accessed 2020-05-11]

32.    Felice M. Two-Spirit. The Canadian Encyclopedia. URL: https://www.thecanadianencyclopedia.ca/en/article/two-spirit [accessed 2020-08-25]

33.    Waite S, Denier N. A Research Note on Canada's LGBT Data Landscape: Where We Are and What the Future Holds. Can Rev Sociol 2019 Feb;56(1):93-117 [FREE Full text] [doi: 10.1111/cars.12232] [Medline: 30793865]

## Abbreviations

**EHR:** electronic health record
**GSSO:** gender, sex, and sexual orientation
**HCV:** hepatitis C virus
**SGM:** sexual and gender minority

XSL•FO
**RenderX**

Original Paper

# Leadership Perspectives on Implementing Health Information Exchange: Qualitative Study in a Tertiary Veterans Affairs Medical Center

Brian E Dixon[1,2,3], PhD, MPA; Cherie Luckhurst[1,4], PhD; David A Haggstrom[1,5,6], MD, MAS

[1]VA HSR&D Center for Health Information and Communication, Richard L Roudebush Veterans Affairs Medical Center, Indianapolis, IN, United States

[2]Department of Epidemiology, Indiana University Richard M Fairbanks School of Public Health, Indianapolis, IN, United States

[3]Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN, United States

[4]Orthopaedic Surgery Research, Riley Hospital for Children, Indiana University School of Medicine, Indianapolis, IN, United States

[5]Division of General Internal Medicine and Geriatrics, Indiana University School of Medicine, Indianapolis, IN, United States

[6]Center for Health Services Research, Regenstrief Institute, Indianapolis, IN, United States

**Corresponding Author:**
David A Haggstrom, MD, MAS
VA HSR&D Center for Health Information and Communication
Richard L Roudebush Veterans Affairs Medical Center
1481 West 10th St
Indianapolis, IN, 46202
United States
Phone: 1 3179882067
Email: dahaggst@iupui.edu

## Abstract

**Background:** The US Department of Veterans Affairs (VA) seeks to achieve interoperability with other organizations, including non-VA community and regional health information exchanges (HIEs).

**Objective:** This study aims to understand the perspectives of leaders involved in implementing information exchange between VA and non-VA providers via a community HIE.

**Methods:** We interviewed operational, clinical, and information technology leaders at one VA facility and its community HIE partner. Respondents discussed their experiences with VA-HIE, including barriers and facilitators to implementation, and the associated impact on health care providers. Transcribed interviews were coded and analyzed using immersion-crystallization methods.

**Results:** VA and community HIE leaders found training to be a key factor when implementing VA-HIE and worked cooperatively to provide several styles and locations of training. During recruitment, a *high-touch approach* was successfully used to enroll patients and overcome their resistance to opting in. Discussion with leaders revealed the high levels of complexity navigated by VA providers and staff to send and retrieve information. Part of the complexity stemmed from the interconnected web of information systems and human teams necessary to implement VA-HIE information sharing. These interrelationships must be effectively managed to guide organizational decision making.

**Conclusions:** Organizational leaders perceived information sharing to be of essential value in delivering high-quality, coordinated health care. The VA continues to increase access to outside care through the VA Maintaining Internal Systems and Strengthening Integrated Outside Networks (MISSION) Act. Along with this increase in non-VA medical care, there is a need for greater information sharing between VA and non-VA health care organizations. Insights by leaders into barriers and facilitators to VA-HIE can be applied by other national and regional networks that seek to achieve interoperability across health care delivery systems.

XSL•FO
RenderX

## *Introduction*

### Background

Health information exchange (HIE) is the electronic transfer of data or information between health care organizations [1], although the format and context of the information may vary. For example, patient information might be used directly in clinical care, or it might inform the management of populations, such as patients attending a clinic for diabetes care [2]. Among US hospitals, 55% reported participating in an organized HIE network that facilitates data sharing [3]. Nonhospital adoption is lower, as 38% of US physicians report sharing information electronically [4].

In 2009, the US Department of Veterans Affairs (VA) launched an interoperability program called Virtual Lifetime Electronic Record Health, which has subsequently been rebranded as the Veterans Health Information Exchange (VHIE) program [5]. The program sought to create a comprehensive, dynamic medical record for each veteran, which is accessible to all VA organizations, regardless of location. Outside the VA, many health care providers are connected to networks that operate within specific geographical regions to exchange health information. To share information, the VA joins with non-VA regional exchanges using the nationwide eHealth Exchange platform [6]. In this paper, HIE efforts between VA and non-VA organizations will be called *VA-HIE*.

Previous research regarding the VA-HIE program has examined the technical infrastructure necessary to enable system interoperability [7-11], the informational value of enabling access to non-VA data [12], the adoption of HIE by patients [13], the impact of HIE on short-term medical costs [14], and the early experiences of clinicians and patients using VA-HIE systems [15]. These studies, like those of the broader US health system, found overall low adoption rates of HIE by both patients and providers. For successful scale-up to a national delivery system that integrates both regional HIE and VA systems, we must identify the implementation practices required for interoperability. Recently, the VA increased its efforts to expand information sharing between its VA medical centers and non-VA community or regional HIEs.

### Objectives

To gain insight into successful interoperability practices, we interviewed organizational leaders associated with a tertiary VA facility following a VA-HIE implementation. The perspectives of VA leaders in information technology, operations, and clinical care and those of community HIE partners were sought to understand the facilitators and barriers to HIE implementation. Insights from leaders within and outside the VA can guide future implementation of HIE in multiple contexts.

## *Methods*

### Study Design

To understand the factors that serve as facilitators or barriers to interoperability, we conducted semistructured interviews with key leaders associated with VA-HIE implementation at the Richard L. Roudebush Veterans Affairs Medical Center (VAMC) in Indianapolis, Indiana.

### Organizational Setting and Implementation Details

The Indianapolis VA was the organizational setting in which the study was conducted, and the Indiana Health Information Exchange (IHIE) was the community-based HIE network [16] with which VA partnered. Further details of the organizational setting and implementation timeline can be found in Multimedia Appendix 1 [16].

### Participants and Recruitment

We used a purposeful sampling approach to recruit leaders in information technology, operational, and clinical roles within Roudebush VAMC and IHIE. All participants were involved with, or impacted by, VA-HIE implementation. Out of 16 invitations, a total of 12 (75%) leaders responded. In total, 9 individuals from the VA and 3 from the IHIE constituted these organizational categories: information technology (4), operations (6), and clinical care (2).

Specific roles in the information technology category included the VA's Chief Health Informatics Officer and systems engineers from the HIE community partner and the VA. Operations roles included the VA's Release of Information (ROI) Officer, Community Coordinator, Director of the Virtual Lifetime Electronic Records Program Office and the CEO of the community HIE partner, and the community partner's client services representative for the VA. Clinical roles consisted of the VA Chief of Pharmacy and the VA Chief of Ambulatory Services. Informed consent to participate in the study was obtained in writing from all participants.

### Data Collection

Interviews were conducted in person or by phone during the period of June 2014 to August 2016 by investigators DH and BD. Interviews lasted about 45 min (range of 22-62 min) and followed a semistructured format. The interview guide was developed by members of the research team, drawing upon knowledge of HIE, clinical practice, and implementation science.

Interviewers posed 3 main categories of questions regarding VA-HIE implementation: (1) how the informant was involved; (2) what the perceived value of the informant's organization was; and (3) what barriers and facilitators the informant perceived. (The interview guide, including specific questions and probes, is included in Multimedia Appendix 2.) After listening to initial responses, interviewers probed using open-ended questions and asked for specific examples. The Institutional Review Board at Indiana University and the VA Human Subjects Committee at Roudebush VAMC approved the project.

### Data Analysis

Interviews were audio recorded, transcribed verbatim, deidentified, and checked for accuracy. All investigators independently read each transcript and initially used open coding [17] to capture the essence of the interview. We developed a preliminary coding dictionary using language derived from the

data [18]. The dictionary subsequently expanded when new ideas were discovered or when previous codes required finer levels of distinction. Each transcript was coded by at least two investigators.

All investigators, along with a research assistant, met regularly to compare coding and discuss emerging topics. We approached the data inductively using the immersion-crystallization approach to understand each participant's unique perspective. Immersion-crystallization refers to investigators immersing themselves into the experiences described in the interview transcripts. Crystallization is the emergence of cohesive insights that capture the ideas expressed by multiple transcripts [19]. Data were managed using QSR NVivo 10 software. Analysis was conducted concurrently with data collection, as specified by grounded theory [20], and to gauge when data saturation had been reached.

## Results

### Overview

Overall, we found that leaders from every organizational area were enthusiastic about HIEs and that they were confident that the VA would eventually *get it right*. Given the changes in VA procedures to implement requirements in the Veterans Access, Choice and Accountability Act of 2014, all interviewees imagined a future when the sharing of medical records would be commonplace, accurate, and secure. Many were proud of the VA's long history of secure information sharing between VA health care organizations across the country. All leaders, including our community HIE partner, imagined that the need for interoperability between information systems (ISs) would continue to increase in the future. From among the leader interview responses, we describe observations across several phases of HIE adoption: recruitment and consent, training, organizational memory, implementation, and sustainability.

### Recruitment and Consent

Recruitment and consent were by far the most frequently discussed topics in our interviews at the VA. Clinical nurses and medical assistants made early efforts at recruitment; however, they had little time to offer detailed explanations of the program to veterans. According to operations leaders, veterans were concerned with how their records would be used, by whom they would be used, and if they would be secure. Eventually, clinical personnel redirected veterans to the ROI office, a division of Health Information Management (HIM) at the VAMC, to learn more about the program. Although clinical staff were relieved of the burden to explain information sharing, recruitment activities increased the administrative load for HIM staff.

All VA leaders with whom we spoke were sensitive to the burden placed on staff by the recruiting and consenting process. Clinical leaders, in particular, noted the impact upon workflow as potentially *disruptive to normal clinical activities* and described clinical personnel as already facing many demands on their time. Likewise, all VA leaders were cognizant of the need for veterans to *opt in* to release their medical information for VA-HIE sharing. In our interviews, operations leaders were

aware that the most successful information sharing systems outside of the VA used an *opt-out* approach, where medical patients were automatically enrolled unless they specifically decided not to release their information. In our study, community HIE leaders described the opt-in process as a barrier to greater enrollment.

The VA seeks to protect the medical information of its patients, and thus, veterans were required to expressly ask to join (opt in) the VA-HIE project and agree to the release of their medical records. Operations leaders agreed that the *opt-in* approach was restrictive but that *it would take an act of congress* (literally) to change. One HIM (operations) leader recalled that opting in was a *multistep process* that was *unexpectedly time intensive*. One full-time clerk was hired to process *consent* documents, which on a peak day could exceed 60 documents.

Several operations leaders also described a successful strategy wherein veterans recruited other veterans; specifically, one lead employee in the ROI office was himself or herself a veteran. This employee, called by leaders a *super recruiter*, interacted with veterans as a peer, understanding and addressing their concerns and quelling their suspicions about the use of their medical data. A leader in operations described the consistent success of veterans recruiting veterans:

> ...*I think being a Veteran and telling them that you're using [the VA-HIE system], I heard this chimed out through the other pilot sites; when some of the Veterans did the recruiting, there's like a camaraderie.*

As the enrollment and consent processes were time intensive, HIM (operations) leaders directed recruiting efforts toward patients who had been seen by outside physicians and were most likely to see the benefit of VA-HIE information sharing. One HIM leader put it this way:

> *We knew who we should target, who were going to have upcoming appointments, who were going to be seen with the non-VA care providers, so that we could...get the biggest bang for our buck.*

HIM (operations) leaders reported that their recruitment rate was among the highest in the country. They attributed this success to finding a process for enrollment that did not burden frontline clinical staff. In addition, allowing a super recruiter to establish a personal connection with each veteran was considered to be essential to the program's success by these leaders. The *high-touch* approach provided greater ease for veterans and thus higher enrollment.

### Training

Training for the launch of the VA-HIE was handled cooperatively by the community HIE and VA trainers:

- The HIE community partner taught group classes to introduce medical assistants, nurses, and physicians to the VA-HIE program. These personnel were shown how to access non-VA data from the VA's electronic health record (EHR) system. Community HIE operations leaders were pleased with the willingness of nurses, technicians, and

assistants to receive the training, which was offered in a classroom setting.

- The VA Community Coordinator (operations) provided one-on-one instruction to a number of staff members who were viewed by their peers as informal leaders. These leaders would help their groups adopt the VA-HIE program to access non-VA information during clinical visits.
- The VA information technology team members gave demonstrations and answered questions in the clinics and at service-level meetings. They created VA-specific brochures and posters to be displayed in the facility.

The intention, according to one clinical leader, was to reach a *critical mass* of trained users. When that mass was reached and personnel began to incorporate data retrieval into their workflows, then training would emerge organically. He said:

> *Often, [training] is between personnel in the clinic. So, once you get enough people, some of it will spread, hopefully.*

Although staff members were enthusiastic about the training, HIM (operations) leaders described some physicians, advanced practice nurses, and other clinical VA personnel as reluctant to participate in the training and to use the product. VA operations leaders did not aggressively promote training or push product use. They believed that the user interface and information retrieval processes were *still rough*. Leaders thought that promoting a rough product would lead to user disappointment. An operations leader described the concern this way:

> *If you get a system that doesn't work well...if the system doesn't work the first time, the doctors aren't gonna mess with it again. They might have their assistant do it, they might tell somebody to do it, but the doctors for the main part are not gonna take the time if it doesn't work right away....So, we were fairly hesitant to go and really push with the doctors.*

## Organizational Memory

Operations leaders described some pessimistic responses of clinical staff to the product launch. These leaders speculated that pessimism resulted from disappointing past experiences with other initiatives. They reported that when several software or web-based projects had been launched in the past few years, the promotion of those launches created high expectations. The projects then disappointed early adopters who found them not to be fully developed. One leader told us:

> *One of the things I have to say, feedback-wise...well, you're advertising something but it's not really going to work until about five years from now. I heard that like a broken record, that "we've been through this before" and it was hard to get people to help launch and use the product.*

The VA has launched many new information technologies and has also been a national leader in quality improvement. Nonetheless, health care providers had mixed experiences with previous roll-outs of health information technology, including early implementation of the personal health record. We found that recent experiences of individuals or even organizational

memory were influential on the implementation of the VA-HIE program.

## Implementation and Adoption

From the beginning of the VA-HIE project, leaders were aware that many groups, internal and external to the VAMC, were affected by its implementation. The integration of 2 EHR systems was a complex organizational task. At least two categories of implementation challenges arose: (1) technical (eg, interoperability) and (2) human (eg, coordination among multiple VA groups in different organizational units) challenges.

### *Technical Challenges*

Information technology leaders reported that no technical challenges arose that required significant effort or delayed the deployment schedule. One operations leader attributed the success of the technical integration to the experience level of the IHIE, stating:

> *[The IHIE] was so advanced, and had years of experience doing health exchange—we were really just another on-boarding process for them...*

Community HIE leaders also reported a favorable integration process. They described a strong and collegial relationship with VA personnel. Furthermore, because of the success of the technical integration, several community coordinators from other VA sites around the country visited Indianapolis to learn best practices for working with an HIE community partner.

VA operations leaders noted that new personnel championed the adoption of the VA-HIE project. Specifically, those who drove the initiative forward had 5 years or less experience at the VA. New personnel, especially nurses, showed enthusiasm about the project and sought ways to incorporate both recruiting and data extraction activities into their tasks. One leader told us:

> *It was kind of the younger ones trying to initiate things and push it forward, believing that it was something of value.*

Operations leaders noted that nurses and pharmacists were the clinical staff members who most often accessed the VA-HIE system and performed data extraction.

However, VA informatics (information technology) leaders reported some challenges with retrieving data following implementation. In most clinics, plans had not been made to incorporate VA-HIE information retrieval into routine care. Personnel were unsure how to proceed. Many physicians were reluctant to change the clinical workflow. One operations leader recounted a session where he assisted a clinician with data retrieval:

> *We'd go on [the system] and do a query and if [the system] sat there, we waited for it. [The clinician said], "I like the system, but I'm not going to sit here for five minutes. It didn't work once. It didn't work twice. I've got to go take care of patients."*

As a result of these early challenges, some HIM operations leaders thought that physicians might underestimate the power and features of the fully functional VA-HIE system. They further

speculated that physician adoption might have been higher had the system been fully operational and user friendly when it was launched.

### Human Challenges

Leaders described clinician uneasiness about relying upon records that were created elsewhere for the purposes of medical decision making. Unknown errors or omissions seemed possible within those outside records. As information sharing between VA and non-VA HIE systems was a relatively new practice, clinical leaders surmised that VA clinicians were concerned about their own liability if medical errors were introduced by inaccurate outside information.

Leaders affirmed the organizational complexity of implementing the VA health care system. They described several different organizational units, system levels, and procedures that were involved in the adoption of VA-HIE information sharing. Interviewees from different departments and settings provided details about their unique perspectives on the implementation process.

In the HIM department, we observed that the process of obtaining patient consent impacted workflows. In addition to explaining the program and answering questions, staff were required to scan the consent forms into the EHR. Medical record staff then recorded the consent in a second system, referred to as the Veteran Authorization Preference system, as well as in a local registry that tracked the number of consents at the facility. Reminders were established in the local registry to re-enroll patients when their current consents expired.

In ambulatory clinics, clinical leaders noted that recruitment tasks competed with existing clinical activities for clinic staff's time, including physicians, nurses, and practice managers. To streamline the workflow, frontline clinical staff members offered VA-HIE information exchange as a service to the veterans who were seen in the clinic. To retrieve medical records, nurses who had received training would use the VA-HIE system with support from the information technology group.

In the ROI office, veterans who visited were asked by operations staff if they wished to enroll in VA-HIE. The Community Coordinator, in concert with health informatics leadership, used several strategies to obtain the consent of veterans for HIE. Oftentimes, these approaches were opportunistic, such as approaching patients who were waiting in line to pick up prescriptions or receive a flu shot. Veterans could also access their records through the MyHealtheVet (MHV) patient portal and give consent through MHV.

From our observations, we derived a novel organizational framework (Figure 1). The figure captures the interconnected web of ISs and human teams and the range of staff, provider, and patient stakeholders who were necessary to integrate 2 robust enterprise health ISs. Although the precise configuration of data systems and teams may vary by project, the framework visualizes key groups that were necessary to successfully implement one VA-HIE partnership. This underlying structure can be applied by other national and regional networks that seek to achieve interoperability across health care delivery systems.

**Figure 1.** Groups and systems important for integration of VA health information exchange at one VA medical center. Access db: Access database; CPRS: Computerized Patient Record System; HIE: Health Information Exchange; HIM: Health Information Management; VA: US Department of Veterans Affairs; VAP: Veteran Authorization Preference registry.
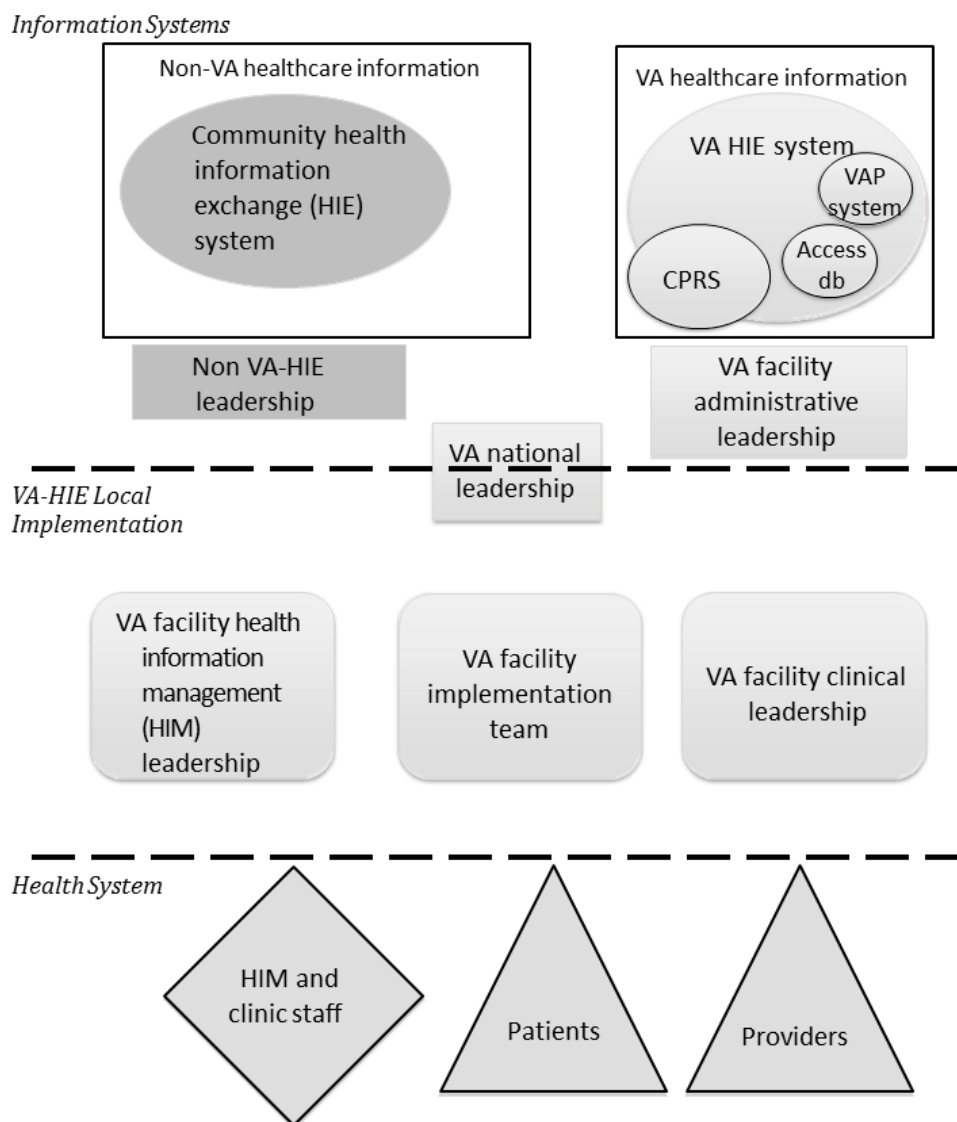


Groups and systems important to Department of Veterans Affairs-health information exchange integration at one VA medical center

In brief, our framework (Figure 1) places consumers of information and the HIM and clinical staff who were in direct contact with the data on the lower level. Information sharing (whether importing patients' records from outside the VA system or allowing them to be exported from the VA system) was mediated by the facility-based groups in the next layer. (This layer also includes the national offices of the Veteran's Health Administration who oversaw the VA-HIE demonstration project.) The top level consists of 2 regional systems that exchanged health care information: the VA facility and its community HIE partner. A detailed discussion of the framework is provided in Multimedia Appendix 3. This framework can help identify key organizational units likely to be involved in successful HIE implementation.

## Sustainability

The biggest sustainability challenge from the perspective of community HIE leaders was how to fund the continued integration of the VA and HIE systems. Most community HIEs charge membership fees to participating health organizations. The VA-HIE demonstration project was, by design, a time-limited initiative wherein community HIE fees were no longer supported by the VA after 2 years. Since then, VA leaders have permitted local HIE subscribers to continue to access VA medical records.

Despite financial concerns, community HIE leaders were in agreement about the value of maintaining interoperability between VA and non-VA ISs. They noted a strong, positive response by non-VA organizations to the availability of the VA medical records. One HIE operations leader told us:

*I think that even just the mention of the VA to our customers gets their attention—it's a source of data that people are interested in....So for us, it's a selling point of our HIE that we have that data....We are glad to have it and we want to continue it.*

Sustainability was perceived by VA clinical leaders to rest upon the added value of exchanged information in the clinic and upon the reliability of the information sharing process. In short, the VA-HIE product must perform in the way that it is intended and provide usable results. One operations leader stated:

*We need to have a product that's prime time ready. It's not because the information isn't there [now]. It is there but we need to have something that's clinical, that's usable in a clinical environment that's fast, that's structured appropriately.*

A VA information technology leader concurred:

*Obviously, it takes a long time—years, decades—whatever, to get to a point that...the system is mature enough to be able to make it work well across all the different data systems.*

## Discussion

### Principal Findings

Overall, VA and community HIE leaders believed in the promise of HIE to facilitate high-quality, coordinated care delivery but nonetheless identified several facilitators and barriers to HIE that will need to be addressed for successful, sustainable interoperability between VA and non-VA providers. Increases in veterans' access to non-VA medical care create a clear need for greater information sharing among VA and non-VA providers. Overall, leaders perceived a strong rationale for integration with non-VA providers to support care coordination and policies that encourage HIE. Organizational leaders also perceived a number of human factors issues that make VA-HIE implementation and workflow challenging from the perspective of frontline providers and staff. These findings provide lessons for other VA sites and other HIE programs throughout the United States and globally.

### Comparison With Previous Work

We observed significant support for the notion by VA and community HIE leaders that VA-HIE, again, is critical to delivering high-quality, coordinated care to veterans. Their support is similar to the findings of Byrne et al [15] who conducted health care provider interviews at other VA-HIE demonstration sites. In this previous study, 96% of veterans felt that VA-HIE would benefit veterans, whereas 71% of VA care providers who used VA-HIE reported positive changes in the care of their patients. Broad support for HIE is likely influenced by policies that create additional incentives to adopt VA-HIE. Community HIE leaders noted that the Veteran's Choice Act of 2014, which expanded access to non-VA medical care for veterans [21], brought greater demand from non-VA providers to access VA records for veteran patients that may be seen in their practice. Similarly, VA operations leaders noted that demand was also increasing in the other direction, that is, VA

providers needed information from non-VA medical records to deliver comprehensive care.

Broader access to non-VA providers will likely increase the chance for fragmentation of care [22] and strengthen the need for VA-HIE [23]. As fragmentation increases in health care delivery, both providers and patients may need to change their mental models from an ownership view of health data to a perspective view that emphasizes continuity of care [24]. To successfully integrate new external health information into clinical work practices, health care providers will need to recognize the potential value of health information from outside their own practice [25,26]. Future research should examine the impact of these policy changes, enabling access to non-VA care for veterans, on the use of VA-HIE and measurable impacts on veterans' health outcomes.

The original VA policy of intentional consent (opt in) was designed to protect patients and preserve the confidentiality of their records. In contrast, most state and regional health organizations assume patient consent; those patients must opt out to restrict sharing. Knowing this, the majority of leaders whom we interviewed believed that the requirement for veterans to explicitly opt in created the single biggest obstacle to patient adoption of the VA-HIE program. The merits and limitations of opting in versus opting out have been discussed extensively [27]. Previous studies have found that patients are concerned about the possible loss of privacy or misuse of their health data [28,29]. Moreover, patients are concerned about losing control over their health records when HIE systems are implemented [30]. Only a narrow majority (58%) of patients believe that the benefits of sharing health information outweigh the risks [31]. Although these trends are changing in the public sector [32], future research should investigate more fully the veteran response to interoperability.

According to leaders, disappointments around past VA information technology initiatives had a meaningful impact on this demonstration project. In the past, providers may have felt that new information technologies had imposed additional time burdens and were imperfectly executed. Although subjective emotional responses may not be foremost in the planning of informatics product releases, our findings reinforce that such considerations may be very important. Many investigators have described the importance of nontechnological elements in evaluating HIE adoption [33-36]. IS literature suggests that individuals' feelings about information technology impact their adoption decisions [37]. Furthermore, researchers have suggested that organizational culture may substantially influence both the implementation of new technologies and the continuance of old systems [38]. Health care workflow often is not driven by efficiency alone but by other considerations, such as individual preferences, or organizational and cultural factors that are important to individuals [39,40].

Clinicians were uneasy and concerned about their own liability if medical decisions were based on inaccurate non-VA information. Questions about the medical liability introduced by shared records have been widely discussed [41] and have highlighted the need for accurate matching of patient identities with patient records across systems. Although many of these

liability concerns have since been addressed (eg, Office of the National Coordinator for Health Information Technology data brief) [32], new issues have emerged. Local laws continue to create many barriers to information exchange [23]. In 2018, more than 2300 state statutes and regulations were associated with electronic health information and its sharing [42], and these laws have introduced many different barriers to the efficient sharing of health information.

In the VA, federal regulations supportive of interoperability have accompanied expansions in veteran access to community care; however, assembling a critical mass of veterans' non-VA health records has remained a formidable challenge. Physicians have had little information available about care received outside the VA because the reach of HIE among veterans has been low (4% have consented to data sharing in 2018). Thus, more than 90% of requests from community HIE partners for veteran health data are rejected because of lack of consent on files [26]. Low adoption of HIE is not unique to the VA; recent reviews show that HIEs remain underutilized, integration between systems is not fully developed, and many barriers remain [43-45].

## Current Developments at VA

Our data were collected during a VA-HIE demonstration project, which was implemented at 9 different VA sites and ended after 2 years. The Department of VA continues to make changes that impact information sharing. The Veteran's Choice Program was replaced in June 2019 by the Veterans Community Care Program, which is similar to Veteran's Choice but is attached to the VA Maintaining Internal Systems and Strengthening Integrated Outside Networks (MISSION) Act. The VA MISSION Act offers veterans new options for health care by further expanding their eligibility requirements to receive care in the community and increasing resources for in-home care providers and tribal health programs [46]. Most relevant to VA's information sharing efforts, the act also provides authority to automatically release veteran EHRs to non-VA care providers [47], and thus, veterans no longer need to opt in to share their records. We anticipate that concerns about patient confidentiality may re-emerge in response to the act's opt-out approach.

Overall, the MISSION Act contains many provisions for improvements in access to health care among veterans. High-quality, coordinated care can be best achieved along with the implementation of an infrastructure of interoperable human-centered technology. The need for an infrastructure that allows access to medical records between organizations—whether public, private, or governmental—has become a national priority [45]. During the COVID pandemic, social isolation has made information sharing even more important, as many clinics and hospitals conduct appointments virtually. Continued examination of both the technical and human challenges of information sharing remains important and timely.

Finally, the VA is currently implementing a new EHR system, a process that will require a decade to fully roll out across all VA facilities. Updating to a modern commercial EHR system will allow VA to make further progress toward its interoperability and HIE goals [48]. The commercial system being implemented is a founding member of the CommonWell Health Alliance, which seeks to establish a nationwide infrastructure for health data exchange [49]. As VA moves forward with its national EHR modernization plan, it should examine the HIE functions afforded through the platform and study how those functions might complement the existing VHIE program. Clinical workflows and technical components differ between the current VA EHR environment and the infrastructure afforded by CommonWell. Understanding these differences and developing approaches that maximize efficiencies for providers and HIM personnel will be critical not only for the success of HIE within the VA but also for the success of its new EHR platform.

## Limitations

We note a few limitations to our study. First, some readers may consider the use of qualitative methods to be a limitation. However, one benefit of a qualitative approach is the real-time capture of dynamic conversations. Interviewers covered a range of topics and were able to dig deeply when warranted. Such fluid revelations of information cannot be achieved when using a static questionnaire. Second, the study was conducted with leaders at one VAMC and its HIE community partner. The IHIE is well established and robust, and thus, the experiences of other VA and HIE leaders may not replicate the partnership described here. Third, each medical center has its own history and culture. VA leaders in our study noted that experience with past VA initiatives might have influenced participation in the VA-HIE project. Other medical centers, having their own history, may respond differently.

## Conclusions

The VA-HIE demonstration project showed how the integration of data across complex networks could be implemented. Leaders at one VAMC and its community partner HIE described the importance of information sharing and its value in providing high-quality patient care. Further discussion with them revealed the daunting levels of complexity VA personnel navigated to send and retrieve information. These VA and HIE leaders discussed the time-intensive process of asking patients to share their medical records. This VA found success in having veterans speak to other veterans, yielding the highest recruitment levels in the country.

Our interviews revealed that the technical compatibility between the 2 systems is not necessarily the major management challenge; rather, it is the coordination of the complex interrelationships among entities within the local and national VA. The synthesis of the observations from organizational leadership responsible for HIE implementation and stakeholders impacted by HIE adoption led us to create a new organizational framework to describe and visualize those relationships. The lessons learned advance implementation science and can be applied by other national and regional networks that seek to achieve interoperability goals across health care delivery systems.

## Acknowledgments

## Authors' Contributions

Study concept and design was done by DH and BD. Acquisition of data was carried out by DH and BD. Analysis and interpretation of data, drafting of the manuscript, and its critical revision for important intellectual content were done by CL, DH, and BD. DH obtained the funding and supervised the study.

## Conflicts of Interest

None declared.

Multimedia Appendix 1
Details of organizational setting and implementation.
[DOCX File , 73 KB - medinform_v9i2e19249_app1.docx ]

Multimedia Appendix 2
Interview guide for Indianapolis VA-HIE evaluation.
[DOCX File , 24 KB - medinform_v9i2e19249_app2.docx ]

Multimedia Appendix 3
Models of entities required to integrate two information systems.
[DOCX File , 194 KB - medinform_v9i2e19249_app3.docx ]

## References

1.    Dixon B. What is health information exchange? In: Health Information Exchange. Health Information Exchange: Navigating and Managing a Network of Health Information Systems Academic Press; 2016:3-20.
2.    Vest JR, Harle CA, Schleyer T, Dixon BE, Grannis SJ, Halverson PK, et al. Getting from here to there: health IT needs for population health. Am J Manag Care 2016 Dec;22(12):827-829 [FREE Full text] [Medline: 27982666]
3.    Holmgren AJ, Patel V, Adler-Milstein J. Progress In Interoperability: Measuring US Hospitals' Engagement In Sharing Patient Data. Health Aff (Millwood) 2017 Oct 01;36(10):1820-1827. [doi: 10.1377/hlthaff.2017.0546] [Medline: 28971929]
4.    Jamoom EW, Yang N. State Variation in Electronic Sharing of Information in Physician Offices: United States, 2015. NCHS Data Brief 2016 Oct(261):1-8. [Medline: 27805548]
5.    Veterans Health Information (VHIE), VHIE Overview. Department of Veterans Affairs. 2020 Dec 1. URL: https://www.va.gov/vhie/ [accessed 2021-01-09]
6.    What is ehealth exchange? eHealth Exchange; 2020. 2020. URL: https://ehealthexchange.org/ [accessed 2021-01-09]
7.    Bouhaddou O, Bennett J, Teal J, Pugh M, Sands M, Fontaine F, et al. Toward a Virtual Lifetime Electronic Record: the Department of Veterans Affairs experience with the Nationwide Health Information Network. AMIA Annu Symp Proc 2012;2012:51-60 [FREE Full text] [Medline: 23304272]
8.    Bouhaddou O, Cromwell T, Davis M, Maulden S, Hsing N, Carlson D, et al. Translating standards into practice: experience and lessons learned at the Department of Veterans Affairs. J Biomed Inform 2012 Aug;45(4):813-823 [FREE Full text] [doi: 10.1016/j.jbi.2012.01.003] [Medline: 22285982]
9.    Lyle J, Bouhaddou O, Botts N, Swall M, Pan E, Cullen T, et al. Veterans Health Administration Experience with Data Quality Surveillance of Continuity of Care Documents: Interoperability Challenges for eHealth Exchange Participants. AMIA Annu Symp Proc 2015;2015:870-879 [FREE Full text] [Medline: 26958223]
10.   Botts N, Bouhaddou O, Bennett J, Pan E, Byrne C, Mercincavage L, et al. Data Quality and Interoperability Challenges for eHealth Exchange Participants: Observations from the Department of Veterans Affairs' Virtual Lifetime Electronic Record Health Pilot Phase. AMIA Annu Symp Proc 2014;2014:307-314 [FREE Full text] [Medline: 25954333]
11.   Bouhaddou O, Bennett J, Cromwell T, Nixon G, Teal J, Davis M, et al. The Department of Veterans Affairs, Department of Defense, and Kaiser Permanente Nationwide Health Information Network exchange in San Diego: patient selection, consent, and identity matching. AMIA Annu Symp Proc 2011;2011:135-143 [FREE Full text] [Medline: 22195064]
12.   Nguyen KA, Haggstrom DA, Ofner S, Perkins SM, French DD, Myers LJ, et al. Medication Use among Veterans across Health Care Systems. Appl Clin Inform 2017 Dec 08;8(1):235-249 [FREE Full text] [doi: 10.4338/ACI-2016-10-RA-0184] [Medline: 28271121]

13. Dixon BE, Ofner S, Perkins SM, Myers LJ, Rosenman MB, Zillich AJ, et al. Which veterans enroll in a VA health information exchange program? J Am Med Inform Assoc 2017 Jan 06;24(1):96-105 [FREE Full text] [doi: 10.1093/jamia/ocw058] [Medline: 27274014]

14. French DD, Dixon BE, Perkins SM, Myers LJ, Weiner M, Zillich AJ, et al. Short-Term Medical Costs of a VHA Health Information Exchange: A CHEERS-Compliant Article. Medicine (Baltimore) 2016 Jan;95(2):e2481 [FREE Full text] [doi: 10.1097/MD.0000000000002481] [Medline: 26765453]

15. Byrne CM, Mercincavage LM, Bouhaddou O, Bennett JR, Pan EC, Botts NE, et al. The Department of Veterans Affairs' (VA) implementation of the Virtual Lifetime Electronic Record (VLER): findings and lessons learned from Health Information Exchange at 12 sites. Int J Med Inform 2014 Aug;83(8):537-547. [doi: 10.1016/j.ijmedinf.2014.04.005] [Medline: 24845146]

16. Overhage JM. Case Study 1 - The Indiana Health Information Exchange. In: Dixon BE, editor. Health Information Exchange: Navigating and Managing a Network of Health Information Systems. Cambridge, Massachusetts: Academic Press; 2016:267-279.

17. Corbin J, Strauss A. Basics of Qualitative Research : Techniques and Procedures for Developing Grounded Theory. In: Qualitative Research. Thousand Oaks, California, United States: SAGE Publications Ltd; Dec 2014:1-456.

18. Handbook of Current Developments in Grounded Theory, Second Edition. Thousand Oaks, California, United States: SAGE Publications Ltd; 2019.

19. Borkan J. Immersion/Crystallization. In: Crabtree BF, Miller WL, editors. Doing Qualitative Research. 2nd ed: SAGE Publishing; 1999:179-194.

20. Thornberg R, Charmaz K. Grounded theory and theoretical coding. In: Sage Handbook of Qualitative Data Analysis. 1st ed: Sage; 2013:153-169.

21. Dixon BE, Haggstrom DA, Weiner M. Implications for informatics given expanding access to care for Veterans and other populations. J Am Med Inform Assoc 2015 Jul 31;22(4):917-920. [doi: 10.1093/jamia/ocv019] [Medline: 25833394]

22. Bourgeois FC, Olson KL, Mandl KD. Patients treated at multiple acute health care facilities: quantifying information fragmentation. Arch Intern Med 2010 Dec 13;170(22):1989-1995. [doi: 10.1001/archinternmed.2010.439] [Medline: 21149756]

23. Adjerid I, Acquisti A, Telang R, Padman R, Adler-Milstein J. The Impact of Privacy Regulation and Technology Incentives: The Case of Health Information Exchanges. Management Science 2016 Apr;62(4):1042-1063. [doi: 10.1287/mnsc.2015.2194]

24. Cebul RD, Rebitzer JB, Taylor LJ, Votruba ME. Organizational fragmentation and care quality in the U.S healthcare system. J Econ Perspect 2008;22(4):93-113. [doi: 10.1257/jep.22.4.93] [Medline: 19791306]

25. Unertl KM, Johnson KB, Gadd CS, Lorenzi NM. Bridging organizational divides in health care: an ecological view of health information exchange. JMIR Med Inform 2013 Oct 29;1(1):e3 [FREE Full text] [doi: 10.2196/medinform.2510] [Medline: 25600166]

26. Donahue M, Bouhaddou O, Hsing N, Turner T, Crandall G, Nelson J, et al. Veterans Health Information Exchange: Successes and Challenges of Nationwide Interoperability. AMIA Annu Symp Proc 2018;2018:385-394 [FREE Full text] [Medline: 30815078]

27. Esmaeilzadeh P, Sambasivan M. Patients' support for health information exchange: a literature review and classification of key factors. BMC Med Inform Decis Mak 2017 Apr 04;17(1):33 [FREE Full text] [doi: 10.1186/s12911-017-0436-2] [Medline: 28376785]

28. Simon SR, Evans JS, Benjamin A, Delano D, Bates DW. Patients' attitudes toward electronic health information exchange: qualitative study. J Med Internet Res 2009;11(3):e30 [FREE Full text] [doi: 10.2196/jmir.1164] [Medline: 19674960]

29. McGraw D, Dempsey JX, Harris L, Goldman J. Privacy as an enabler, not an impediment: building trust into health information exchange. Health Aff (Millwood) 2009;28(2):416-427 [FREE Full text] [doi: 10.1377/hlthaff.28.2.416] [Medline: 19275998]

30. Kim K, Joseph J, Ohno-Machado L. Comparison of consumers' views on electronic data sharing for healthcare and research. J Am Med Inform Assoc Jul 2015;22(4):821-830. [doi: 10.1093/jamia/ocv014] [Medline: 25829461]

31. Perera G, Holbrook A, Thabane L, Foster G, Willison DJ. Views on health information sharing and privacy from primary care practices using electronic medical records. Int J Med Inform 2011 Feb;80(2):94-101. [doi: 10.1016/j.ijmedinf.2010.11.005] [Medline: 21167771]

32. Patel V, Hughes P, Barker W, Moon L. Trends in Individuals? Perceptions regarding Privacy and Security of Medical Records and Exchange of Health Information. ONC Data Brief, no.33. Office of the National Coordinator for Health Information Technology: Washington DC February 2016:2012-2014.

33. Adler-Milstein J, Jha AK. Health information exchange among U.S. hospitals: who's in, who's out, and why? Healthc (Amst) 2014 Mar;2(1):26-32. [doi: 10.1016/j.hjdsi.2013.12.005] [Medline: 26250086]

34. Walker J, Pan E, Johnston D, Adler-Milstein J, Bates DW, Middleton B. The value of health care information exchange and interoperability. Health Aff (Millwood) 2005 Jan;Suppl Web Exclusives:W5-10. [doi: 10.1377/hlthaff.w5.10] [Medline: 15659453]

35. Esmaeilzadeh P, Sambasivan M. Health Information Exchange (HIE): A literature review, assimilation pattern and a proposed classification for a new policy approach. J Biomed Inform 2016 Dec;64:74-86 [FREE Full text] [doi: 10.1016/j.jbi.2016.09.011] [Medline: 27645322]

36.    Vest JR. More than just a question of technology: factors related to hospitals' adoption and implementation of health
        information exchange. Int J Med Inform 2010 Dec;79(12):797-806. [doi: 10.1016/j.ijmedinf.2010.09.003] [Medline:
        20889370]

37.    Hu X, Lin Z, Whinston AB, Zhang H. Hope or Hype: On the Viability of Escrow Services as Trusted Third Parties in
        Online Auction Environments. Information Systems Research 2004 Sep;15(3):236-249. [doi: 10.1287/isre.1040.0027]

38.    Kim H, Chan H, Chan Y, Gupta S. Understanding the balanced effects of belief and feeling on information systems
        continuance. In: Proceedings of the 2004 International Conference on Information Systems. 2004 Presented at: 2004
        International Conference on Information Systems; 2004; Washington, DC, USA.

39.    Lehmann C, Unertl K, Rioth M, Lorenzi N. Change Management for the Successful Adoption of Clinical Information
        Systems. In: Finnell JT, Dixon BE, editors. Clinical Informactics Study Guide. Switzerland: Springer; 2016:435-446.

40.    Ozkaynak M, Unertl K, Johnson S, Brixey J, Haque S. Clinical Workflow Analysis, Process Redesign,Quality Improvement.
        In: Finnell JT, Dixon BE, editors. Clinical Informatics Study Guide. Switzerland: Springer; 2016:135-161.

41.    Dullabh P, Molfina M. Liability Coverage for Regional Health Information Organizations: Lessons from the AHRQ-Funded
        State and Regional Demonstration Projects in Health Information Technology and Other Community Efforts. Prepared by
        the National Opinion Research Center, AHRQ National Resource Center for Health Information Technology. In: AHRQ
        Publication No. 09-0071-EF. Rockville, Maryland: Agency for Healthcare Research and Quality; 2009.

42.    Schmit C, Sunshine G, Pepin D, Ramanathan T, Menon A, Penn M. Transitioning From Paper to Digital: State Statutory
        and Regulatory Frameworks for Health Information Technology. Public Health Rep 2017;132(5):585-592 [FREE Full text]
        [doi: 10.1177/0033354917722994] [Medline: 28813620]

43.    Furukawa MF, King J, Patel V, Hsiao C, Adler-Milstein J, Jha AK. Despite substantial progress In EHR adoption, health
        information exchange and patient engagement remain low in office settings. Health Aff (Millwood) 2014
        Sep;33(9):1672-1679. [doi: 10.1377/hlthaff.2014.0445] [Medline: 25104827]

44.    Vest JR, Kash BA. Differing Strategies to Meet Information-Sharing Needs: Publicly Supported Community Health
        Information Exchanges Versus Health Systems' Enterprise Health Information Exchanges. Milbank Q 2016 Mar;94(1):77-108
        [FREE Full text] [doi: 10.1111/1468-0009.12180] [Medline: 26994710]

45.    Interoperability. Office of the National Coordinator for Health Information Technology. 2019. URL: https://www.healthit.gov/
        topic/interoperability

46.    VA launches new health care options under MISSION Act. S Dept of Veterans Affairs, Office of Public Affairs & Media
        Relations, News Release. 2019. URL: https://www.va.gov/opa/pressrel/pressrelease.cfm?id=5264 [accessed 2020-01-09]

47.    VHIE Sharing Options. US Dept of Veterans Affairs, Health Care, Veterans Health Information Exchange (VHIE). 2019.
        URL: https://www.va.gov/VHIE/VHIE_Sharing_Options.asp [accessed 2020-06-06]

48.    EHR Modernization. Department of Veterans Affairs. URL: https://www.ehrm.va.gov/ [accessed 2021-01-13]

49.    CommonWell Health Alliance. URL: https://www.cerner.com/commonwell-health-alliance [accessed 2021-01-13]

## Abbreviations

**EHR:** electronic health record
**HIE:** health information exchange
**HIM:** Health Information Management
**IHIE:** Indiana Health Information Exchange
**IS:** information system
**MHV:** MyHealtheVet
**MISSION:** Maintaining Internal Systems and Strengthening Integrated Outside Networks
**ROI:** release of information
**VA:** veterans affairs
**VAMC:** Veterans Affairs Medical Center
**VHIE:** Veterans Health Information Exchange

Original Paper

# Prediction of Prolonged Length of Hospital Stay After Cancer Surgery Using Machine Learning on Electronic Health Records: Retrospective Cross-sectional Study

Yong-Yeon Jo[1], PhD; JaiHong Han[2], MD; Hyun Woo Park[1], PhD; Hyojung Jung[1], BS; Jae Dong Lee[1], MS; Jipmin Jung[3], MS; Hyo Soung Cha[3], PhD; Dae Kyung Sohn[4], MD, PhD; Yul Hwangbo[1], MD, PhD

[1]Healthcare AI Team, National Cancer Center, Goyang, Republic of Korea
[2]Department of Surgery, National Cancer Center, Goyang, Republic of Korea
[3]Cancer Data Center, National Cancer Control Institute, National Cancer Center, Goyang, Republic of Korea
[4]Center for Colorectal Cancer, Research Institute and Hospital, National Cancer Center, Goyang, Republic of Korea

**Corresponding Author:**
Yul Hwangbo, MD, PhD
Healthcare AI Team
National Cancer Center
323 Ilsan-ro
Ilsandong-gu
Goyang, 10408
Republic of Korea
Phone: 82 10 8885 2812
Email: yulhwangbo@ncc.re.kr

## Abstract

**Background:**   Postoperative length of stay is a key indicator in the management of medical resources and an indirect predictor of the incidence of surgical complications and the degree of recovery of the patient after cancer surgery. Recently, machine learning has been used to predict complex medical outcomes, such as prolonged length of hospital stay, using extensive medical information.

**Objective:**   The objective of this study was to develop a prediction model for prolonged length of stay after cancer surgery using a machine learning approach.

**Methods:**   In our retrospective study, electronic health records (EHRs) from 42,751 patients who underwent primary surgery for 17 types of cancer between January 1, 2000, and December 31, 2017, were sourced from a single cancer center. The EHRs included numerous variables such as surgical factors, cancer factors, underlying diseases, functional laboratory assessments, general assessments, medications, and social factors. To predict prolonged length of stay after cancer surgery, we employed extreme gradient boosting classifier, multilayer perceptron, and logistic regression models. Prolonged postoperative length of stay for cancer was defined as bed-days of the group of patients who accounted for the top 50% of the distribution of bed-days by cancer type.

**Results:**   In the prediction of prolonged length of stay after cancer surgery, extreme gradient boosting classifier models demonstrated excellent performance for kidney and bladder cancer surgeries (area under the receiver operating characteristic curve [AUC] >0.85). A moderate performance (AUC 0.70-0.85) was observed for stomach, breast, colon, thyroid, prostate, cervix uteri, corpus uteri, and oral cancers. For stomach, breast, colon, thyroid, and lung cancers, with more than 4000 cases each, the extreme gradient boosting classifier model showed slightly better performance than the logistic regression model, although the logistic regression model also performed adequately. We identified risk variables for the prediction of prolonged postoperative length of stay for each type of cancer, and the importance of the variables differed depending on the cancer type. After we added operative time to the models trained on preoperative factors, the models generally outperformed the corresponding models using only preoperative variables.

**Conclusions:**   A machine learning approach using EHRs may improve the prediction of prolonged length of hospital stay after primary cancer surgery. This algorithm may help to provide a more effective allocation of medical resources in cancer surgery.

XSL•FO
RenderX

**KEYWORDS**

postoperative length of stay; cancer surgery; machine learning; electronic health records

## Introduction

Cancer is a major burden on public health worldwide [1], and the amount of health care resources associated with its treatment is constantly increasing [2]. The major strategies of cancer treatment include surgery, chemotherapy, and radiation therapy, with surgery being the most common treatment approach. Compared with other cancer-related management strategies, cancer surgery requires greater use of health care resources and, consequently, greater medical costs [3]. Postoperative length of stay (POLOS) in the hospital is one of the reasons for the cost increase. As patients with cancer are discharged after full recovery from surgery, POLOS is also an indirect indicator of surgical recovery and postoperative complications in patients with cancer. That is, a prolonged POLOS (PPOLOS) indicates a delayed recovery after cancer surgery.

In previous literature, factors associated with PPOLOS have been evaluated for several cancer surgeries [4-8], and risk factors such as age, malnutrition, underlying diseases (including diabetes, cardiovascular diseases, renal dysfunction, and respiratory disease), and common blood count results (such as neutrophil-lymphocyte ratio, albumin, and hemoglobin) have been reported. However, the majority of studies have used a small number of subjects and have not evaluated a wide variety of clinical factors. Thus, there are many limitations to clinical application of the results of those studies.

Currently, most medical institutions store electronic health records (EHRs) and use them to improve the quality and efficiency of hospitals [9,10]. Many recent studies using EHRs have reported that machine learning–based models outperform statistical models in predicting outcomes and adverse events [11,12].

In this study, we assessed whether PPOLOS of patients with cancer can be predicted with machine learning approaches using EHR data and evaluated the effect of preoperative factors on the prediction of PPOLOS for each type of cancer.

## Methods

### Data Source and Subjects

Our retrospective study was conducted using EHR data from the Korea Cancer Big Data Platform (K-CBP), which was constructed in the National Cancer Center, Goyang, Republic of Korea. Details of the K-CBP have been described elsewhere [13]. Briefly, the K-CBP is a multidatabase framework that contains various medical information including clinical and genomic data and medical images. In this study, de-identified clinical data obtained from patients with cancer who visited the National Cancer Center were used. We extracted data from the K-CBP from 61,743 subjects with 19 cancer types who underwent primary cancer surgery between January 1, 2000, and December 31, 2017. The inclusion criteria for patients were as follows: (1) age ≥18 years, (2) surgery performed with general anesthesia, and (3) first instance of surgery for primary cancer. We excluded subjects who had emergency cancer surgery, cancer removal with local anesthesia, surgery for multiple primary cancers, or missing or typo-filled records for surgery, pathology, and hospitalization. Cancer types with fewer than 100 total cases were also excluded. Finally, we retrieved data from 42,751 subjects with 17 cancer types, including lip, oral cavity, and pharynx (oral; International Classification of Diseases codes C00-C14); esophagus (C15); stomach (C16); colon and rectum (colon; C18-C20); liver (C22); gallbladder and biliary tract (gallbladder; C23 and C24); pancreas (C25); larynx (C32); lung (C33 and C34); breast (C50); cervix uteri (C53); corpus uteri (C54); ovary (C56); prostate (C61); kidney (C64); bladder (C67); and thyroid (C73).

### Variables from EHRs

We examined several variables from diverse categories within EHRs, such as records of surgeries, blood tests, and medications, as well as pathologic reports and nursing charts. We only used medical data recorded within 6 months prior to surgery. For data on underlying diseases, only preoperative evaluation data were used. In the case of repeated data such as blood and biochemical tests, only the data recorded just before surgery was used in the analysis. For simplicity of interpretation, we reorganized variables into five major categories as follows: (1) surgical and cancer factors, (2) underlying diseases and functional laboratory assessments, (3) general assessments, (4) medications, and (5) social factors. Each major category consisted of one to five subcategories. Details of variables are described in Table 1. There were two types of missing values in our variables: missing numeric values were replaced by the middle value, and missing categorical values were replaced with "value unknown." We conducted the min-max normalization for obtained whole variables. It should be noted that we basically used preoperative variables for the prediction of PPOLOS. However, tumor staging represented by T/N stage—based on the TNM staging system of the American Joint Committee on Cancer [14,15]—was extracted from pathologic reports because of the lack of structured T/N stage information in preoperative images. In addition, we obtained the operative time as a typical intraoperative factor and analyzed its effect on the prediction of PPOLOS.

**Table 1.** Preoperative clinical variables.

| Major category and subcategory | Variables |
|---|---|
| **Surgical and cancer factors** | |
| Surgery | • Types of surgery<br>• Co-operations<br>• Surgeons |
| Cancer stage | • T/N stages |
| **Underlying diseases and functional laboratory assessments** | |
| Underlying diseases and related laboratory parameters | • Liver diseases (history of liver disease, hepatitis viral tests, aspartate aminotransferase, alanine aminotransferase, bilirubin, alkaline phosphatase, gamma-glutamyl transferase)<br>• Diabetes mellitus (history of diabetes, $HbA_{1c}$[a], glucose, urine glucose)<br>• Renal disease (history of renal disease, BUN[b], creatinine)<br>• Cardiac disease (history of cardiac disease)<br>• Hypertension (history of hypertension)<br>• Allergic disease (history of allergic disease)<br>• Tuberculosis (history of tuberculosis)<br>• Cancer (history of cancer)<br>• Mental disorder (history of mental disorder) |
| Cardiopulmonary functions | • Pulmonary function (FVC[c], $FEV_1$[d])<br>• Cardiac function (EF[e], E/A[f], RVSP[g]) |
| Nutritional factors | • Degree of appetite<br>• Albumin, globulin, A/G[h] ratio, protein<br>• Cholesterol (total, LDL[i], HDL[j], triglyceride)<br>• Lymphocyte count |
| Inflammatory factors | • hs-CRP[k], ESR[l], fibrinogen |
| Initial laboratory parameters | • Blood count (except lymphocyte count)<br>• Electrolytes, chemistry tests<br>• Urinalysis<br>• Coagulation tests<br>• Hormone tests<br>• ABO blood type |
| **General assessments** | |
| Demographic characteristics and anthropometric factors | • Age<br>• Sex<br>• Height, weight, BMI<br>• Ambulation, ECOG[m] performance<br>• Type of admission<br>• History of previous operation<br>• Family history of diseases<br>• Degree of diseases insight |
| Vital signs | • Blood pressure (systolic, diastolic)<br>• Body temperature<br>• Breath rate<br>• Pulse rate |
| Substance exposure | • Alcohol<br>• Smoking<br>• Alternative therapy |

| Major category and subcategory | Variables |
|---|---|
|     Symptoms | • Gastroenteric, cardiovascular, respiratory, neurologic, dermatologic, and urinary symptoms<br>• Sleep and fatigue<br>• Mood<br>• Pain |
| **Medications** | |
|     Drugs | • Medications |
| **Social factors** | |
|     Family | • Marriage, child, cohabitation |
|     Education | • Level of education |
|     Religion | • Type of religion |
|     Occupation | • Type of job |

[a]$HbA_{1c}$: hemoglobin $A_{1c}$.

[b]BUN: blood urea nitrogen.

[c]FVC: forced vital capacity.

[d]$FEV_1$: forced expiratory volume in the first second of expiration.

[e]EF: ejection fraction.

[f]E/A: ratio of the early (E) to late (A) ventricular filling velocities.

[g]RVSP: right ventricular systolic pressure.

[h]A/G ratio: albumin to globulin ratio.

[i]LDL: low-density lipoprotein.

[j]HDL: high-density lipoprotein.

[k]hs-CRP: high-sensitivity C-reactive protein.

[l]ESR: erythrocyte sedimentation rate.

[m]ECOG: Eastern Cooperative Oncology Group.

### Definition of PPOLOS

In the literature, PPOLOS is defined in a variety of ways [7,8,16,17]. This study focused on predicting which patients with cancer will use a significant amount of hospital resources. Therefore, the PPOLOS study group was defined as the subset of patients who used 50% of the total ward after surgery. Specifically, we calculated the total number of postoperative bed-days by considering the respective length of stay between surgery and discharge for patients with each type of cancer. Next, we arranged the patients by POLOS from shortest to longest. Then, we defined the long-term hospitalized patient group, which occupied half of the total hospital bed-days, as the PPOLOS group.

### Models

To predict PPOLOS, we employed three models: (1) extreme gradient boosting (XGB) classifier [18], (2) multilayer perceptron (MLP) [19], and (3) logistic regression (LR). XGB classifier is one of the most widely used machine learning algorithms. It is a high-performance classifier based on gradient boosting that trains decision trees in succession such that residuals of earlier trees are corrected by later ones. MLP is a type of feed-forward neural network in which all computation is directed from the input layer to the output layer. The model

is built on the architecture of at least three layers, with one input layer, variable hidden layers, and one output layer. Backpropagation is used to find optimal layer weights for the model [20]. LR is a commonly used classification algorithm to assign observations to a discrete set of classes. Unlike the majority of LR algorithms yielding continuous values, its outputs are converted by the sigmoid function into probabilities mapped to the classes. These models have been utilized in numerous medical and clinical studies to analyze EHRs, vital signals, and images, as well as to support medical decisions [21-24]. In our study, the MLP model consisted of a self-dot attention layer and two fully connected layers. We evaluated the performance of the model using 5-fold cross-validation. In each fold, training and test sets were divided in an 8:2 ratio.

## *Results*

### Ethics Statement

The research protocol was approved by the Institutional Review Board of the National Cancer Center (IRB No. NCC2018-0113). All data used in this retrospective study were de-identified.

### Characteristics of the Subjects

Multimedia Appendix 1 shows the characteristics of each cancer population. Stomach cancer surgery (n=8929) was the most

common surgery in this study, followed by breast (n=8918), colon (n=7449), thyroid (n=5071), lung (n=4455), and liver (n=1342) cancer surgeries. The average age of the patients was 56.6 years, and women accounted for 55.75% (23,835/42,751) of the total cancer cases. Oral (mean 22.2 days, SD 22.3 days, median 16.9 days), esophageal (mean 22.1 days, SD 22.5 days, median 15.8 days), gallbladder (mean 20.7 days, SD 14.8 days, median 16.9 days), and pancreatic (mean 21.0 days, SD 15.1 days, median 16.9 days) cancers were associated with relatively long POLOS, whereas thyroid (mean 3.3 days, SD 2.2 days, median 3.0 days) and breast (mean 5.4 days, SD 6.5 days, median 4.1 days) cancers were associated with relatively short POLOS. The respective PPOLOS thresholds and proportions of patients with PPOLOS for each cancer type were as follows: stomach (10 days; 2481/8929, 27.80%), breast (6 days; 2354/8918, 26.40%), colon (11 days; 2143/7449, 28.77%), thyroid (4 days; 781/5071, 15.40%), lung (12 days; 1195/4455, 26.28%), liver (15 days; 320/1342, 25.34%), prostate (9 days; 312/1054, 29.60%), ovary (18 days; 266/1016, 26.18%), kidney (9 days; 162/767, 21.12%), esophageal (24 days; 184/761, 24.18%), cervix uteri (16 days; 150/706, 21.25%), corpus uteri (12 days; 120/535, 22.43%), oral (27 days; 113/528, 21.40%), gallbladder (25 days; 127/499, 25.45%), pancreatic (23 days; 99/365, 27.12%), bladder (11 days; 35/233, 15.02%), and larynx (31 days; 24/123, 19.51%).

## Prediction Performance

[Multimedia Appendix 2](#) shows the performance of our models in predicting PPOLOS with four metrics: accuracy, specificity, sensitivity, and area under the receiver operating characteristic curve (AUC). When evaluating the AUC metrics for our XGB classifiers, the models performed excellently for kidney and bladder cancers (AUC >0.85). A moderate performance (AUC 0.70-0.85) was observed for stomach (AUC 0.83), breast (AUC 0.83), colon (AUC 0.71), thyroid (AUC 0.79), prostate (AUC 0.78), cervix uteri (AUC 0.78), corpus uteri (AUC 0.79), and oral (AUC 0.79) cancers. In contrast, the models had relatively low performance for lung, liver, ovary, esophageal, gallbladder, pancreatic, and larynx cancers (AUC <0.7).

Receiver operating characteristic (ROC) curves of major cancers are shown in [Figure 1](#). For cancers with fewer than 4000 cases, we found that classification performance did not vary significantly between the different models. However, for cancers with more than 4000 cases (stomach, breast, colon, thyroid, and lung cancers), the performance of XGB classifiers was superior to that of the other models. For the metric of sensitivity, which represents the prediction of cases with PPOLOS, MLP showed better performance than the other methods.

**Figure 1.** Receiver operating characteristic (ROC) curves of all models for major cancers. For each subfigure, a legend shows the average area under the ROCs with their standard deviations on 5-fold cross-validations for the models. Solid lines represent the plots of the average area under the ROCs, covering the plots of all area under the ROCs by its shaded regions. MLPAtt: multilayer perceptron with attention mechanism; XGB: extreme gradient boosting.
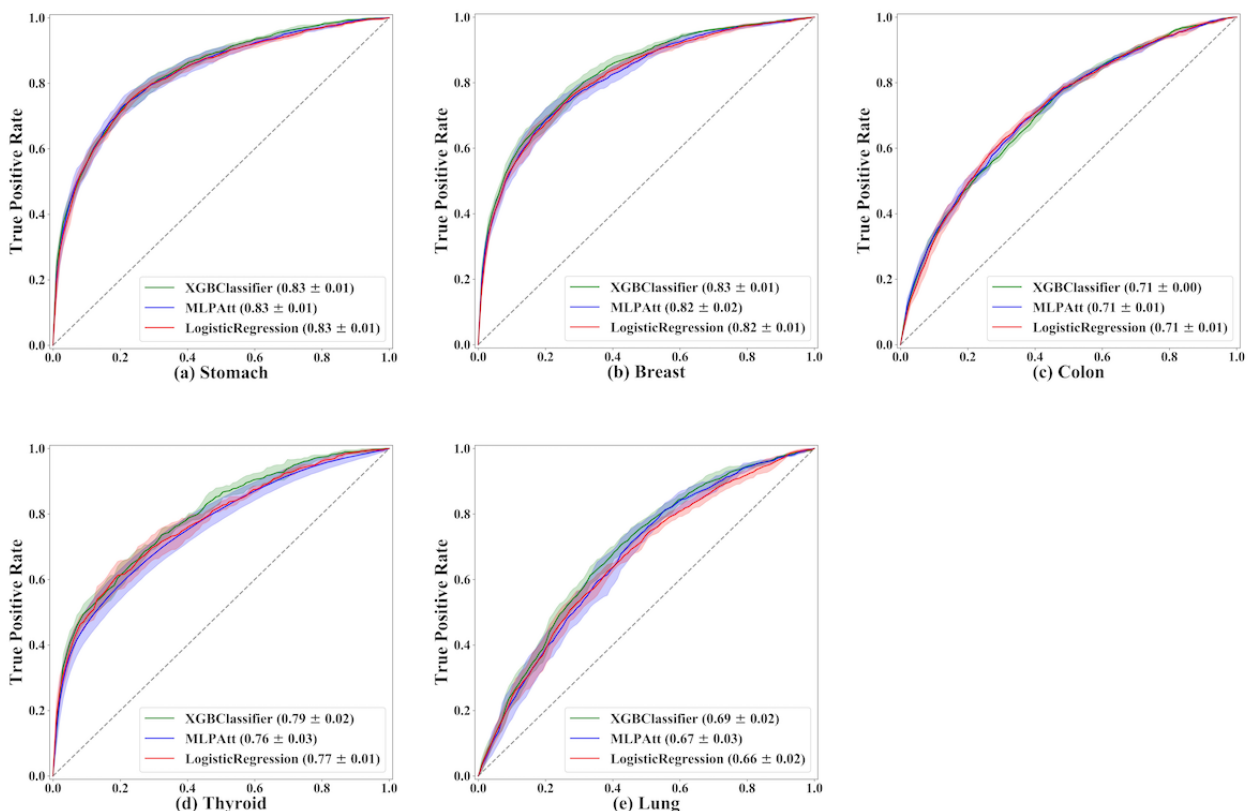


## Risk Factors of PPOLOS

We identified the most important variables for each model by examining the respective coefficient or attention score.

[Multimedia Appendix 3](#) shows the top 10 important variables from the models of the five cancers with the highest number of patients (stomach, breast, colon, thyroid, and lung cancers).

For each type of cancer, various risk factors were identified in the three models. The top 10 risk factors identified in the five cancers in the XGB classifier model were as follows:

- stomach cancer: albumin and globulin, urinary symptoms, surgeries (total gastrectomy and laparoscopy-assisted distal gastrectomy), forced expiratory volume in the first second of expiration, absolute neutrophil count, zolpidem use, and N stage;
- breast cancer: urinary symptoms, surgeries (modified radical mastectomy and breast-conserving surgery), surgeon, globulin, famotidine use, N stage, marriage, and metoclopramide use;
- colon cancer: surgeon, co-operation, albumin, surgeries (abdominoperineal resection and laparoscopic anterior resection), urinary symptoms, marriage, N stage, and urine white blood cell count;
- thyroid cancer: N stage, urinary symptoms, surgery (total thyroidectomy), albumin and globulin, ejection fraction, surgeon, drinking, and marriage; and

- lung cancer: albumin and globulin, sex, nonsmoker, absolute neutrophil count, theophylline use, route of admission, marriage, and hemoglobin.

No universal set of risk factors was present in subjects with PPOLOS, as the importance of a given variable was dependent on both the type of cancer and the model used.

## Contribution of the Variable Group to the PPOLOS Prediction

We plotted all variable scores derived from XGB classifier for nine types of cancer with the largest subject populations in Figure 2. In this figure, a bar represents the cumulative scores in a major category divided into colors corresponding to subcategories, with the sum of their cumulative scores equal to 1. We found that various variables contribute to the prediction of PPOLOS, which are different for each type of cancer.

**Figure 2.** Variable scores derived from extreme gradient boosting (XGB) classifier for the top nine cancers of the patient population. Each subfigure shows cumulative scores of all variables derived from XGB classifier for a single cancer, where each bar is segmented into subcategories by colors and represents the cumulative score of a major category.
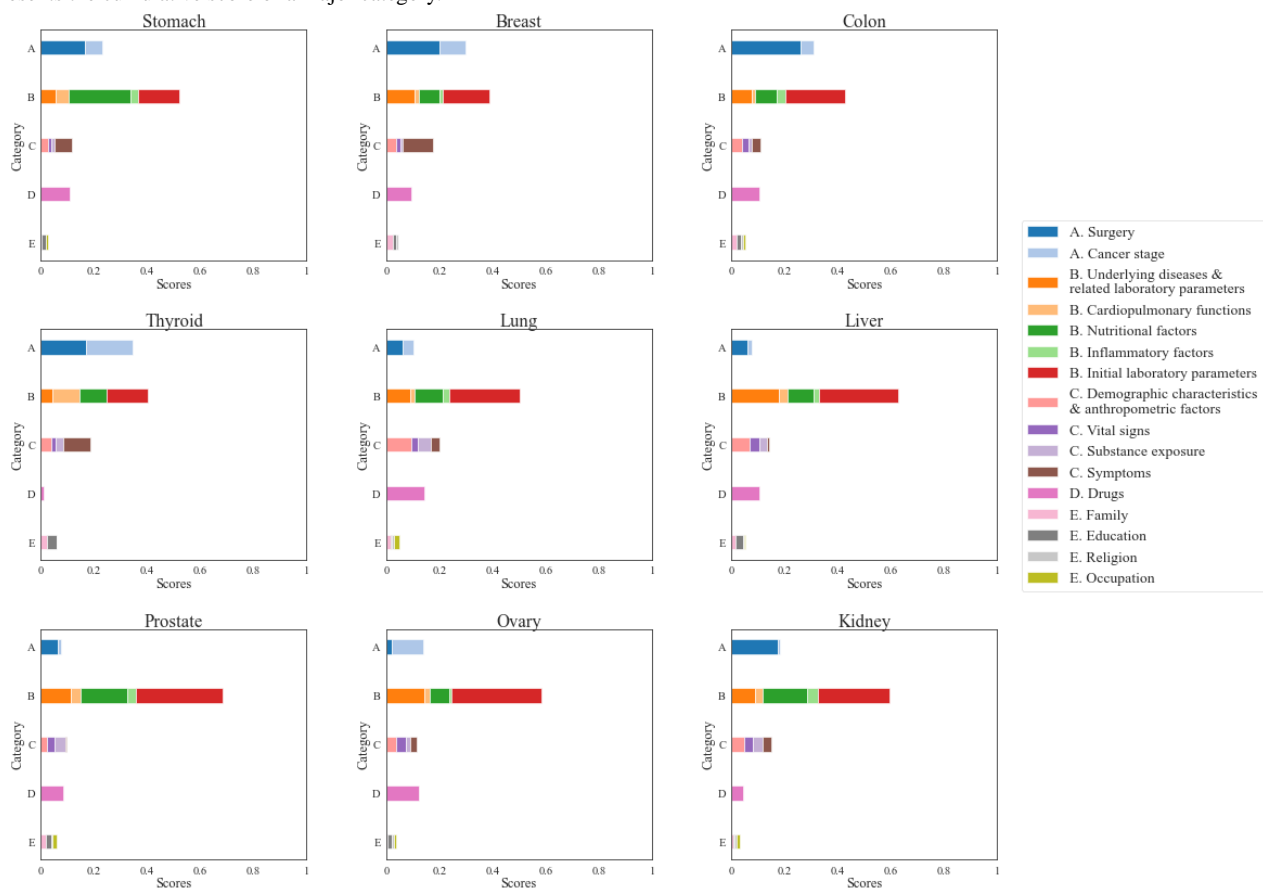


We note that variables in major category A (ie, surgical and cancer factors) that occupy more than 20% of the total proportion influence the prediction of PPOLOS for the top four cancers of the patient population (ie, stomach, breast, colon, and thyroid) more than for the other five cancers. Major category B (ie, underlying diseases and functional laboratory assessments) contains the most influential factors for all nine cancers.

## Impact of the Operative Time

To evaluate the impact of intraoperative factors on the PPOLOS prediction, we incorporated operative time, a representative indicator of surgery quality, to the models trained on preoperative factors. We evaluated changes in the classification performance of PPOLOS in the model including the operative time. The average AUC increased from 0.74 to 0.76 for all models. Figure 3 shows the prediction performance of XGB

classifiers. The yellow bar shows the AUC of the XGB classifier trained with only preoperative variables and the blue bar shows the AUC of the model trained with the operative time in addition to preoperative variables. The model trained with preoperative variables and operative time generally outperformed the models trained without operative time. For bladder and larynx cancer, adding operative time to the models had no benefit in predicting PPOLOS.

**Figure 3.** Performance of the extreme gradient boosting (XGB) classifier model for the prediction of prolonged postoperative length of stay by cancer type. The yellow bar indicates the model trained only using preoperative variables and the blue bar represents the model trained with operative time in addition to preoperative variables. AUC: area under the receiver operating characteristic curve.
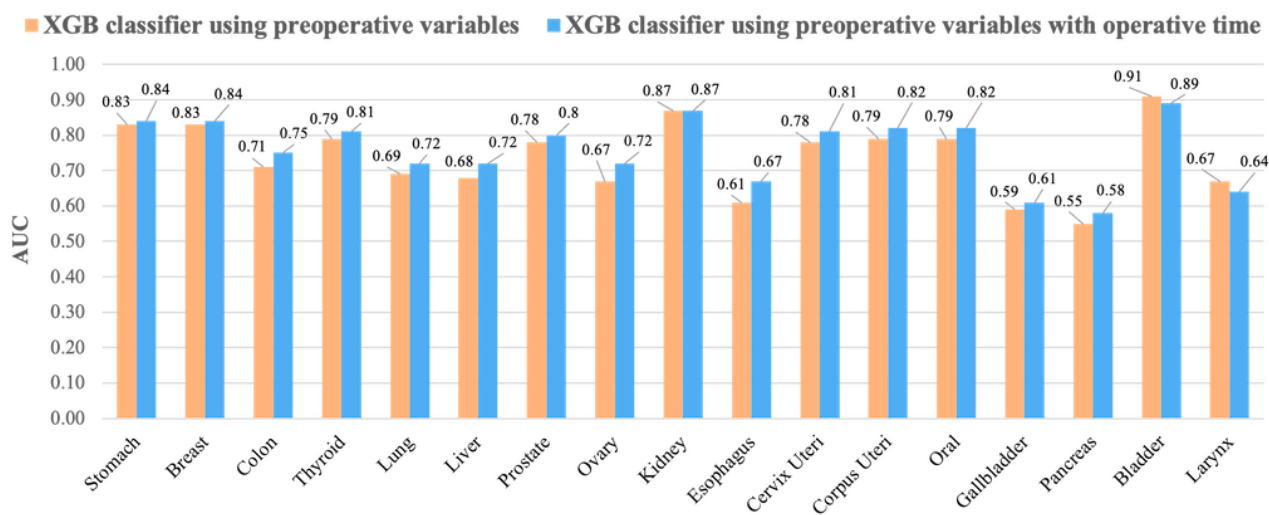


## Discussion

### Principal Findings

In cancer surgery, POLOS is both an important indicator for hospital resource use and an indirect predictor of the incidence of surgical complications and recovery of systemic conditions [25,26]. To allocate resources related to cancer surgery properly and predict the time required for recovery after surgery, an evaluation of whether POLOS can be predicted using preoperative data must be performed. Previous statistical studies have focused on identifying risk factors of PPOLOS from among the main variables. However, to develop an application that works on EHRs, an engineering approach is needed. To predict complex outcomes such as PPOLOS, data containing as many variables as possible must be used and data processing must be limited to the application in the medical field.

In the present study, we showed that PPOLOS could be partially predicted using preoperative data from EHRs for various cancer types. Acceptable predictive performance of PPOLOS (AUC >0.8) was observed for stomach, breast, prostate, kidney, and bladder cancers. For lung, liver, ovarian, esophageal, and pancreatic cancers, the predictive performance of PPOLOS was relatively low. During surgeries performed on patients with stomach, breast, prostate, kidney, and bladder cancers, cancer metastasis beyond the affected organ is relatively uncommon and the extent of surgical removal is generally considered to be uniform. In other words, in surgeries for these cancers, organ removal is the most common surgical method, and patients with cancer beyond the organ are often not candidates for surgery as the initial treatment strategy. However, surgeries for lung, liver, ovarian, esophageal, and pancreatic cancers are considered to vary significantly in practice. We assume that the prediction of recovery after surgery and PPOLOS may be possible for cancer surgery with insignificant variations in the surgical methods or with limited extent of the surgical field.

We identified the top-ranking variables associated with PPOLOS for major cancers and confirmed that the following factors correlated with PPOLOS: malnutrition (albumin and globulin), cancer stage, type of surgery, pulmonary function, and BMI [4-8]. Doxofylline and theophylline, which were used for treatment of pulmonary diseases, were associated with PPOLOS after stomach and lung cancer surgeries, respectively. Digestive drugs (famotidine, metoclopramide, and others) and pain medications (acetaminophen and tramadol) also correlated with PPOLOS for various cancers. It could be interpreted that the underlying conditions associated with the use of drugs correlate with PPOLOS, but further research is needed to confirm that the effects of certain drugs contribute to PPOLOS. We further identified that social factors—including marriage, job, and education—affect the hospital discharge time. In a previous study [27], marital status was found to be a factor affecting health care utilization among Medicare beneficiaries.

We categorized the factors that affect PPOLOS and visually identified that there are differences in the relative weight of the factors affecting PPOLOS by cancer types (Figure 2). The effects of surgical factors were relatively high in surgeries for stomach, breast, colon, and kidney cancers. The cancer stage contributed the most to the determination of PPOLOS after thyroid, breast, and ovarian cancer surgeries. For liver cancer, underlying diseases and related laboratory parameters were a major factor when determining PPOLOS. Nutritional factors largely contributed to determining PPOLOS for stomach cancer. Compared with other cancer surgeries, subjective symptoms

were an important factor in predicting PPOLOS after breast and thyroid cancer surgeries.

In this study, we aimed to predict the length of the hospital stay after surgery. However, owing to various factors occurring during surgery, it is difficult to determine POLOS. As it is difficult to evaluate the events that occur during surgery using quantitative data from EHRs, we analyzed the effect of operative time. It was observed that the predictive performance of PPOLOS increased markedly for colon, liver, ovarian, and esophageal cancer surgeries. It is believed that a model that predicts POLOS more effectively can be generated by combining preoperative data with intraoperative data, such as vital signs during anesthesia, loss of blood, and surgical instruments used.

Predictive modeling using data from EHRs is expected to improve the quality of health care and allocation of medical resources. However, studies using conventional statistical models have mainly focused on identifying risk factors for length of stay in hospital. Statistical models have limitations in processing numerous unrefined variables and in their application to real-world data. In recent years, machine learning has been used to develop predictive models [11,12]. In this study, XGB classifier and MLP showed slightly better performance than the LR model for surgeries of stomach, breast, colon, thyroid, and lung cancers, which each had more than 4000 cases. Therefore, we believe that machine learning models will be actively used as tools for predicting complex outcomes such as POLOS in the medical field.

One limitation of our study pertains to the fact that variables of data derived from the EHRs of a single cancer center in the Republic of Korea were used. Another limitation is that we used typical methods such as XGB classifier, MLP, and LR. For future study, we need to consider using multicenter EHR data and other methods for analysis. Also, we analyzed data from patients undergoing cancer surgery over a period of 18 years, during which there were likely to have been changes in patient characteristics, clinical practices (such as surgical methods), and patient care after surgery. These temporal trends may have confounded our models' performance.

If our research results are advanced, we expect to be able to create a model that predicts POLOS before surgery. Following that, it may be possible to build an application into EHRs that can automatically determine the patient's surgery day by considering the capacity of the ward.

## Conclusions

In our retrospective study, we developed models that predict PPOLOS in patients with cancer and analyzed variables affecting PPOLOS. This approach could help to provide more efficient allocation of medical resources in cancer surgery by embedding machine learning models into the EHR system to support decision making for hospital management.

## Authors' Contributions

YYJ, JHH, and YH conceived the study design. HWP and YH surveyed previous literature. HWP, HJJ, and JMJ acquired EHR data from the data warehouse in the National Cancer Center. YYJ developed and evaluated the models. All authors analyzed and discussed the results. YYJ, JHH, and YH wrote the manuscript. All authors reviewed and approved the final manuscript.

## Conflicts of Interest

None declared.

Multimedia Appendix 1
Characteristics of the cancer population.
[PDF File (Adobe PDF File), 103 KB - medinform_v9i2e23147_app1.pdf ]

Multimedia Appendix 2
Prediction performance of prolonged postoperative length of stay.
[PDF File (Adobe PDF File), 115 KB - medinform_v9i2e23147_app2.pdf ]

Multimedia Appendix 3
Top 10 variables for training the models by cancer type.
[PDF File (Adobe PDF File), 165 KB - medinform_v9i2e23147_app3.pdf ]

## References

1. Global Burden of Disease Cancer Collaboration, Fitzmaurice C, Abate D, Abbasi N, Abbastabar H, Abd-Allah F, et al. . [doi: 10.1001/jamaoncol.2019.2996] [Medline: 31560378]

2.  Mariotto AB, Yabroff KR, Shao Y, Feuer EJ, Brown ML. Projections of the cost of cancer care in the United States: 2010-2020. J Natl Cancer Inst 2011 Jan 19;103(2):117-128 [FREE Full text] [doi: 10.1093/jnci/djq495] [Medline: 21228314]

3.  Warren JL, Yabroff KR, Meekins A, Topor M, Lamont EB, Brown ML. Evaluation of trends in the cost of initial cancer treatment. J Natl Cancer Inst 2008 Jun 18;100(12):888-897 [FREE Full text] [doi: 10.1093/jnci/djn175] [Medline: 18544740]

4.  Laky B, Janda M, Kondalsamy-Chennakesavan S, Cleghorn G, Obermair A. Pretreatment malnutrition and quality of life - association with prolonged length of hospital stay among patients with gynecological cancer: a cohort study. BMC Cancer 2010 May 25;10:1-6 [FREE Full text] [doi: 10.1186/1471-2407-10-232] [Medline: 20497581]

5.  Gohil R, Rishi M, Tan BHL. Pre-operative serum albumin and neutrophil-lymphocyte ratio are associated with prolonged hospital stay following colorectal cancer surgery. Br J Med Med Res 2014 Jan 01;4(1):481-487 [FREE Full text] [doi: 10.9734/BJMMR/2014/5444#sthash.FpqnVF8X.dpuf] [Medline: 24278850]

6.  Rickard MJFX, Dent OF, Sinclair G, Chapuis PH, Bokey EL. Background and perioperative risk factors for prolonged hospital stay after resection of colorectal cancer. ANZ J Surg 2004;74(1-2):4-9. [doi: 10.1046/j.1445-1433.2003.02887.x] [Medline: 14725696]

7.  Wright CD, Gaissert HA, Grab JD, O'Brien SM, Peterson ED, Allen MS. Predictors of prolonged length of stay after lobectomy for lung cancer: a Society of Thoracic Surgeons General Thoracic Surgery Database risk-adjustment model. Ann Thorac Surg 2008 Jun;85(6):1857-1865. [doi: 10.1016/j.athoracsur.2008.03.024] [Medline: 18498784]

8.  Agrawal S, Chen L, Tergas AI, Hou JY, St Clair CM, Ananth CV, et al. Identifying modifiable and non-modifiable risk factors associated with prolonged length of stay after hysterectomy for uterine cancer. Gynecol Oncol 2018 Jun;149(3):545-553. [doi: 10.1016/j.ygyno.2018.03.048] [Medline: 29559171]

9.  Lee J, Kuo Y, Goodwin JS. The effect of electronic medical record adoption on outcomes in US hospitals. BMC Health Serv Res 2013 Feb 01;13(1):1-7 [FREE Full text] [doi: 10.1186/1472-6963-13-39] [Medline: 23375071]

10.  Park Y, Han D. Current Status of Electronic Medical Record Systems in Hospitals and Clinics in Korea. Healthc Inform Res 2017 Jul;23(3):189-198 [FREE Full text] [doi: 10.4258/hir.2017.23.3.189] [Medline: 28875054]

11.  Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes. JAMA Netw Open 2020 Jan 03;3(1):e1918962-e1918962 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.18962] [Medline: 31922560]

12.  Mortazavi BJ, Bucholz EM, Desai NR, Huang C, Curtis JP, Masoudi FA, et al. Comparison of Machine Learning Methods With National Cardiovascular Data Registry Models for Prediction of Risk of Bleeding After Percutaneous Coronary Intervention. JAMA Netw Open 2019 Jul 03;2(7):e196835-e196835 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.6835] [Medline: 31290991]

13.  Cha HS, Jung JM, Shin SY, Jang YM, Park P, Lee JW, et al. The Korea Cancer Big Data Platform (K-CBP) for Cancer Research. Int J Environ Res Public Health 2019 Jun 28;16(13):2290 [FREE Full text] [doi: 10.3390/ijerph16132290] [Medline: 31261630]

14.  Greene FL, Page DL, Fleming ID, et al. Esophagus. In: American Joint Committee on Cancer (AJCC) Cancer Staging Manual, 6th edition. New York, NY: Springer; 2002:167-178.

15.  Edge SB, Byrd DR, Compton CC, Fritz AG, Greene FL, Trotti A, editors. American Joint Committee on Cancer (AJCC) Cancer Staging Manual, 7th edition. Chicago: Springer Inc; 2010.

16.  Collins TC, Daley J, Henderson WH, Khuri SF. Risk factors for prolonged length of stay after major elective surgery. Ann Surg 1999 Aug;230(2):251. [doi: 10.1097/00000658-199908000-00016] [Medline: 10450740]

17.  Chen Y, Scholten A, Chomsky-Higgins K, Nwaogu I, Gosnell JE, Seib C, et al. Risk Factors Associated With Perioperative Complications and Prolonged Length of Stay After Laparoscopic Adrenalectomy. JAMA Surg 2018 Nov 01;153(11):1036-1041 [FREE Full text] [doi: 10.1001/jamasurg.2018.2648] [Medline: 30090934]

18.  Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: XGBoost: A Scalable Tree Boosting System. New York, NY, United States: Association for Computing Machinery; 2016 Presented at: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 13-17 August; San Francisco California USA p. 785-794. [doi: 10.1145/2939672.2939785]

19.  Svozil D, Kvasnicka V, Jir^í P. Introduction to multi-layer feed-forward neural networks. Chemometrics and Intelligent Laboratory Systems 1997 Dec;39(1):43-62 [FREE Full text] [doi: 10.1016/s0169-7439(97)00061-0]

20.  Hecht-Nielsen R. Neural Networks for Perception: Computation, Learning, and Architectures. In: Harry W, editor. III-3. Theory of the Backpropagation Neural Network. USA: Academic Press; 1992:65-93.

21.  Dhruva SS, Ross JS, Mortazavi BJ, Hurley NC, Krumholz HM, Curtis JP, et al. Association of Use of an Intravascular Microaxial Left Ventricular Assist Device vs Intra-aortic Balloon Pump With In-Hospital Mortality and Major Bleeding Among Patients With Acute Myocardial Infarction Complicated by Cardiogenic Shock. JAMA 2020 Feb 25;323(8):734-745 [FREE Full text] [doi: 10.1001/jama.2020.0254] [Medline: 32040163]

22.  Hu H, Wang H, Wang F, Langley D, Avram A, Liu M. Prediction of influenza-like illness based on the improved artificial tree algorithm and artificial neural network. Sci Rep 2018 Mar 20;8(1):4895 [FREE Full text] [doi: 10.1038/s41598-018-23075-1] [Medline: 29559649]

23.    Zanini R, Colombini E. Castro MCFd, editors. 2019 Oct Presented at: Parkinson's Disease EMG Signal Prediction Using Neural Networks; 6-9 Oct; IEEE International Conference on Systems, Man and Cybernetics (SMC) p. 2446-2453. [doi: 10.1109/smc.2019.8914553]

24.    Zeng X, An J, Lin R, Dong C, Zheng A, Li J, et al. Prediction of complications after paediatric cardiac surgery. Eur J Cardiothorac Surg 2020 Feb 01;57(2):350-358. [doi: 10.1093/ejcts/ezz198] [Medline: 31280308]

25.    McAleese P, Odling-Smee W. The effect of complications on length of stay. Ann Surg 1994 Dec;220(6):740-744. [doi: 10.1097/00000658-199412000-00006] [Medline: 7986140]

26.    Khan NA, Quan H, Bugar JM, Lemaire JB, Brant R, Ghali WA. Association of postoperative complications with hospital costs and length of stay in a tertiary care center. J Gen Intern Med 2006 Feb;21(2):177-180 [FREE Full text] [doi: 10.1111/j.1525-1497.2006.00319.x] [Medline: 16606377]

27.    Pandey KR, Yang F, Cagney KA, Smieliauskas F, Meltzer DO, Ruhnke GW. The impact of marital status on health care utilization among Medicare beneficiaries. Medicine (Baltimore) 2019 Mar;98(12):e14871. [doi: 10.1097/MD.0000000000014871] [Medline: 30896632]

## Abbreviations

**AUC:** area under the receiver operating characteristic curve
**EHR:** electronic health record
**K-CBP:** Korea Cancer Big Data Platform
**LR:** logistic regression
**MLP:** multilayer perceptron
**POLOS:** postoperative length of stay
**PPOLOS:** prolonged postoperative length of stay
**ROC:** receiver operating characteristic
**XGB:** extreme gradient boosting

Original Paper

# Using Machine Learning to Collect and Facilitate Remote Access to Biomedical Databases: Development of the Biomedical Database Inventory

Eduardo Rosado[1], BSc; Miguel Garcia-Remesal[1], PhD; Sergio Paraiso-Medina[1], PhD; Alejandro Pazos[2], MD, PhD; Victor Maojo[1], MD, PhD

[1]Biomedical Informatics Group, School of Computer Science, Universidad Politecnica de Madrid, Madrid, Spain

[2]Grupo de Redes de Neuronas Artificiales y Sistemas Adaptativos - Imagen Médica y Diagnóstico Radiológico, Department of Computer Science and Information Technologies, Faculty of Computer Science, University of A Coruña, A Coruña, Spain

**Corresponding Author:**
Victor Maojo, MD, PhD
Biomedical Informatics Group
School of Computer Science
Universidad Politecnica de Madrid
Campus de Montegancedo, s/n
Madrid, 28660
Spain
Phone: 34 699059254
Email: vmaojo@fi.upm.es

## Abstract

**Background:** Currently, existing biomedical literature repositories do not commonly provide users with specific means to locate and remotely access biomedical databases.

**Objective:** To address this issue, we developed the Biomedical Database Inventory (BiDI), a repository linking to biomedical databases automatically extracted from the scientific literature. BiDI provides an index of data resources and a path to access them seamlessly.

**Methods:** We designed an ensemble of deep learning methods to extract database mentions. To train the system, we annotated a set of 1242 articles that included mentions of database publications. Such a data set was used along with transfer learning techniques to train an ensemble of deep learning natural language processing models targeted at database publication detection.

**Results:** The system obtained an F1 score of 0.929 on database detection, showing high precision and recall values. When applying this model to the PubMed and PubMed Central databases, we identified over 10,000 unique databases. The ensemble model also extracted the weblinks to the reported databases and discarded irrelevant links. For the extraction of weblinks, the model achieved a cross-validated F1 score of 0.908. We show two use cases: one related to "omics" and the other related to the COVID-19 pandemic.

**Conclusions:** BiDI enables access to biomedical resources over the internet and facilitates data-driven research and other scientific initiatives. The repository is openly available online and will be regularly updated with an automatic text processing pipeline. The approach can be reused to create repositories of different types (ie, biomedical and others).

## Introduction

Since the inception of the web, the amount of information available online has dramatically increased. Such an explosion can be mainly observed in the biomedical area, where the publication of the Human Genome Project [1] led to a myriad of new primary and translational projects. The latter produced a vast amount of additional "omics" information that needed to be remotely found, accessed, collected, managed, analyzed, and used. This acceleration in data production has been observed in biology and bioinformatics in particular [2,3].

XSL•FO
**RenderX**

To facilitate access to such a plethora of information, thousands of different databases were created by many scientists to exchange their knowledge and data with other colleagues and institutions. The number of database publications is continuously increasing; therefore, giving visibility and access to those resources can be a complicated task. A study on the usage of databases and software in articles available at PubMed Central (PMC) [4] found that the top 5% most popular resources accounted for 47% of all citations. In comparison, 70% of all detected resources were referenced just once. This focus on a few popular resources suggests a wasted opportunity for researchers to benefit from many informatics tools designed and published to support scientific research and clinical practice.

A successful biomedical information classification initiative was the Unified Medical Language System (UMLS) [5]. Led by Lindberg, Humphreys, McCray, and staff from the National Library of Medicine (NLM), UMLS's original mission was to facilitate computer programs in understanding and accessing biomedical literature. However, the identification of databases in scientific articles is not straightforward within UMLS. This initiative focuses on the description of scientific research and the grouping of data under big repositories for clinical and genetic data, such as the Systematized Nomenclature of Medicine (SNOMED) [6] and Online Mendelian Inheritance in Man (OMIM) [7], while neglecting smaller databases that comprise the vast majority of the available resources.

Currently, most efforts toward detecting and including new databases in repositories have relied on manual approaches. Such a strategy involving human resources cannot scale properly. Previous work in manual database compilation includes, for instance, the Database of Databases (DoD2007) [8]. The DoD2007 repository has increased over the years, having reached a total of 1082 molecular biology databases at the time of writing. The journal Nucleic Acids Research (NAR) releases a yearly update on their molecular biology database collection, with the 2020 publication [9] containing a total of 148 scientific articles either presenting a new database or reporting an update to a previously existing database. Another study [10] collected a set of 112 widely used human-related biological databases. Fairsharing (formerly Biosharing) [11] is a regularly updated, curated, and crowdsourced collection of life sciences resources. As of May 2020, it contains 1470 databases as claimed on their webpage.

The resources mentioned have been of great use for the research community. However, using manual labor for such collection efforts is a costly endeavor. Figure 1 shows the number of new publications indexed in PubMed each year, a density distribution instead of a cumulative one. As can be seen, the publication rate has been accelerating continuously, especially since the start of the new millennium. In 1950, PubMed registered 85,792 new publications; in 2019, however, 1,392,830 new publications were registered. Such numbers imply an increase of 1623% in publication rate. Since the year 2011, more than one million articles have been registered on the platform each year.

**Figure 1.** Evolution of new publications in PubMed (ie, the MEDLINE database) per year from 1950 to 2018.
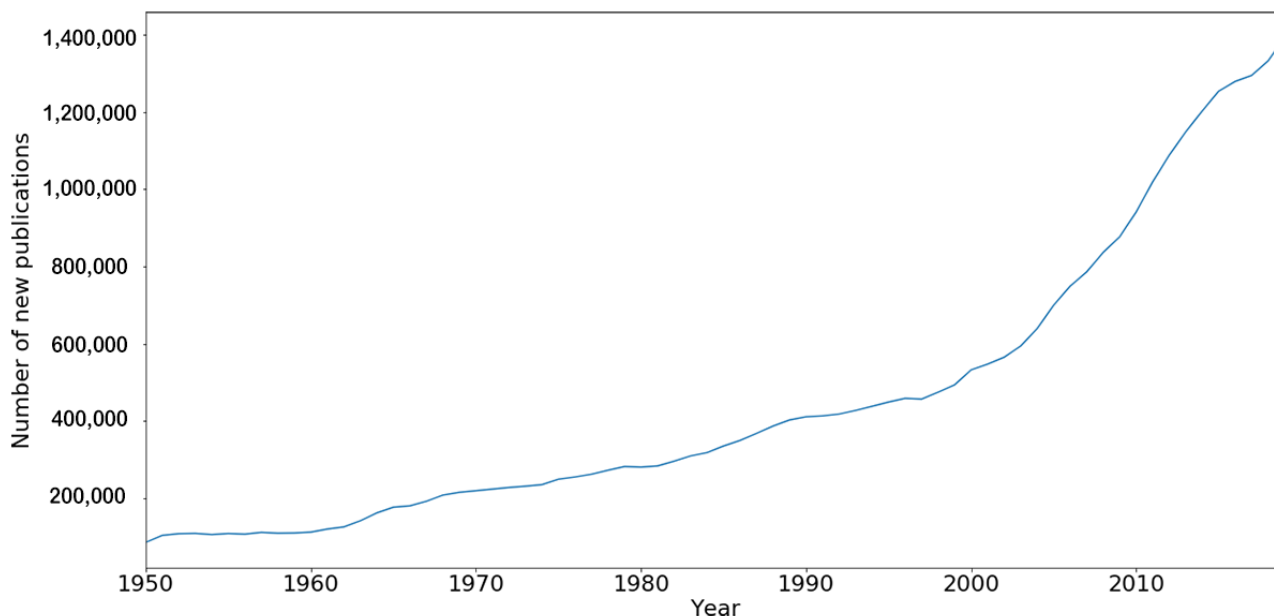


In this context, we consider automatic methods for information extraction to be a good alternative, if not a necessary one. Here we present the Biomedical Database Inventory (BiDI), a collection of databases automatically extracted from the scientific literature by applying deep learning methods for natural language processing (NLP).

## Methods

### Model Architecture

To detect databases in the literature, we have created an ensemble of NLP models adapted from different reported architectures [12-14] for masked word prediction.

In Figure 2, we show a simplified version of the training process for masked word prediction. Initially, we replace a word from

the sentence with the special token [MASK]. Each input token is passed through a token embedding layer, thus converting each token into a vector representation. At this stage, [MASK] is just another input token and is represented by a single embedding, which is the same for all [MASK] tokens. These vectors are general word embeddings, which are randomly initialized and trained jointly with the rest of the network using gradient descent. The latter propagates through the entire network to the input token embeddings, including the [MASK] token, and, thus, they are updated just like any other model parameter.

**Figure 2.** The pretraining setup for the masked word (w) prediction architecture. Only one encoder block is shown. The sample sentence only contains three words. However, in practice each sentence has a length of 512, adding padding to the right when necessary. [MASK] is a special token that is used to replace a word from the sentence.
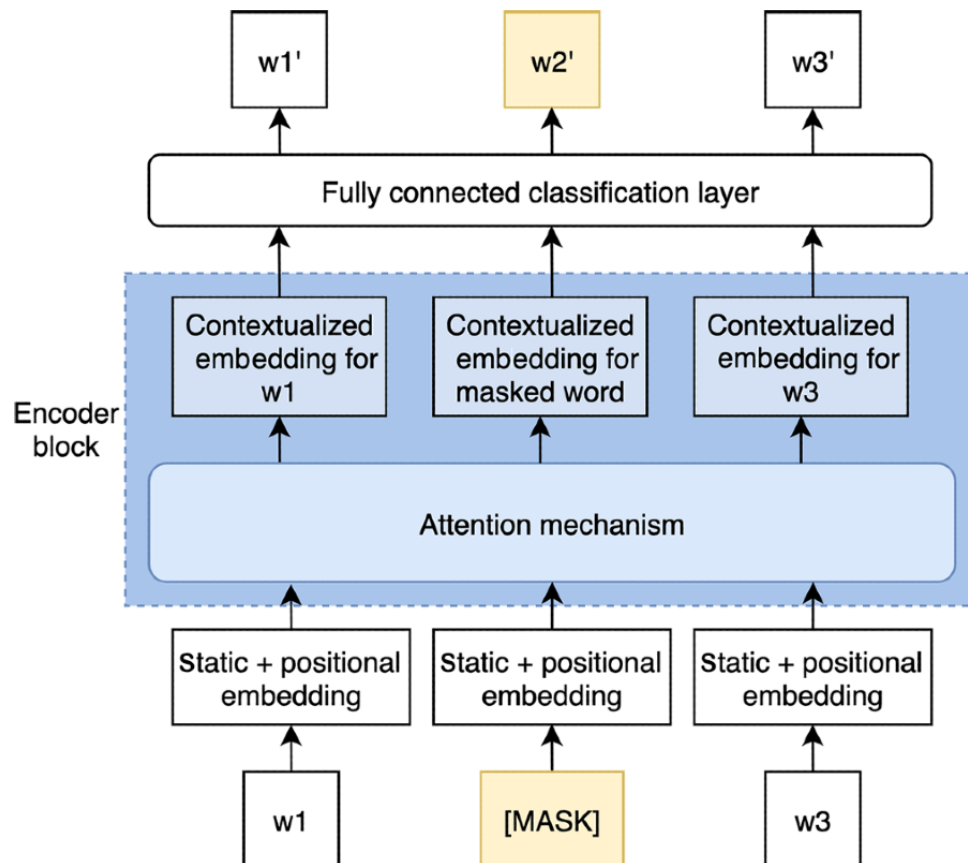
The network also learns positional embeddings, which refer to the token sequence of input. These are represented as a combination of sine and cosine functions of different frequencies, which provide the model with information regarding the token's situation. Finally, each token is replaced by the sum of the static token embedding and the positional embedding.

Then, we apply the attention mechanism by which a contextualized embedding is produced for each word. This embedding is a weighted sum of all input vectors multiplied by a value matrix. The weight factor would be the attention score computed for each word. Finally, each of the contextualized vectors is fed to a fully connected layer to produce a new word. The loss is calculated as the cross-entropy loss between the masked word and the model's output in that position, discarding the predictions for all nonmasked terms. We refer to Liu et al [12] and Devlin et al [14] for specific details about the underlying architecture.

By masking 15% of every sentence repeatedly with different masking schemes each time, the model learns to predict any word in any sentence. When the training process involves millions of sentences, the result is a general-purpose language model capable of capturing complex dependencies between words. We shall remark that the particular architecture we applied has a total of 12 encoder blocks, stacked one on top of another, and a replicated attention mechanism with 12 attention heads on each encoder block. Therefore, 12 different attention schemes are applied simultaneously to focus on different input aspects.

Our task is an instance of sentence classification. Therefore, we do not need to consider the final contextualized embeddings of each word. Instead, we can perform a forward pass on the model and extract only the first embedding, which corresponds to the special classification token [CLS]. The [CLS] token goal is to mark the start of a sentence, and it can be easily fine-tuned to represent global information about the whole input. In Figure 3, we represent our fine-tuning process.

**Figure 3.** Fine-tuning the setup for our model. Only one encoder block is shown. The sample sentence only contains three words (w). However, in practice each sentence has a length of 512, adding padding to the right when necessary. [CLS]: classification token.
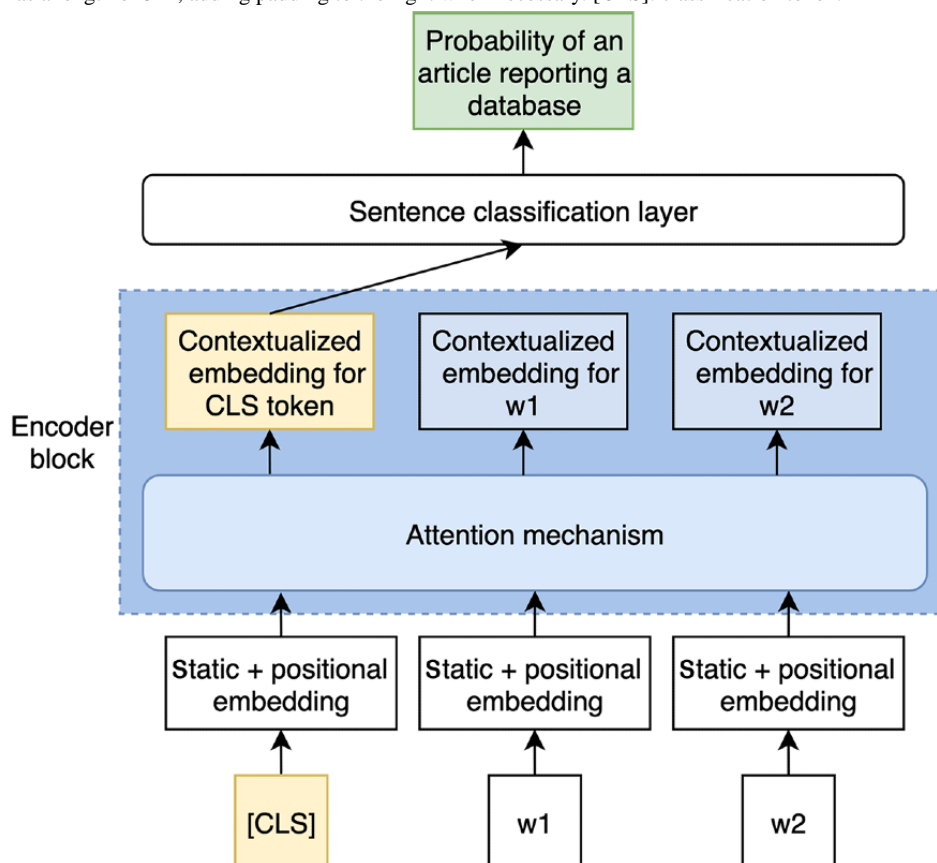


As shown in the figure, we extract the [CLS] token and feed it to a new, fully connected classification layer to obtain a probability for the whole sentence. However, we update the parameter's weights on the classification layer and every parameter in the encoder blocks by backpropagation. Therefore, the model transfers the relevant information regarding our task into the [CLS] embedding. Each encoder block adds information to that vector before feeding it to the classification layer.

We trained three different models under the same architecture. First, we developed a model to detect a database publication considering only the title of the article. Then we trained a second model to perform the same task with sentences extracted from the abstract. After considering the particular structure of a title sentence, we committed to this design, which usually contains vital information about the article. We aimed to train a model

that focused on title sentences to extract all the information about them. Finally, we trained a model on the task of database link classification (ie, to differentiate between sentences with a link to a database homepage from sentences with a link unrelated to the database). This third model allows us to extract the right link from every article, achieving our goal of directly linking articles to data.

In Figure 4, we represent the complete procedure, which consists of an ensemble of the three models. The first step is to classify an article by the title; only when it is classified as a negative sample do we resort to the abstract sentences for confirmation. The third model is applied to sentences containing a link only if either the first or the second model delivered a positive label for the article.

**Figure 4.** Pipeline for article classification.
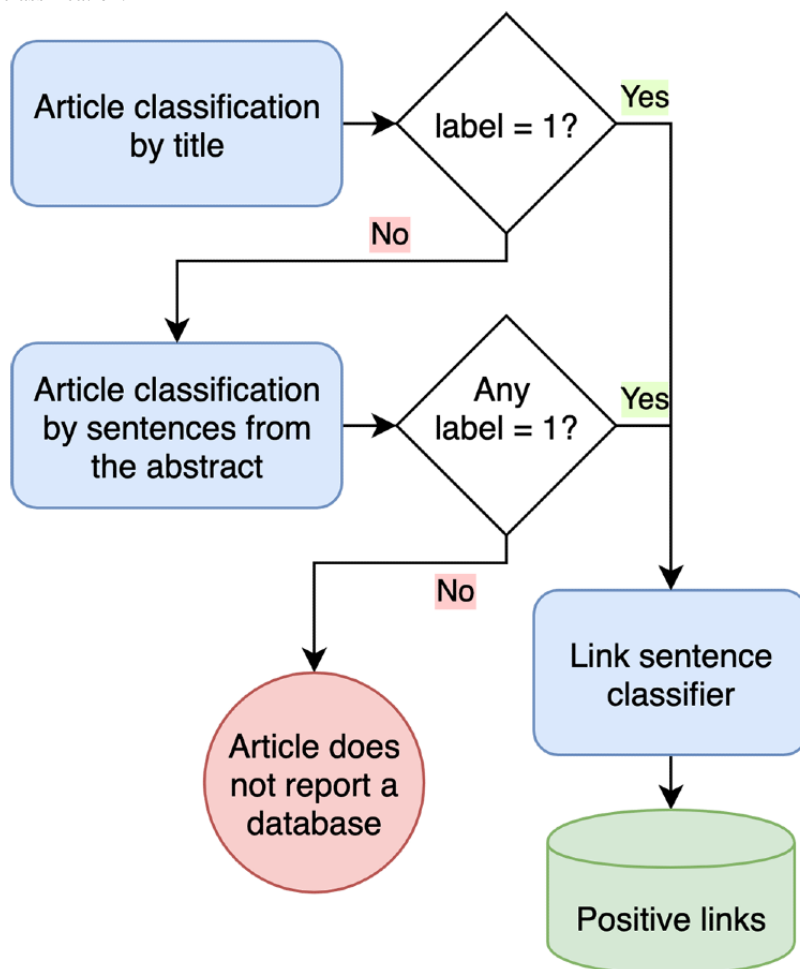


## Data Collection and Preprocessing

We performed database publication detection on two scientific article repositories: PubMed and PMC. For PubMed, we downloaded a total of 12,615,511 articles with abstracts. For PMC, we downloaded the Open Access Subset with a total of 2,710,216 articles. Aside from the number of manuscripts, the main difference between these data sources is that PMC offers full-text articles, while PubMed only provides titles and abstracts.

Article data were subject to a series of preprocessing steps before feeding them to the model for training. The process is depicted in Figure 5.

**Figure 5.** Data preprocessing steps. The subtasks inside each step were executed in order from top to bottom. BERT: Bidirectional Encoder Representations from Transformers; NLTK: Natural Language Toolkit.
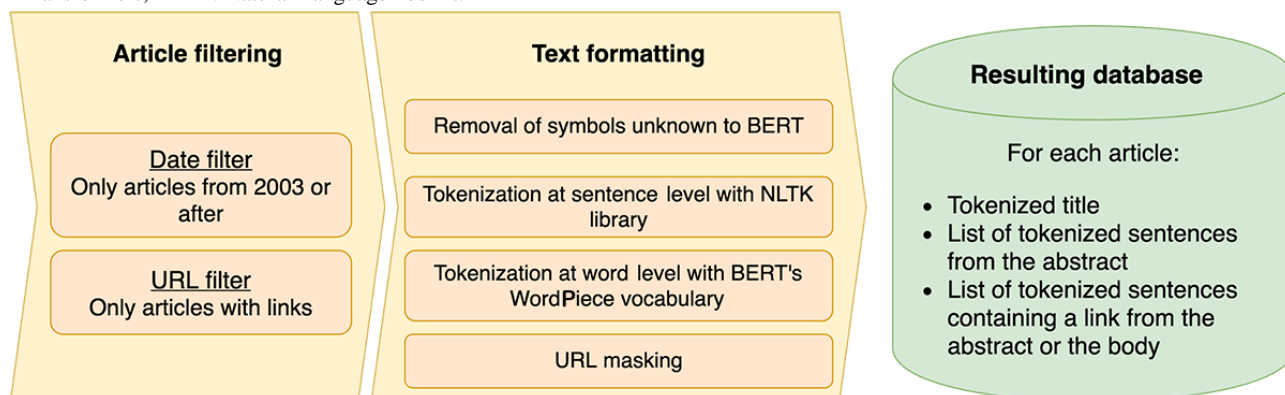


We now describe each step in the preprocessing pipeline:

1. Article filtering. We only considered articles starting from 2003 since this was the year that marked the completion of the Human Genome Project [1], a milestone for the life sciences research community. We also removed articles not containing at least one URL, as we aimed to link the publication to the database. The URLs were extracted by applying a regular expression, using the substrings "http" and "www" as anchors.

2. Text formatting:

a. Removal of symbols. The model used a subword vocabulary of about 30,000 items. Such a vocabulary included all the letters in the English language, making it possible for the model to represent any English word, even if it was never seen by the model before, as, ultimately, it can decompose it into each letter. For languages with an alphabet different from the one used for the English language, however, the coverage was not guaranteed. We removed such foreign characters.

b. Sentence tokenization. We used the Natural Language Toolkit library for Python (Python Software Foundation) to split the whole text into smaller sentences. This library provides a rule-based system that carries out this partition process while maintaining each sentence's semantic integrity.

c. Word tokenization. Words in the text were tokenized according to WordPiece tokens, expanding every article's word count each time a word was split into several subwords.

d. URL masking. In our experiments, we found that URL links are hard to process for the model. We wanted the model to know that a link is present in the sentence. For this reason, we decided to mask all URLs with the keyword "link." At the same time, we stored the original URL links to retrieve them later.

In the end, the preprocessing stage produced a database with three different fields for each article. The first one contained the title sentence, the second one consisted of all the abstract sentences, and the third one included all sentences containing a link. The third field's sentences were extracted from the abstract for PubMed articles and from the abstract and the body for PMC articles. The preprocessing stage resulted in a reduced set of papers: 24,437 manuscripts for PubMed and 450,777 for PMC. The PubMed data set reduction was quite drastic, as it is not a common practice to include a link to the resource in the abstract. This is a typical pattern that, in our opinion, restricts the capabilities of search engines to find such information and direct access to the reported databases, since PubMed only provides the abstract.

## Building a Labeled Data Set

We compiled three annotated data sets, one for each of the models we trained. All data sets were developed from the same set of 1242 articles. We created labels for each analyzed manuscript for its title, abstract, and link sentences before considering the next one. All tags were either positive or negative, as in binary classification. The annotation was performed by two human annotators with biomedical informatics backgrounds—one of them also has formal training in the medical field. Each human annotator independently annotated the data sets. The degree of agreement regarding the annotations' results for each category (ie, titles, abstract sentences, and link sentences) was assessed using Cohen κ. We obtained an almost perfect consensus for each of the categories.

In Table 1, we provide the count of positive and negative annotations performed. As can be seen, there was a significant data imbalance favoring the negative samples. The articles' real distribution was actually even more skewed toward negative samples. Therefore, we believe that our data sets suitably represented the actual data distribution. Due to this imbalance, it was not easy to manually find positive articles. We actively facilitated the annotation of positive examples by discarding any negative examples we found after sampling 1000 manuscripts (ie, after the 1000th sample, we only added positive examples to the data set). We continued sampling until we collected a reasonable number of positive samples so that the model could learn the syntactic and semantic patterns of the target group.

**Table 1.** Positive and negative samples, the total number of labels for each data field, and annotator agreement evaluation per category. In all cases, the unit for the count is a sentence.

| Category | Positive samples, n (%) | Negative samples, n (%) | Total, N (%) | Annotator agreement, Cohen κ |
|---|---|---|---|---|
| Titles | 320 (25.8) | 922 (74.2) | 1242 (100) | 1 |
| Abstract sentences | 775 (7.5) | 9535 (92.5) | 10,310 (100) | 0.991 |
| Link sentences | 441 (34.3) | 845 (65.7) | 1286 (100) | 0.988 |

Table 2 shows some sample sentences extracted from the training set. The columns show the articles' annotation, while the rows indicate the annotation per sentence (ie, titles and abstract sentences). A positively annotated sentence implies that the manuscript contains a database. Conversely, negatively annotated sentences can be found in positive and negative articles. Specifically, the sentence "Mobility changes in response to COVID-19" is the title of an article that presents a free access database to the community [15]; however, the title does not show evidence of this and, therefore, receives a negative label. Most of the abstract sentences do not present direct evidence of mentioning a database. On average, we found that only 2 sentences from the abstracts of papers that confirmed to be reporting a database were also positively annotated.

**Table 2.** Examples of trained sentences. Annotation of title and abstract sentences extracted from annotated articles.

| Category and annotation of the trained set | Positive article (database found) | Negative article (database not found) |
| --- | --- | --- |
| Positive title | CoV2ID: Detection and therapeutics oligo database for SARS-CoV-2 | N/A[a] |
| Negative title | Mobility changes in response to COVID-19 | Predicting care and repercussions for caregivers of surgical patients at home |
| Positive abstract sentence | We have created a comprehensive manually curated database of circular RNAs associated with diseases. | N/A |
| Negative abstract sentence | Recent studies have shown the role of circRNAs in a number of diseases. | A non-randomized and consecutive sample of 317 informal caregivers of surgical patients with abdominal surgery was included in the study. |

[a]N/A: not applicable. No positive sentence from this category could be found in articles where no database was found.

## Results

### Overview

In Figure 6, we present the performance of the three models in their particular tasks. We assessed the models with a 5-fold cross-validation approach. We show the receiver operating characteristic curves and the associated area under the curve (AUC) values in the figure.

**Figure 6.** Receiver operating characteristic curve and area under the curve (AUC) values for the 5-fold cross-validation experiment on the three trained models. The dotted diagonal lines represent random chance. (A) Title classification; (B) abstract classification; (C) link classification.
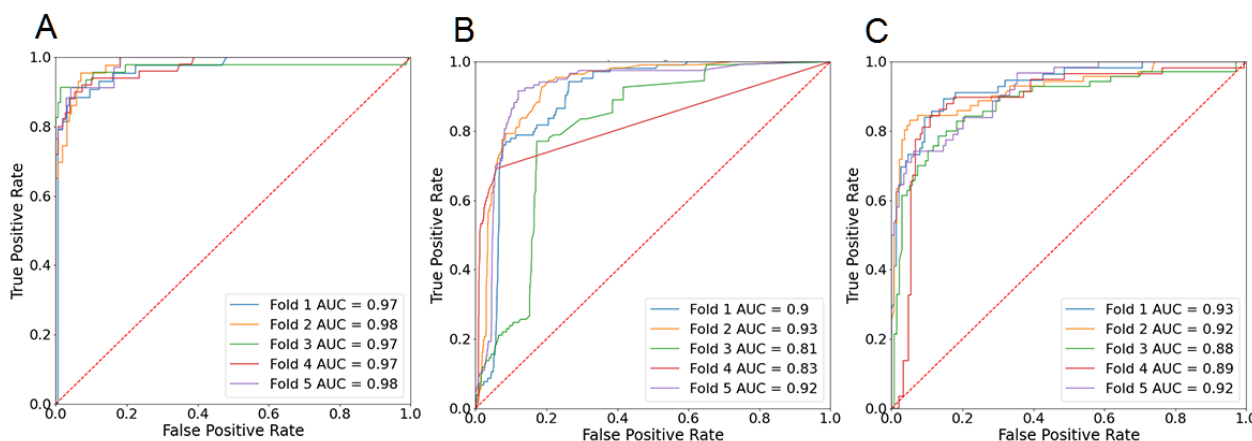


As can be seen, the title classification model obtained very high AUC values for every fold, all of them above 0.95, showing high precision and recall scores under the different data partitions. Conversely, the model for abstract sentences achieved high AUC values (above 0.90) on some of the divisions but not on all of them, as two of them presented a score ranging between 0.80 and 0.85. Finally, the link sentences model showed consistent performance, with every AUC score within a 0.03 distance from 0.90.

We also evaluated the whole system by applying the ensemble model on database publication detection, following the algorithm described in Figure 4, and comparing the results to those obtained with its individual components alone, namely the title model and the abstract sentences. This validation was made with a single partition of 30% and 70% for training over the samples.

Note that this task was not the same as the one performed in the previous experiment. We illustrate this with an example: if an article shows evidence of reporting a new database in the abstract but not in the title, such a title is assigned a negative label. In contrast, the paper as a whole is labeled as a positive sample. When we apply the title model to the aforementioned article title, if the model outputs a negative prediction, it is considered a hit in the cross-validation experiment. However, it is a miss for the ensemble model evaluation experiment since the predicted label does not match the expected class.

In Table 3, we present the results regarding precision, recall, and F1 score. The highest precision was obtained with the title model, but this model also yielded the lowest recall score. On the other hand, the abstract sentences model had a lower precision but achieved a better F1 score. Finally, we can see from the table that the ensemble model yielded the best recall and the best F1 score, showing a higher overall performance than both individual models.

**Table 3.** Precision, recall, and F1 scores for the title sentence model, the abstract sentences model, the ensemble model on the article classification task, and the weblinks.

| Category | Precision | Recall | F1 score |
|---|---|---|---|
| Title model | 0.960 | 0.795 | 0.870 |
| Abstract model | 0.911 | 0.923 | 0.917 |
| Ensemble (title + abstract) | 0.900 | 0.959 | 0.929 |
| Weblinks | 0.922 | 0.893 | 0.908 |

To build the BiDI database, we applied the ensemble model to all the articles collected from PubMed, PMC, and the COVID-19 Open Research Dataset (CORD-19). The total number of manuscripts that received a positive annotation, after removing duplicates, was 10,417: 5354 from PubMed, 5001 from PMC, and 62 from CORD-19. It is important to note that we found very few false positive cases in this list. One example was the article with the PubMed identifier 32226598, which was positive due to some sentences having reported computer resources and because it included an "Associated Data" category on the article's header, but it was empty (ie, "Not applicable").

## Use Cases

BiDI provides a search engine based on the Medical Subject Headings (MeSH) vocabulary [16]. This taxonomy, included in the UMLS Metathesaurus, is a controlled vocabulary, a collection of medical-related terms. The NLM uses MeSH terms to classify PubMed articles. As of January 2020, the taxonomy contains more than 29,000 elements. By allowing the user to filter papers by these terms, we enable the application of very specialized queries.

To demonstrate the utility provided by BiDI, we now present two use cases. Let us suppose that we want to find and access data repositories on single nucleotide polymorphisms for specific ethnic populations. Through BiDI, we performed a text search by typing "single nucleotide polymorphism," accepting matches from either the title or the abstract. BiDI returned a total of 230 articles with associated databases. We then applied MeSH term filters to narrow the search. We selected the "Ethnic Groups" MeSH term, and after proceeding with the query expansion, BiDI presented two papers reporting databases relevant to our query. The final results are shown in Figure 7.

**Figure 7.** Search results for single nucleotide polymorphisms databases in the Biomedical Database Inventory (BiDI) web platform.



Similarly, suppose that we want to find genomic databases related to coronavirus. We performed a text search by typing "coronavirus." A total of 16 results were returned, including repositories with heterogeneous data types, such as fatality rates or Twitter messages regarding coronavirus. To narrow the search, we selected the "Genomics" MeSH term. The query returned three articles reporting coronavirus genomic databases.

## *Discussion*

The number of publications and associated databases included in BiDI is one order of magnitude higher than those of manually collected database repositories. The possibility to perform automatic and regular updates is also a significant advantage. In that sense, our NLP system can analyze up to 42 articles per second on an average commercial graphics card, therefore updating thousands of manuscripts in minutes instead of hours or days, as would be expected if a team of human curators did the work.

Concerning general-purpose repositories such as PubMed, we consider BiDI a complement to them. Moreover, a PubMed weblink is provided along with every article when clicking on the title. BiDI cannot be directly compared to PubMed, as the former is a specialized subset of the latter, and it is not intended to be a replacement but an extension of PubMed. BiDI's objective is to provide fast access to databases and their associated articles; therefore, a link to the data is included along with every paper. The PubMed platform does not directly acknowledge the resources presented in articles since its goal is to provide generic access to a large number of manuscripts.

Regarding performance, BiDI has shown high precision and recall scores with a training data set of moderate size. We can only expect better performance and generalization with more training data given the superior data scalability provided by deep learning models.

Given that we only considered articles providing URLs, the resulting repository acts as a direct link between the papers and the actual data sets; it then essentially becomes a bridge between research and data, as proposed in earlier studies. In particular, Hoogerwerf et al [17] described the efforts made by the OpenAIRE initiative [18] to promote discipline-independent linking practices between publications, data, project information, and researchers. BiDI complies with these guidelines and, in the future, it could be expanded to integrate project metadata and author information.

BiDI's mission is to make scientific database resources easier to find and easy to access to facilitate biomedical scientists' and clinicians' routine work. Many authors currently do not provide easy access to their experimental data, after a deidentification process to prevent personal data rights violations. In this context, many initiatives have been launched to increase data sharing in science. For instance, the Findable, Accessible, Interoperable, and Reusable (FAIR) Principles [19] defined a set of recommendations focused on improving findability, accessibility, interoperability, and scientific data reusability. BiDI contributes mainly to the first two principles and aims to extend its service to the community as more researchers align themselves with the FAIR Principles and share their data to catalyze scientific discovery.

We may think of medical imaging and the "omics" fields as obvious candidates to benefit from BiDI, due to the amount of data generated by experiments in those areas. However, almost any domain can benefit from our repository to a certain extent. Others can easily reimplement the approach to create their own search tools. Ultimately, BiDI may enable the reuse of biomedical resources and facilitate data-driven research and other scientific initiatives.

We integrated BiDI into an automatic text processing pipeline to update the repository regularly. BiDI is openly available online [20].

## Acknowledgments

## Authors' Contributions

ER, MGR, and VM designed the study and the methods. MGR, SPM, and VM collected and analyzed the data. ER, MGR, SPM, AP, and VM interpreted the results and wrote the article. VM, SPM, and AP revised the manuscript. All authors approved the final draft of the document. The corresponding author had full access to all of the data in the study and submitted the manuscript to be considered for publication.

## Conflicts of Interest

None declared.

## References

1. Collins F, Morgan M, Patrinos A. The Human Genome Project: Lessons from large-scale biology. Science 2003 Apr 11;300(5617):286-290. [doi: 10.1126/science.1084564] [Medline: 12690187]
2. Marx V. Biology: The big challenges of big data. Nature 2013 Jun 13;498(7453):255-260. [doi: 10.1038/498255a] [Medline: 23765498]
3. Greene CS, Tan J, Ung M, Moore JH, Cheng C. Big data bioinformatics. J Cell Physiol 2014 Dec;229(12):1896-1900 [FREE Full text] [doi: 10.1002/jcp.24662] [Medline: 24799088]
4. Duck G, Nenadic G, Filannino M, Brass A, Robertson DL, Stevens R. A survey of bioinformatics database and software usage through mining the literature. PLoS One 2016;11(6):e0157989 [FREE Full text] [doi: 10.1371/journal.pone.0157989] [Medline: 27331905]
5. Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Yearb Med Inform 1993(1):41-51 [FREE Full text] [Medline: 27668467]
6. Cornet R, de Keizer N. Forty years of SNOMED: A literature review. BMC Med Inform Decis Mak 2008 Oct 27;8 Suppl 1:S2 [FREE Full text] [doi: 10.1186/1472-6947-8-S1-S2] [Medline: 19007439]
7. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2005 Jan 01;33(Database issue):D514-D517 [FREE Full text] [doi: 10.1093/nar/gki033] [Medline: 15608251]
8. Babu PA, Udyama J, Kumar RK, Boddepalli R, Mangala DS, Rao GN. DoD2007: 1082 molecular biology databases. Bioinformation 2007 Oct 12;2(2):64-67 [FREE Full text] [doi: 10.6026/97320630002064] [Medline: 18188423]

9.   Rigden D, Fernández XM. The 27th annual Nucleic Acids Research database issue and molecular biology database collection. Nucleic Acids Res 2020 Jan 08;48(D1):D1-D8 [FREE Full text] [doi: 10.1093/nar/gkz1161] [Medline: 31906604]

10.  Zou D, Ma L, Yu J, Zhang Z. Biological databases for human research. Genomics Proteomics Bioinformatics 2015 Feb;13(1):55-63 [FREE Full text] [doi: 10.1016/j.gpb.2015.01.006] [Medline: 25712261]

11.  McQuilton P, Gonzalez-Beltran A, Rocca-Serra P, Thurston M, Lister A, Maguire E, et al. BioSharing: Curated and crowd-sourced metadata standards, databases and data policies in the life sciences. Database (Oxford) 2016;2016:baw075 [FREE Full text] [doi: 10.1093/database/baw075] [Medline: 27189610]

12.  Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A robustly optimized BERT pretraining approach. arXiv Preprint posted online on July 26, 2019. [FREE Full text]

13.  Vaswani A. Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). 2017 Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); December 4-9, 2017; Long Beach, CA p. 5999-6009 URL: https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

14.  Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on May 24, 2019 [FREE Full text]

15.  Warren MS, Skillman SW. Mobility changes in response to COVID-19. arXiv. Preprint posted online on March 31, 2020 [FREE Full text]

16.  Lipscomb C. Medical Subject Headings (MeSH). Bull Med Libr Assoc 2000 Jul;88(3):265-266 [FREE Full text] [Medline: 10928714]

17.  Hoogerwerf M, Lösch M, Schirrwagen J, Callaghan S, Manghi P, Iatropoulou K, et al. Linking data and publications: Towards a cross-disciplinary approach. Int J Digit Curation 2013 Jun 14;8(1):244-254 [FREE Full text] [doi: 10.2218/ijdc.v8i1.257]

18.  Rettberg N, Schmidt B. OpenAIRE - Building a collaborative open access infrastructure for European researchers. LIBER Q 2012 Nov 29;22(3):160 [FREE Full text] [doi: 10.18352/lq.8110]

19.  Wilkinson M, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]

20.  BiDI - Biomedical Database Inventory. Madrid, Spain: Biomedical Informatics Group, Universidad Politécnica de Madrid; 2020. URL: http://gib.fi.upm.es/bidi/ [accessed 2021-02-14]

## Abbreviations

**AUC:** area under the curve
**BiDI:** Biomedical Database Inventory
**[CLS]:** classification token
**CORD-19:** COVID-19 Open Research Dataset
**DoD2007:** Database of Databases
**FAIR:** Findable, Accessible, Interoperable, and Reusable
**MeSH:** Medical Subject Headings
**NAR:** Nucleic Acids Research
**NLM:** National Library of Medicine
**NLP:** natural language processing
**OMIM:** Online Mendelian Inheritance in Man
**PMC:** PubMed Central
**SNOMED:** Systematized Nomenclature of Medicine
**UMLS:** Unified Medical Language System

XSL•FO
**RenderX**

Original Paper

# Medication Management Service for Old Age Homes in Hong Kong Using Information Technology, Automation Technology, and the Internet of Things: Pre-Post Interventional Study

Kei Hong So[1,2], BPharm (Hons), MCP; Cheuk Wun Ting[1], BPharm (Hons); Chui Ping Lee[1], PharmD; Teddy Tai-Ning Lam[1], PharmD, PhD; Sau Chu Chiang[2], BPharm (Hons), MHA; Yin Ting Cheung[1], PhD

[1]School of Pharmacy, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, Hong Kong (China)
[2]Hong Kong Pharmaceutical Care Foundation, Hong Kong, Hong Kong (China)

**Corresponding Author:**
Yin Ting Cheung, PhD
School of Pharmacy
Faculty of Medicine
The Chinese University of Hong Kong
8th Floor, Lo Kwee-Seong Integrated Biomedical Sciences Building
Area 39, The Chinese University of Hong Kong, N.T. Shatin
Hong Kong
Hong Kong (China)
Phone: 852 39436833
Email: yinting.cheung@cuhk.edu.hk

## Abstract

**Background:** Innovation in technology and automation has been increasingly used to improve conventional medication management processes. In Hong Kong, the current practices of medication management in old age homes (OAHs) are time consuming, labor intensive, and error prone. To address this problem, we initiated an integrated medication management service combining information technology, automation technology, and the Internet of Things in a cluster network of OAHs.

**Objective:** This pilot study aimed to evaluate the impact of the medication management program on (1) medication management efficiency, (2) medication safety, and (3) drug wastage in OAHs. We compared the time efficiency and the reductions in medication errors and medication wastage in OAHs before and at least 2 weeks after the implementation of the program.

**Methods:** From November 2019 to February 2020, we recruited 2 OAHs (serving 178 residents) in Hong Kong into the prospective, pre-post interventional study. The interventional program consisted of electronic medication profiles, automated packaging, and electronic records of medication administration. Using 3-way analysis of variance, we compared the number of doses prepared and checked in 10-minute blocks before and after implementation. We received anonymous reports of medication errors from OAH staff and analyzed the results with the Fisher exact test. We also calculated the quantity and cost of wasted medications from drug disposal reports.

**Results:** The number of doses prepared and checked in 10-minute blocks significantly increased postimplementation (pre: 41.3, SD 31.8; post: 70.6, SD 22.8; $P<.001$). There was also a significant reduction in medication errors (pre: 10/9504 doses, 0.1%; post: 0/5731 doses; $P=.02$). The total costs of wasted medications during January 2020 in OAH 1 (77 residents) and OAH 2 (101 residents) were HK $2566.03 (US $328.98) and HK $5249.48 (US $673.01), respectively.

**Conclusions:** Our pilot study suggested that an innovative medication management program with information technology, automation technology, and Internet of Things components improved the time efficiency of medication preparation and medication safety for OAHs. It is a promising solution to address the current limitations in medication management in OAHs in Hong Kong.

**KEYWORDS**

XSL•FO
**RenderX**

## Introduction

Reports from the literature have shown that medication errors in long-term care facilities are common and can have deleterious consequences. One study of 865 long-term care facilities in Japan found that the incidence rate of medication errors was 40.0 per 1000 residents [1]. A local study in Hong Kong reported that 46.4% of drug-related problems in old age homes (OAHs) were drug administration errors [2]. To address this problem, user-friendly information technology (IT) systems and electronic medication administration records (eMARs) have been increasingly used in long-term care facilities as they allow more reliable input and more rapid retrieval of medical information [3-5]. A systematic review summarized that adopting IT in healthcare services could successfully decrease medication errors, increase adherence to guidelines, and enhance surveillance [6]. While some less recent studies detected medication errors still existing after using automated dispensing system, Beobide Telleria et al found a reduction of 91% in dispensing errors after using an automated packaging system in seven nursing homes in Spain [3,7,8]. The use of automated packaging and dispensing systems has effectively improved safety in the dispensing and administration of solid drugs in nursing homes.

In Hong Kong, people 65 years or older constituted 15.9% of the total population in 2016. Among people in this age group, 8.1% were living in non-domestic households such as one of the 734 licensed OAHs [9,10]. Improving the drug distribution model in OAHs is one of the top priorities in Hong Kong as it is often associated with medication wastage. In most cases, OAH residents have regular medical follow-ups from the hospitals and clinics of the Hospital Authority (HA), which is part of the public health care sector [2]. As the time between each follow-up visit is long [11], 6-month to 1-year supplies of medications are typically dispensed at a single setting to ensure that there is sufficient medication to last until the next follow-up visit. However, as older patients may experience frequent symptom changes that lead to subsequent adjustment or early discontinuation of medications, a large quantity of surplus medication is wasted [11]. One study calculated that the total extrapolated annual cost of medications requiring disposal in all OAHs in Hong Kong was over HK $5,800,000 (US $743,589), including more than 10 million units of oral solid preparations that were estimated to be discarded [11].

Other than being associated with high wastage, the conventional medication preparation process is also labor intensive, time consuming, and error prone. In Hong Kong, OAH nursing staff manually prepare and check all medications, and all medication management activities (medication preparation, checking, and administration) in OAHs are made with reference to and recorded on their in-house paper medication administration records (MARs). This laborious process can result in medication errors and compromise patient safety. In a local survey, 36.8% of the OAHs reported that there was at least one incident of incorrect drug administration over the past 5 years [12]. There were 126 officially reported cases of suspected medication incidents in OAHs from 2008 to 2010 [13].

To address the limitations of the current medication management process in OAHs, the Hong Kong Pharmaceutical Care Foundation Ltd (HKPCF) [14], a local nongovernmental organization, initiated the Integrated Old Age Home Medication Management Program ("the Program") in 2019. Funded by a philanthropic foundation, the Program aims to utilize various technological systems to provide an integrated medication management service to a cluster network of OAHs. This initiative is innovative and, to our knowledge, the first of its kind in Hong Kong.

To date, few studies have systematically evaluated both the efficiency and safety of such comprehensive medication management services. Therefore, we carried out a pilot study to evaluate the impact of the Program on (1) medication management efficiency, (2) medication safety, and (3) drug wastage in OAHs.

In this paper, we described the components of the medication management program and compared the time efficiency and the reductions in medication errors and medication wastage in OAHs before and at least 2 weeks after the implementation of the program. Based on this pilot study, we discussed potential limitations of the program and highlighted directions for future research.

## Methods

### Design and Study Population

The pilot study was a prospective and interventional study that included preintervention and postintervention evaluation of efficiency outcomes and medication error rates. Time-motion methods [15] were used to quantify the number of doses prepared and checked during the medication handling processes. Recruitment of OAHs started in August 2019, and pilot data collection ended in April 2020. Figure 1 summarizes the timeline of Program implementation and data collection for the 2 OAHs that we recruited for this pilot study.

Approval was obtained from the Survey and Behavioural Research Ethics Committee of the Chinese University of Hong Kong (reference number: SBRE-19-106) (Text S1 in Multimedia Appendix 1). We obtained verbal consent from the staff members observed in the OAHs and the HKPCF prior to the data collection period.

**Figure 1.** Timeline of implementation of the Integrated Old Age Home Medication Management Program and study data collection of the two Old Age Homes recruited. eMAR: electronic medication administration record; OAH: old age home; SMMS: SafeMed Medication Management System.
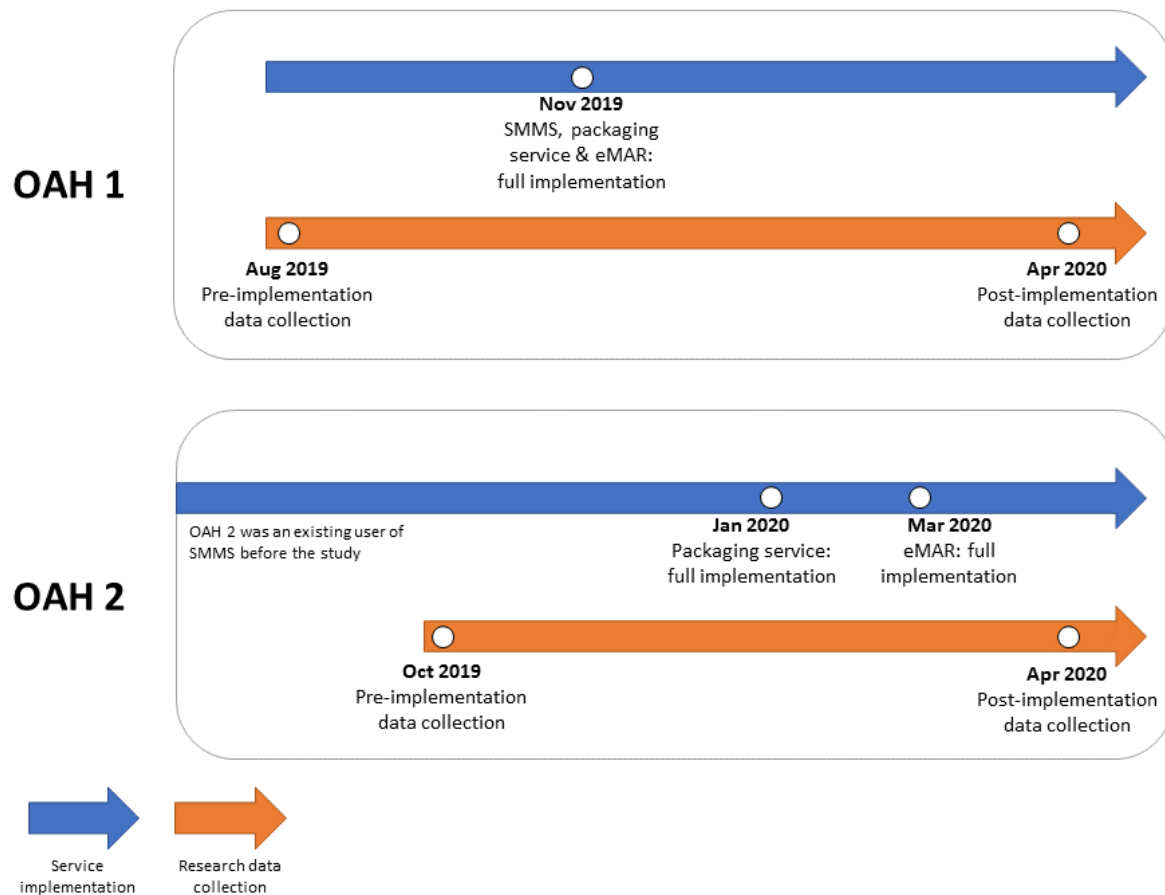


## Intervention

The Program consisted of three components combining elements of IT, automation technology (AT), and the Internet of Things (IoT): (1) electronic medication profiles in the SafeMed Medication Management System (SMMS); (2) a centralized automation-assisted multidose packaging service using the Automated Tablet Dispensing and Packaging System (ATDPS); and (3) an eMAR app on mobile tablet devices. SMMS is an electronic medication management system developed by pharmacists at HKPCF for maintaining updated medication profiles of OAH residents with the support of a comprehensive drug database. ATDPS consists of an automated packaging machine that generates multidose drug pouches containing tablets or capsules to be administered to individual residents during different medication rounds. Automated multidose packaging is done with ATDPS such that medications of a patient at a particular time of administration are packaged in the same drug pouch. Automated packaging, instead of manual packaging, is adopted to diminish human errors under the fatigue of repetitive procedures of dispensing and packaging, hence improving medication safety. All beneficiary OAHs received all three components as the intervention except existing users of the SMMS, for which only (2) and (3) were provided in addition to the SMMS as the intervention.

Figure 2 summarizes the medication management workflow before and after the implementation of the Program. Before implementation of the Program (left), dispensed medications collected from hospitals and clinics are recorded, prepared, checked, and administered manually by OAH nursing staff. After implementation (right), nursing staff create medication information electronically using SMMS. The medications are collected by staff of the centralized dispensing hub for automated packaging and verification. Packaged medications are sent back to OAHs for checking and administration.

Throughout the process of implementing the intervention, HKPCF pharmacists provided training, regular follow-up, and technical assistance for OAH nursing staff to familiarize themselves with the use of the new system. In the postimplementation phase, "medication preparation" of daily medications (including automation-assisted packaging and additional verification of packaging errors) was conducted by HKPCF in a centralized medication management hub ("the hub"). Medications packaged in drug pouches were delivered to the OAHs in batches of 2 to 3 days of medication supply. Paper MARs were replaced by the eMAR app, which possesses features such as bar-coded patient identification and electronic signatures, for "medication checking" and "medication administration." Figure S1 in Multimedia Appendix 1 summarizes the flow of medications and information, and Text S2 in Multimedia Appendix 1 describes in detail the workflow and components of the Program.

**Figure 2.** Medication management process before and after implementation of the Integrated Old Age Home Medication Management Program. AT: automation technology; eMARs: electronic medication administration records; IoT: Internet of Things; IT: information technology; OAH: old age home.
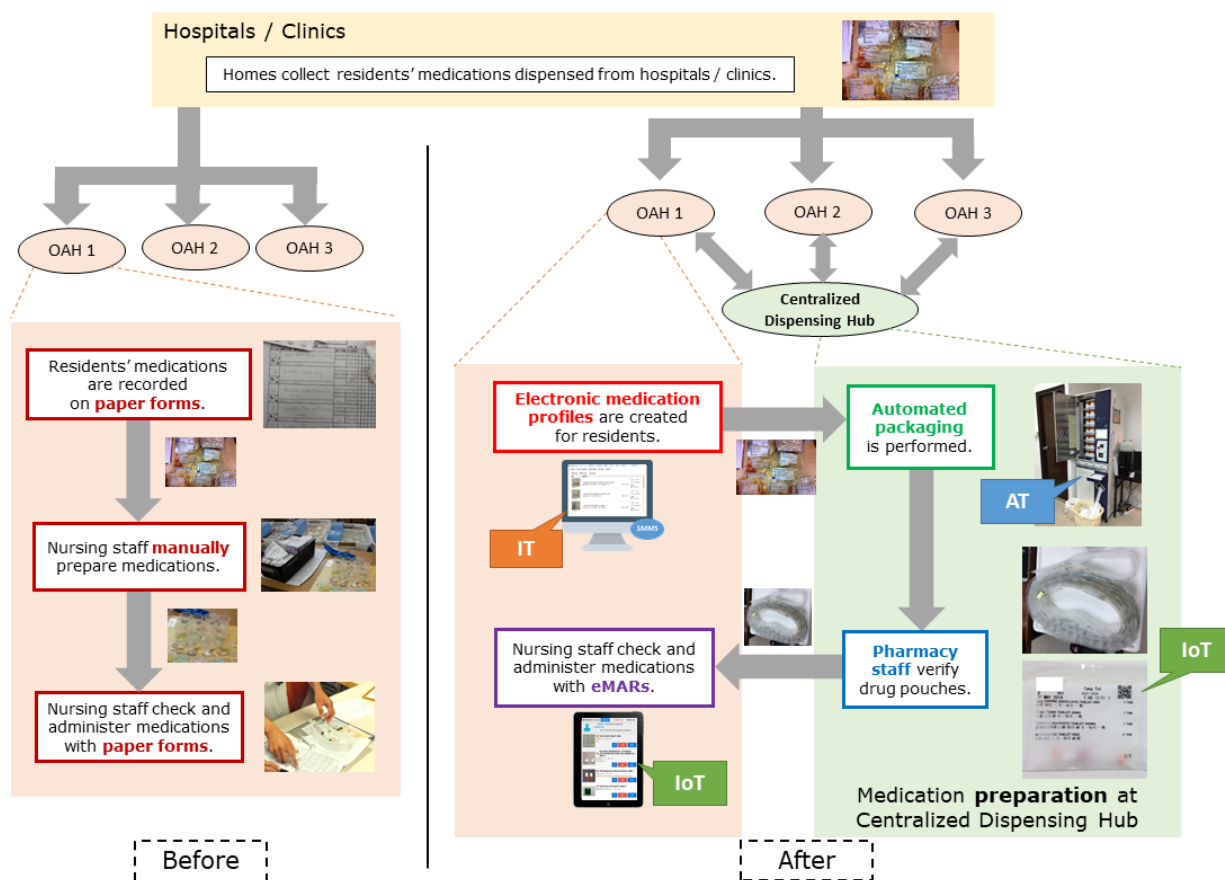


## Primary Outcome: Time Efficiency

The primary outcome of the study was the time efficiency of medication preparation and checking, represented by the mean number of doses prepared and checked in 10-minute blocks. The time-motion approach, which has been widely adopted to evaluate the efficiency of workflow, was used to observe the processes of medication preparation and checking.

During data collection, all staff members performed their duties as usual, and an investigator or a helper of the research team videotaped the processes. An investigator viewed the videos and calculated the number of doses prepared or checked in 10-minute blocks in the preimplementation and postimplementation phases. Time during which staff members were interrupted during the medication management processes was excluded. The postimplementation data were collected after at least 2 weeks of run-in time of full service implementation.

## Secondary Outcomes: Medication Safety and Medication Wastage

The secondary outcomes were the number of medication errors and medication wastage. We evaluated the impact on medication safety by measuring the number of medication errors captured within 5 to 10 days during both the preimplementation and postimplementation phases. In the postimplementation phase, we collected data at least 2 weeks after the full implementation of the Program.

The OAH nursing staff voluntarily reported all medication errors identified during medication checking and administration using an anonymous incident-reporting form in Chinese, modified from the Medication Risk Management Report in *Guidelines on Drug Management in Residential Care Homes 2018* [16]. Errors were calculated as the proportion of problematic doses over the total number of doses processed during the period of data collection from the SMMS.

We retrieved data on medication wastage in January 2020 from the drug disposal reports in the SMMS, which estimated the number of dispensed drugs that were not administered to the patient due to early discontinuation of treatment. Under the centralized automated packaging service of the Program, any unused oral solid medications (pills, tablets, capsules, sustained-release tablets, etc) that are originally dispensed for one patient would be repackaged and supplied to another patient that requires the same medication. Hence, we made a reasonable assumption that wastage could be almost completely eliminated when medications were centralized for redistribution to all residents and that the calculated cost would reflect the cost of wasted medication that could potentially be avoided after the implementation of the Program. In the drug disposal reports, medications were classified according to dosage forms and therapeutic classes in the British National Formulary (BNF) [17]. Based on residents' prescriptions and the unit prices of medications derived from HA data, the quantity and cost of wasted medications in the OAHs were calculated to quantify the scale of medication wastage.

## Data Processing and Statistical Analysis

We used SPSS, version 24 (IBM Corp), as the software for statistical analysis. We summarized the baseline characteristics of each OAH with descriptive statistics. We compared the number of doses prepared and checked in 10-minute blocks in the 2 recruited OAHs before and after the implementation of the intervention using 3-way analysis of variance (ANOVA). By calculating the standard deviation, we observed the variations within samples for medication preparation and checking for each OAH before and after implementation. To detect any significant changes in the number of medication errors, we conducted the Fisher exact test based on the number of correct and incorrect doses before and after implementation. We defined a statistically significant difference as a *P* value of less than .05.

## Results

### Characteristics of OAHs

For this pilot study, we recruited 2 out of the 8 OAHs participating in the Program. Table 1 shows the baseline characteristics of the 2 recruited OAHs (referred as OAH 1 and OAH 2). The age of residents, number of staff members, staffing ratios, and number of medications per resident were similar in both OAHs. OAH 1 was a new user of the SMMS, whereas OAH 2 was an existing user of the SMMS at the time of enrolment into the Program.

**Table 1.** Baseline characteristics of OAHs.

| Characteristics | OAH[a] 1 | OAH 2 |
|---|---|---|
| Care level[b] [18] | Care and attention home for the elderly | Care and attention home for the elderly |
| Mode of operation[c] [18] | Subvented home | Subvented home |
| District | Sham Shui Po | Kowloon City |
| Residents, n | 77 | 101 |
| **Age of residents (years)** | | |
| Mean (SD) | 87.2 (6.9) | 86.4 (8.9) |
| Median (IQR) | 88 (8) | 87 (12) |
| Range | 67-104 | 67-105 |
| **Gender of residents, n (%)** | | |
| Male | 18 (23.4) | 36 (35.6) |
| Female | 59 (76.6) | 65 (64.4) |
| **Staff members, n [18]** | | |
| Total staff members[d] | 52.3 | 52.75 |
| Nurses | 9 | 7 |
| Health workers | 2 | 6 |
| Staff-to-resident ratio | 1:1.5 | 1:1.9 |
| Nursing staff–to-resident ratio[e] | 1:7 | 1:7.8 |
| **Oral medications per resident, n** | | |
| Mean (SD) | 8.3 (3.6) | 9.8 (3.7) |
| Median (IQR) | 8 (5) | 10 (4) |
| Range | 0-21 | 1-21 |
| Status of using SMMS[f] | New user | Existing user |

[a]OAH: old age home.

[b]Based on care level, OAHs in Hong Kong were classified into Homes for the Elderly (for older people with no or mild impairment), Care and Attention Homes for the Elderly (for older people with moderate impairment), and Nursing Homes (for older people with severe impairment).

[c]Based on mode of operation, OAHs in Hong Kong were classified into Subvented Homes, Contract Homes, Non-profit-making Self-financing Homes, and Private Homes.

[d]The number of staff members may not be an integer because some staff members were part-time employees.

[e]Nursing staff refers to nurses and health workers.

[f]SMMS: SafeMed Medication Management System.

## Primary Outcome: Time Efficiency

The mean number of doses prepared in 10-minute blocks for OAH 1 increased from 13.1 (SD 3.7) in the preimplementation phase to 77.1 (SD 9.1) in the postimplementation phase, while for OAH 2, there was an increase from 35.3 (SD 7.7) to 84.8 (SD 15.3) doses. The mean number of doses checked in 10-minute blocks in RCHE 1 increased from 24.7 (SD 12.6) to 67.2 (SD 13.0), but in OAH 2, it decreased from 92.0 (SD 7.1) to 53.1 (SD 33.7) (Table S1 in Multimedia Appendix 1). In other words, the time required to prepare 1000 doses of medications had dropped from 560.6 (SD 320.1) minutes to 126.3 (SD 18.5) minutes (Table S2 in Multimedia Appendix 1).

The results of a 3-way ANOVA (Figures 3 and 4; Tables S3 and S4 in Multimedia Appendix 1) showed the mean number of doses processed in 10 minutes (error bars denote standard error), stratified by (a) preimplementation versus post implementation, (b) OAH 1 versus OAH 2, (c) pre- and post-comparison stratified by OAH, (d) pre- and post- comparison stratified by process (medication preparation/checking), and (e) pre- and post- comparison stratified by both OAH and process. Each of the main effect or interaction terms presented is statistically significant ($P<.001$) in the 3-way ANOVA analysis. Results showed a significant increase ($P<.001$) in the number of doses prepared and checked in 10-minute blocks after the implementation of the Program. These numbers also increased significantly ($P<.001$) in OAH 1 alone during the postimplementation phase. However, fewer doses were processed during checking in OAH 2 postimplementation.

**Figure 3.** Three-way analysis of variance results (considering the effect of one or two factors). OAH: old age home.
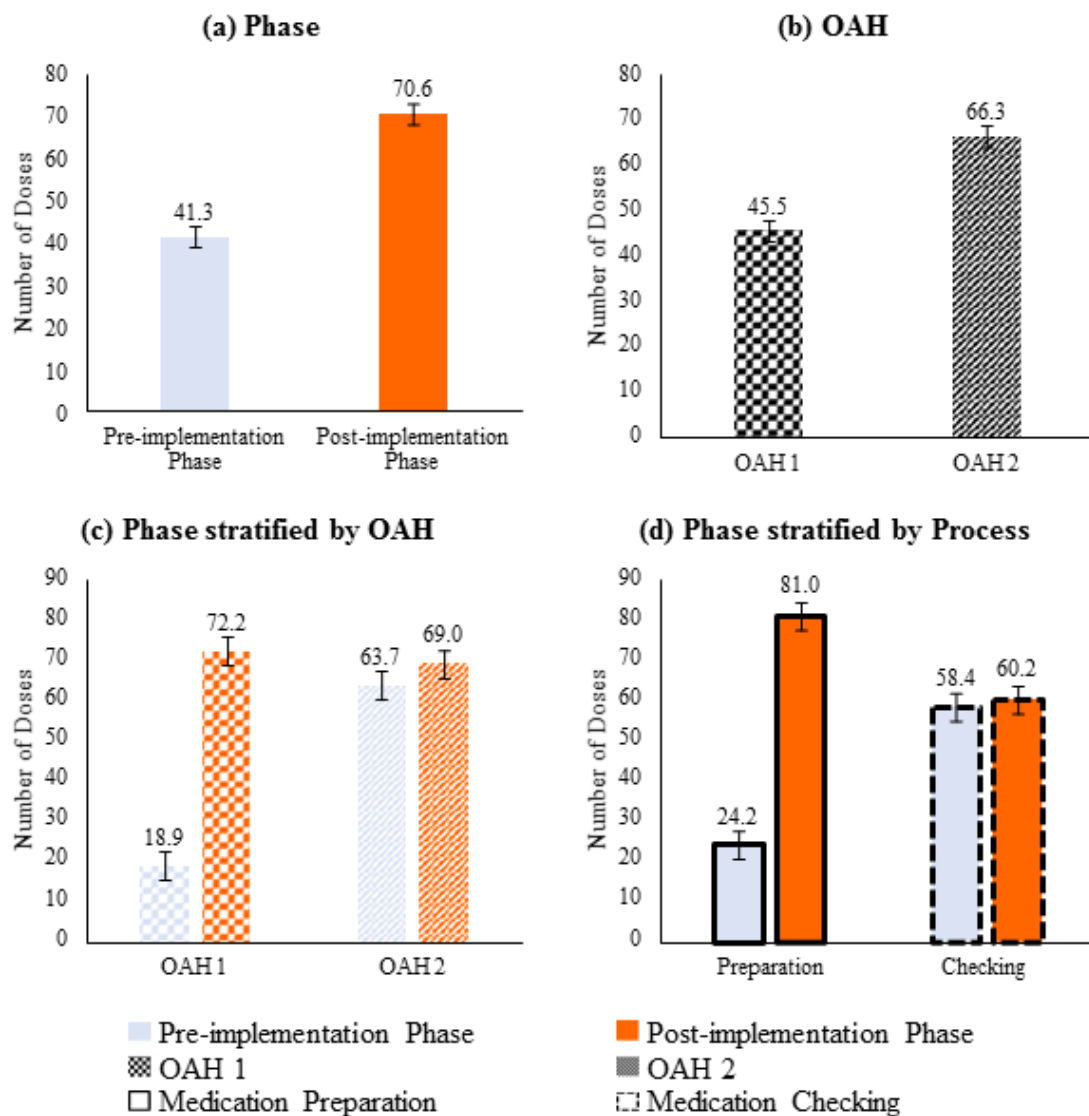
**Figure 4.** Three-way analysis of variance results (considering the effect of three factors). OAH: old age home.
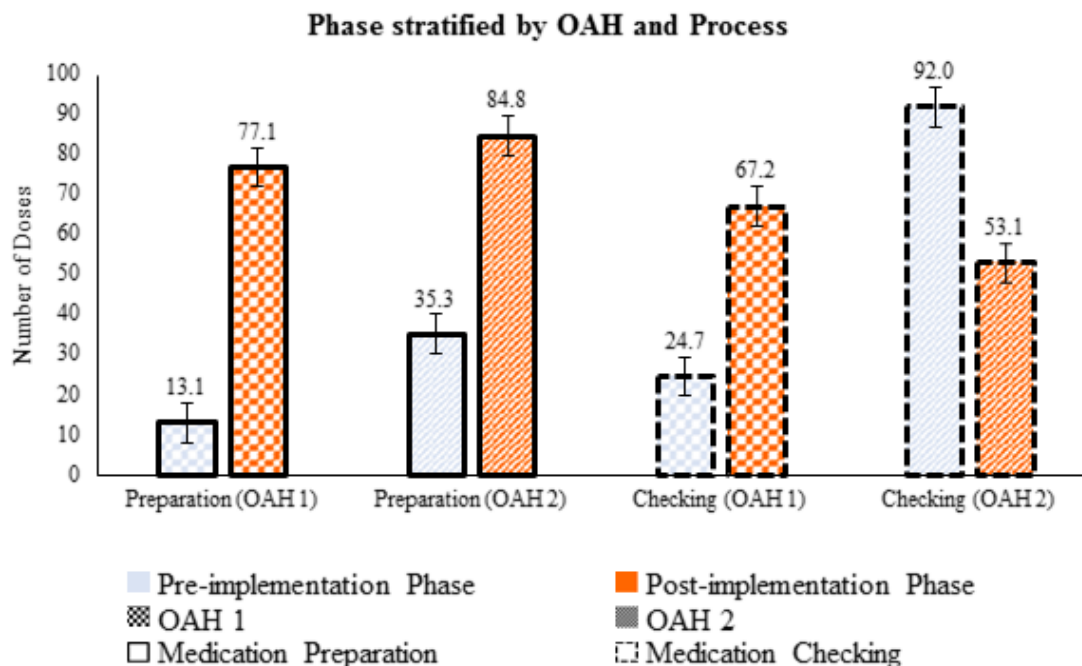


## Secondary Outcome: Medication Errors

Due to time constraints and the expressed preference of OAH 1, we only collected the medication error data from OAH 2. During the data collection period in the preimplementation phase, a total of 9504 doses were handled, among which the nursing staff reported medication errors for 10 doses during medication checking and none during medication administration (Table 2). We classify the errors according to the nature of the errors and the medications involved in Tables S5 and S6 of Multimedia Appendix 1. We emphasize that the OAH nursing

staff corrected all of these errors before medication administration, so they should be considered as "near misses" and did not result in actual errors.

During the data collection period in the postimplementation phase, there were a total of 5731 doses, among which nursing staff reported no medication errors during medication checking and medication administration (Table 2). Comparing the results before and after the intervention, we found a significant reduction ($P$=.02) in the number of medication errors using the Fisher exact test.

**Table 2.** Number of wrong and correct doses in preimplementation and postimplementation phases.

| Phase | Doses in OAH[a] 2, n | | |
|---|---|---|---|
| | Wrong | Correct | Total |
| Preimplementation phase | 10 | 9494 | 9504 |
| Postimplementation phase | 0 | 5731 | 5731 |

[a]OAH: old age home.

## Secondary Outcome: Medication Wastage

In January 2020, the total costs of wasted medications in OAH 1 (77 residents) and OAH 2 (101 residents) were HK $2566.03 (US $328.98) and HK $5249.48 (US $673.01), respectively. We estimated that the annual costs of wasted medications in OAH 1 and OAH 2 were HK $30,792.36 (US $3947.74) and HK $62,993.76 (US $8076.12), respectively, which were equivalent to HK $399.90 (US $51.27) per resident in OAH 1 and HK $623.70 (US $79.96) per resident in OAH 2.

Based on the dosage forms (Table 3), 3199 and 6762 tablets and capsules were wasted in OAH 1 and OAH 2, respectively,

in January 2020. In terms of costs, the greatest proportion of wasted medications in OAH 2 was oral solid medications, constituting 52.8% (HK $2769.41 or US $355.05) of wasted medications. Oral solid medications also contributed to 27.3% (HK $701.51 or US $89.94) of wasted medications in OAH 1.

Based on the therapeutic classes in the BNF (Table S7 in Multimedia Appendix 1), drugs for central nervous system disorders made up the drug class with the highest cost of wasted medications in OAH 1 (HK $1328.24 or US $170.29; 51.8% of the total cost). In OAH 2, the group with the highest cost of wasted medication was drugs for nutrition and blood disorders (HK $1613.70 or US$ 206.88; 30.7% of the total cost).

**Table 3.** Medication wastage based on dosage forms.

| Dosage form (unit) | OAH[a] 1 | | OAH 2 | |
|---|---|---|---|---|
| | n | Cost, HK $ (US $[b]) | n | Cost, HK $ (US $[b]) |
| Oral solid medications (tablets/capsules) | 3199 | 701.51 (89.94) | 6762 | 2769.41 (355.05) |
| Oral liquid medications (mL) | 4355 | 158.13 (20.27) | 3625 | 207.78 (26.64) |
| Inhaled medications (puffs) | 1385 | 279.63 (35.85) | 1076 | 110.94 (14.22) |
| Topical medications (grams) | 0 | 0 (0.00) | 0 | 0 (0.00) |
| Transdermal patches (pieces) | 62 | 1013.95 (129.99) | 0 | 0 (0.00) |
| Suppositories (suppositories) | 72 | 255.71 (32.78) | 65 | 230.85 (29.60) |
| Eye drops (drops) | 342 | 75.19 (9.64) | 1142 | 325.9 (41.78) |
| Sprays (doses) | 208 | 64.35 (8.25) | 0 | 0 (0.00) |
| Mouthwashes (mL) | 300 | 7.48 (0.96) | 0 | 0 (0.00) |
| Insulin injections (insulin units) | 0 | 0 (0.00) | 3004 | 278.05 (35.65) |
| Other parenteral injections (μg) | 3000 | 10.08 (1.29) | 6240 | 1326.55 (170.07) |
| Total | N/A[c] | 2566.03 (328.98) | N/A | 5249.48 (673.01) |

[a]OAH: old age home.

[b]US $1=HK $7.8.

[c]N/A: not applicable.

## *Discussion*

### Principal Findings

To the best of our knowledge, this is one of the first studies to examine the efficiency and safety of a medication management service integrating the use of IT, AT, and the IoT through three components: electronic medication profiles, centralized automated multidose packaging, and eMARs. Our study found that medications were more efficiently prepared in the hub and more efficiently checked in the OAHs after implementation of the Program.

The use of technology can drastically improve the workflow of medication management in OAHs. In a paper-based workflow, OAH staff need to flip through the paper MARs and read the instructions on the medication labels one by one during medication preparation and checking. They also need to manually take out the medication bags one by one and dispense them into suitable containers during medication preparation. With the implementation of eMARs, nursing staff can check the packaged pouches using a dedicated electronic device app showing the patient's medication regimen, with authentic photographs to help them identify the correct medications. The data from the residents' electronic profiles are rapidly sent from SMMS to ATDPS, which can package up to 60 drug pouches per minute [19]. Therefore, the IT- and AT-assisted workflow is undeniably more efficient than the traditional paper-based and manual workflow, as shown by the increased number of doses checked postimplementation in OAH 1. However, we note that the findings for OAH 2 showed otherwise; this may be due to the insufficient run-in time (2 weeks in OAH 2 versus more than 1 month in OAH 1), leading to reduced efficiency during the initial implementation phase of the Program. One comprehensive systematic review on the effect of health

information technology on quality and efficiency of health care highlighted that data on time use and efficiency metrics were inconsistent [6]. The impact of automation on improving time efficiency requires long-term prospective evaluation.

The traditional mode of operation for medication management in OAHs puts a disproportionately heavy burden on nursing home staff. In Hong Kong, there is neither a statutory nor an administrative requirement to have any pharmacists or dispensers (pharmacy technicians) in OAHs [20]. To save costs, Health Workers (registered non-nurse OAH employees who have attended an approved short training course) [20] are frequently assigned as the main personnel for handling medications despite their inadequate training for this responsibility. Therefore, nursing staff often spend considerable effort and time spotting and preventing medication errors. After the implementation of the Program, the responsibilities for medication preparation and verification were transferred to pharmacists and dispensers in the hub. They are professionally trained in the handling of medications, and their expertise may have contributed to the improvement in efficiency, as well as the reduction in medication errors.

The Program also enables more effective use of manpower in OAHs for direct patient care services. Conventional time-consuming medication management activities place a huge burden on nursing staff members, who are also expected to provide other care services for residents. In a study in Australia, medication preparation and administration constituted approximately one-quarter of nurses' 8-hour shifts in OAHs [21]. Upon implementation of the Program, the nursing staff in this study were completely relieved of their medication preparation duties. Therefore, the Program helped the OAH nursing staff to spare more time for the provision of

higher-quality, patient-oriented care services for residents, in turn improving the residents' quality of life.

The reported medication error rate of 0.1% was already quite low in the preimplementation phase. However, in the postimplementation phase, the Program further significantly reduced the error rate to zero. This result matched with the study of Beobide Tellería et al, which found a reduction of 91% in dispensing errors after using ATDPS in 7 OAHs in Spain [3]. Several features of the Program may have contributed to the reduced rate of medication errors. First, the orderly arrangement of electronic medication profiles in the SMMS reduces transcription errors. Second, the assistance of the automated ATDPS, together with its unique features such as using a photoelectric sensor to ensure that medications are correctly sorted into the pouches, can reduce fatigue from repetitive processes that were traditionally carried out manually, thus improving accuracy. Finally, value-added services in the hub can also help reduce errors. Pharmacists apply their professional expertise to check for medication-related problems, such as the duplication of medication therapies, incorrect dosages, potential drug-drug interactions, and inappropriate long-term use of medications, upon receipt of new prescriptions in the SMMS and medications from OAHs. The pharmacists and dispensers can also further verify the contents of the drug pouches before delivery to the OAHs to ensure the accuracy of the medications packaged by the ATDPS. This multipronged approach provides additional layers of safety nets to reduce medication errors to a minimum and improve medication safety for the residents of OAHs.

We observed a considerable quantity of wasted medication in the OAHs. This problem not only increases expenditure in the public health care system, but also poses a threat to environmental safety from a public health perspective. The extra quantity of medication also puts pressure on the storage capacity of OAHs, as space is scarce in Hong Kong. Oral solid medications accounted for a particularly large proportion of wastage. Under the centralized automated packaging service of the Program, medication wastage could thus be potentially reduced, as unused oral solid medications originally dispensed for one patient can now be packaged and supplied to another patient who requires the same medication. Our pilot data suggest that technology may help to reduce the cost associated with wastage. Future work should be dedicated to evaluating the tangible benefits of the Program.

Despite the potential benefits demonstrated in this pilot study, the implementation of the Program was not without challenges. First, the mentality and learning capacity of OAH staff members were major barriers, as the Program brought a change in their usual workflow. Nursing staff might prefer to adhere to their familiar routine. A study showed that around 40% of nursing staff reported difficulty in making changes for improvement [22]. Many OAH nursing staff are also older and might be less familiar with the technologies used in the Program. Adequate training, on-site support, sufficient run-in time, changes in management, and leadership from the senior management of the OAHs are needed to help nursing staff adapt to the new service model in the Program.

Financial sustainability is also a major challenge in continuing the Program. Without extra funding support after the end of the Program, the centralized packaging service alone would cost HK $300 (US $38.5) per resident per month. None of the participating OAHs were willing to pay for such a service out of pocket due to tight budget constraints. Specifically, we need to identify strategies to fund the Program by harnessing the actual savings from reduced medication wastage, reduced dispensing and preparation work, and the elimination of certain redundant steps in the medication management process. For example, dispensing is done in both HA pharmacies and the hub (ie, it is duplicated). To streamline the current model, we propose that medications should be provided by the HA in bulk directly to the hub instead of dispensed individually to each patient at each medical follow-up. We anticipate that in the long run, this proposed model will make more cost-effective use of human and other resources, while fully unlocking the potential savings of unused medications. We will also negotiate with relevant stakeholders (including the HA and Social Welfare Department) and policy makers to advocate a more streamlined service model.

The results of this pilot study should be interpreted in light of a few important limitations. Compared with the study of Beobide Tellería et al [3], which recruited 7 OAHs, the inclusion of only 2 OAHs in this study, both of the same type, may not fully reflect the effect of the Program on other OAHs of different natures and types. The duration of data collection was relatively short and may not have sufficiently covered the variations in time efficiency and medication errors on different days. The 2-week run-in time of implementation before data collection was apparently not sufficient for the staff of OAH 2 to familiarize themselves with the new workflow. The COVID-19 pandemic caused delays to the implementation of the Program and data collection, limiting the duration and extent of the study. However, despite the modest sample size and short observation period, the pilot data demonstrated statistically significant positive outcomes.

In the time-motion study, there may have been social desirability bias, as the OAH staff might have felt scrutinized under the camera and behaved differently from usual. To reduce the pressure felt by OAH staff during data collection, only 10-minute excerpts were recorded, which may not have fully represented the time efficiency throughout the entire process. The interoperation of medication safety outcomes may have been limited by the small number of doses and underreporting of medication errors. Especially in the preimplementation phase, the OAH staff may have felt reluctant to report errors due to already heavy workloads or fear of negative consequences. Lastly, the drug disposal reports were based on residents' prescription data entered in the SMMS, implying that the actual quantity of medication administered to patients may not have been completely accurate.

The Program consists of combining the three distinct components of SMMS, ATDPS, and eMAR. Hence, we are unable to differentiate the effect of each individual intervention in this pilot study. However, we can potentially observe differences in outcomes among other OAHs at different implementation phases of the program. The study only focused

on the measurable aspects of medication preparation and checking, which may not have fully reflected the less quantifiable effects of the Program. It also did not evaluate the effect of the Program on medication administration, as it was not appropriate to videotape residents during administration and no attempt was made to detect administration errors. However, we anticipate that the use of eMARs probably enhanced the safety and efficiency of medication administration. The impact of other features of the intervention (Table S8 in Multimedia Appendix 1) on improving the conventional medication management process should be evaluated in future studies.

## Future Work

Our future work includes recruiting a wider variety of OAHs in terms of care level, staff ratio, occupancy rate, and mode of operation. This would allow us to delineate whether the complexity of residents' medications and the working practices of nursing staff can influence the outcomes of the Program. The unmeasurable impact of the Program would be evaluated through qualitative means, such as focus group discussions and structured interviews, to gather the perspectives of OAH staff on the user-friendliness of the Program. To capture the changes in OAH staff's roles and responsibilities more comprehensively, we would explore how time that was relieved from medication packing and checking is redistributed to fulfil other nursing tasks during the postimplementation phase. A cost-benefit analysis would also be conducted to determine the best approach to implement this Program in a resource-limited setting. The eventual goal is to explore the prospects of adapting and applying this service model on other settings that involve complicated medication management, such as rehabilitation centers, hospices, schools for students with special needs, and other types of long-term care facilities.

## Conclusion

An innovative medication management program with IT, AT, and IoT components improved the time efficiency of medication preparation and medication safety for OAHs. With streamlining of the workflow in the future, the Program might also reduce medication wastage. Therefore, the Program is a promising solution to address the current limitations in medication management in OAHs in Hong Kong. Our future work includes validating these findings prospectively in a larger sample of OAHs over a longer time horizon. Both the tangible and intangible impact of the Program on improving patient safety, efficiency, and staff satisfaction would be explored. From a service perspective, negotiation with relevant stakeholders and policy makers would be targeted at advocating a more streamlined service model so that OAHs can achieve safer and more efficient medication management service.

## Authors' Contributions

KHS and CWT made equal contributions as co-first authors. SCC and YTC made equal contributions as senior authors. All authors contributed to the study concept and design and interpreted the data. KHS, CWT, and SCC acquired subjects and data. KHS, CWT, TTNL, and YTC analyzed data. KHS and CWT prepared the manuscript. All authors revised the manuscript and approved the final version. SCC solicited funding for the project.
The sponsor of the Program, the Li Ka Shing Foundation, had no role in the data collection, analysis, or reporting of the study.

## Conflicts of Interest

KHS is employed as pharmacist by Hong Kong Pharmaceutical Care Foundation Ltd (HKPCF), which is responsible for development and implementation of the Integrated Old Age Home Medication Management Program and its components. SCC is the Director of Hong Kong Pharmaceutical Care Foundation Ltd (HKPCF). The authors do not have financial competing interest. All other authors declared no conflicts.

Multimedia Appendix 1
Supplementary Materials.
[DOCX File , 429 KB - medinform_v9i2e24280_app1.docx ]

## References

1.      Kaneko S, Okada M. Medication management factors associated with medication errors at Japanese long-term care facilities. J Appl Med Sci 2018;7(1):1-10 [FREE Full text]

2.      Wong AY, Chan PW, Chan EW, Cheng M, Wong IC. The Pharmaceutical Services to the Elderly in the Old Aged Homes in Hong Kong: A Scoping Exercise. Hong Kong Pharm J 2013 Aug 20;20(1):15 [FREE Full text]

3.      Beobide Telleria I, Ferro Uriguen A, Miró Isasi B, Martínez Arrechea S, Genua Goena MI. The impact of automation on the safety of drug dispensing in nursing homes. Farm Hosp 2018 Jul 01;42(4):141-146 [FREE Full text] [doi: 10.7399/fh.10949] [Medline: 29959837]

4.      Alenius M, Graf P. Use of Electronic Medication Administration Records to Reduce Perceived Stress and Risk of Medication Errors in Nursing Homes. Comput Inform Nurs 2016 Jul;34(7):297-302. [doi: 10.1097/CIN.0000000000000245] [Medline: 27270628]

5.      Cousein E, Mareville J, Lerooy A, Caillau A, Labreuche J, Dambre D, et al. Effect of automated drug distribution systems on medication error rates in a short-stay geriatric unit. J Eval Clin Pract 2014 Oct;20(5):678-684 [FREE Full text] [doi: 10.1111/jep.12202] [Medline: 24917185]

6.      Chaudhry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, et al. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Ann Intern Med 2006 May 16;144(10):742-752. [Medline: 16702590]

7.      Baril C, Gascon V, St-Pierre L, Lagacé D. Technology and medication errors: impact in nursing homes. Int J Health Care Qual Assur 2014;27(3):244-258. [doi: 10.1108/IJHCQA-03-2013-0029] [Medline: 25786188]

8.      van den Bemt PMLA, Idzinga JC, Robertz H, Kormelink DG, Pels N. Medication administration errors in nursing homes using an automated medication dispensing system. J Am Med Inform Assoc 2009;16(4):486-492 [FREE Full text] [doi: 10.1197/jamia.M2959] [Medline: 19390109]

9.      Census and Statistics Department. 2016 Population By-census Thematic Report: Older Persons. 2018. URL: https://www.statistics.gov.hk/pub/B11201052016XXXXB0100.pdf [accessed 2021-01-20]

10.     Social Welfare Department. Social Welfare Services in Figures–2018 Edition. 2018. URL: https://www.swd.gov.hk/storage/asset/section/296/tc/swdfig2018(Fast_web_view).pdf [accessed 2020-09-12]

11.     Wong AY, Chan PW, Chan EW, Cheng M, Wong IC. Drug Wastage among the Elderly Living in Old Aged Homes in Hong Kong. Hong Kong Pharm J 2013 May 25;20(1):16-19 [FREE Full text]

12.     Leung EMF. The Project on Accreditation System for Residential Care Services for the Elders in Hong Kong - Five-year review report (2005-2010). Hong Kong: The Hong Kong Association of Gerontology; 2012.

13.     Wong A, Chan P, Chan E, Lam M, Wong I. Pharmaceutical Services of Elderly Patients in Hong Kong: Scoping study to identify and analyse the nature of pharmaceutical services provided to the elderly in Hong Kong - Report to the Pharmaceutical Society of Hong Kong. 2013 Jan 01. URL: http://www.ps.org.hk/uploads/files/Pharmaceutical%20Services%20of%20Elderly%20Patients%20in%20Hong%20Kong.pdf [accessed 2020-09-12]

14.     Hong Kong Pharmaceutical Care Foundation. About Us - Hong Kong Pharmaceutical Care Foundation. URL: http://www.pcfhk.org/en/service [accessed 2020-09-12]

15.     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 01;49:292-299 [FREE Full text] [doi: 10.1016/j.jbi.2014.02.017] [Medline: 24607863]

16.     Social Welfare Department. Guidelines on Drug Management in Residential Care Homes 2018. 2018. URL: https://www.swd.gov.hk/storage/asset/section/2984/en/DrugMgtGuideline2018_Chpt1_2_3.pdf [accessed 2020-09-12]

17.     Joint Formulary Committee. BNF 73 (British National Formulary) March 2017. United Kingdom: The Pharmaceutical Press; 2017.

18.     Social Welfare Department. Search for Residential Care Homes for the Elderly. 2019. URL: https://elderlyinfo.swd.gov.hk/en/ltc_search/rcs [accessed 2020-04-16]

19.     Yuyama Co, Ltd. Litrea III. URL: https://www.yuyamarx.com/wp-content/uploads/2018/07/Litrea3.pdf [accessed 2020-09-12]

20.     Social Welfare Department. Code of Practice for Residential Care Homes (Elderly Persons) March 2013 (Revised Edition). 2015. URL: https://www.swd.gov.hk/doc/LRB/LORCHE/CodeofPractice_E_201303_201503R5.pdf [accessed 2020-09-12]

21.     Qian S, Yu P, Hailey D. Nursing staff work patterns in a residential aged care home: a time-motion study. Aust Health Rev 2016 Nov;40(5):544-554. [doi: 10.1071/AH15126] [Medline: 26615222]

22.     Hughes CM, Lapane KL. Nurses' and nursing assistants' perceptions of patient safety culture in nursing homes. Int J Qual Health Care 2006 Aug;18(4):281-286. [doi: 10.1093/intqhc/mzl020] [Medline: 16855298]

## Abbreviations

**ANOVA:** analysis of variance
**AT:** automation technology
**ATDPS:** Automated Tablet Dispensing and Packaging System
**BNF:** British National Formulary
**eMAR:** electronic medication administration record
**HA:** Hospital Authority
**HKPCF:** Hong Kong Pharmaceutical Care Foundation Ltd
**IoT:** Internet of Things

**IT:** information technology
**MAR:** medication administration record
**OAH:** old age home
**SMMS:** SafeMed Medication Management System

Original Paper

# Adapting Bidirectional Encoder Representations from Transformers (BERT) to Assess Clinical Semantic Textual Similarity: Algorithm Development and Validation Study

Klaus Kades[1,2*], MSc; Jan Sellner[1,3*], MSc; Gregor Koehler[1], MSc; Peter M Full[1,4], BSc; T Y Emmy Lai[1,5], MSc; Jens Kleesiek[1,2,3,6*], MD, PhD; Klaus H Maier-Hein[1,2,3,4*], PhD

[1]German Cancer Research Center (DKFZ), Heidelberg, Germany

[2]Partner Site Heidelberg, German Cancer Consortium (DKTK), Heidelberg, Germany

[3]Helmholtz Information and Data Science School for Health, Karlsruhe/Heidelberg, Germany

[4]Heidelberg University, Heidelberg, Germany

[5]Hochschule Mannheim, University of Applied Sciences, Mannheim, Germany

[6]Institute for Artificial Intelligence in Medicine (IKIM), University Medicine Essen, Essen, Germany

[*]these authors contributed equally

**Corresponding Author:**
Klaus Kades, MSc
German Cancer Research Center (DKFZ)
Im Neuenheimer Feld 280
Heidelberg, 69120
Germany
Phone: 49 6221420
Email: k.kades@dkfz.de

## *Abstract*

**Background:** Natural Language Understanding enables automatic extraction of relevant information from clinical text data, which are acquired every day in hospitals. In 2018, the language model Bidirectional Encoder Representations from Transformers (BERT) was introduced, generating new state-of-the-art results on several downstream tasks. The National NLP Clinical Challenges (n2c2) is an initiative that strives to tackle such downstream tasks on domain-specific clinical data. In this paper, we present the results of our participation in the 2019 n2c2 and related work completed thereafter.

**Objective:** The objective of this study was to optimally leverage BERT for the task of assessing the semantic textual similarity of clinical text data.

**Methods:** We used BERT as an initial baseline and analyzed the results, which we used as a starting point to develop 3 different approaches where we (1) added additional, handcrafted sentence similarity features to the classifier token of BERT and combined the results with more features in multiple regression estimators, (2) incorporated a built-in ensembling method, *M*-Heads, into BERT by duplicating the regression head and applying an adapted training strategy to facilitate the focus of the heads on different input patterns of the medical sentences, and (3) developed a graph-based similarity approach for medications, which allows extrapolating similarities across known entities from the training set. The approaches were evaluated with the Pearson correlation coefficient between the predicted scores and ground truth of the official training and test dataset.

**Results:** We improved the performance of BERT on the test dataset from a Pearson correlation coefficient of 0.859 to 0.883 using a combination of the M-Heads method and the graph-based similarity approach. We also show differences between the test and training dataset and how the two datasets influenced the results.

**Conclusions:** We found that using a graph-based similarity approach has the potential to extrapolate domain specific knowledge to unseen sentences. We observed that it is easily possible to obtain deceptive results from the test dataset, especially when the distribution of the data samples is different between training and test datasets.

*(JMIR Med Inform 2021;9(2):e22795)* doi:10.2196/22795

XSL•FO
**RenderX**

# Introduction

Every day, hospitals acquire large amounts of textual data which contain valuable information for medical decision processes, research projects, and many other medical applications [1]. However, the huge quantity of reports is unsuitable for manual examination, and automatic access is hindered by the unstructured nature of the data [2]. Natural Language Understanding can help to tackle this problem by automatically extracting relevant information from textual data [3,4]. In this paper, we will focus on a subtask of Natural Language Understanding called Semantic Textual Similarity, which evolved within Natural Language Understanding as a dedicated research question aiming to address tasks like question answering, semantic information retrieval, and text summarization [5-9].

In the clinical domain, Semantic Textual Similarity has the potential to ease clinical decision processes (eg, by highlighting crucial text snippets in a report), query databases for similar reports, assess the quality of reports, or be used in question answering applications [1]. Furthermore, clinical reports are often of poor quality due to time limitations or due to the fact that many text snippets are simply copy-pasted from other reports [10,11]. This introduces low-quality data samples that make it harder for Natural Language Understanding algorithms to extract relevant information. In this context, Semantic Textual Similarity can be a key processing step when dealing with redundant text snippets [2].

State-of-the-art Natural Language Processing (NLP) methods for assessing the Semantic Textual Similarity of nonclinical data are developed and benchmarked based on the Semantic Textual Similarity benchmark, which compromises the SemEval Semantic Textual Similarity tasks from 2012 to 2017 [5] and is part of the General Language Understanding Evaluation dataset. In order to strengthen the development of Natural Language Processing tools for clinical and biomedical text data, which are often not publicly available, the team of the National NLP Clinical Challenges (n2c2), formerly known as i2b2 NLP Shared Tasks, has issued several tasks and organized challenges since 2006. This paper reports our participation in track 1, "n2c2/OHNLP Track on Clinical Semantic Textual Similarity," of the 2019 National NLP Clinical Challenges. We present the 3 submitted systems, a further best performing variation of the different approaches, and a statistical analysis of the dataset. The aim of the track in which we participated was to predict the Semantic Textual Similarity between two clinical sentences. A similar task was already tackled in the BioCreative/OHNLP 2018 ClinicalSTS track [3,4].

The winners of Track 1: 'n2c2/OHNLP 2018 Track on Clinical Semantic Textual Similarity' [3] proposed 4 systems that combined string, entity, and number similarity features with deep learning features. In their best performing system, the winners trained a ridge regression model based on the prediction score of 8 independently trained models [3,12]. The second best performing team proposed an approach using Attention-Based Convolutional Neural Networks and Bidirectional Long Short Term Memory networks [3].

In recent years, the general Natural Language Processing domain made a huge step forward with the breakthrough of transfer learning which allows leveraging semantic knowledge from huge amounts of unlabeled text data. That is, a model can be pretrained on enormous unlabeled text data with multiple unsupervised tasks. The trained model captures a universal language representation and can be effectively fine-tuned on different downstream tasks. For example, the 2018 language model, Bidirectional Encoder Representations from Transformers (BERT), introduced a multilayer bidirectional Transformer that is trained on a massive amount of text in two unsupervised tasks: (1) next sentence prediction and (2) masked word prediction. To use the model for further downstream tasks, it is usually enough to add a linear layer on top of the pretrained model to achieve state-of-the-art performance for the desired downstream tasks [13]. Since the introduction of the Transformer and BERT, new variations of the original models perform even better by (1) introducing more pretraining tasks [14-16], (2) employing multitask learning approaches [17], and (3) combining the aforementioned approaches [18].

The application of pretrained models like BERT on clinical data comes with the question if the model can handle domain-specific nuances. One proposed approach to handling domain-specific nuances is to use transfer learning to adapt the model to clinical data [19-21]. Another approach is to incorporate already existing methods. To investigate the extent to which BERT can handle domain-specific nuances, we examined the performance of BERT on subgroups of sentences and found that it performed modestly with sentence types that were simply structured and highly specific (eg, sentences which prescribe medications). Based on these findings, we created 3 approaches with the aim to address the diversity of clinical sentences present in the given data.

To summarize our contributions (see Figure 1), we show that the use of BERT on clinical data can be enriched by the following:

- a simple modification of the BERT architecture by adding additional similarity features and employing a built-in ensembling method.
- a graph-based similarity approach for a subset of structured sentences in which the knowledge of the training set is extrapolated to unseen sentence pairs of the test set.

Additionally, we show that statistically analyzing the data reveals differences between the training and test datasets. This analysis made the process of interpreting the results easier.

The code to reproduce the results of this paper is available online [22].

**Figure 1.** Overview of our pipeline for the different approaches. Blue boxes denote feature sets (see Multimedia Appendix 3) or scores which are also used as features, and the framed boxes denote processing steps. Bold scores correspond to our three submissions. [CLS] represents a classifier token which is used by BERT for sentence level downstream tasks [13]. The medication graph works only on a subset of the data with input from the Voting Regression. New scores are predicted and replaced with other scores only for this subset. BERT: Bidirectional Encoder Representations from Transformers.
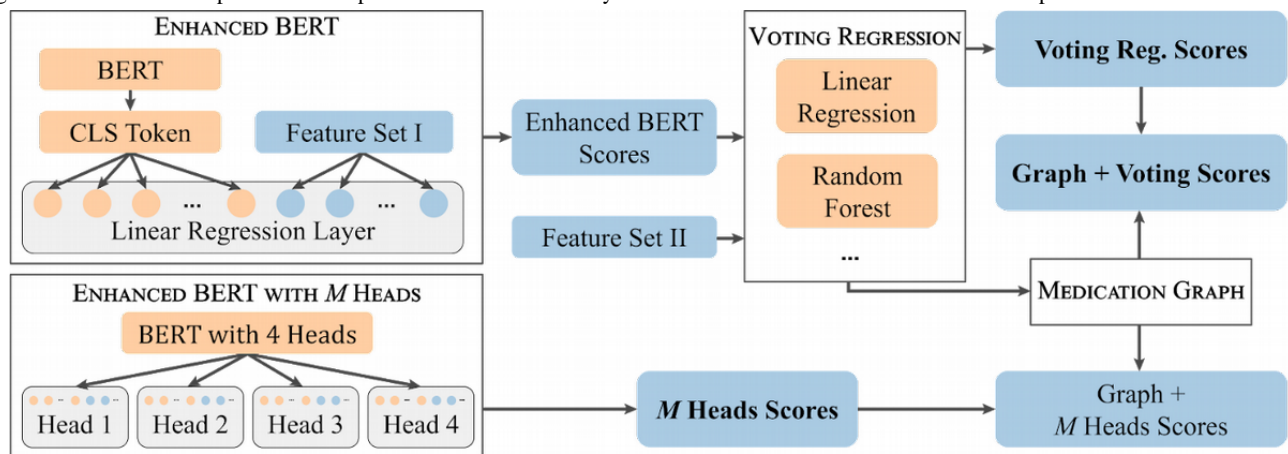


## Methods

### Overview

Our methods were developed and tested on the data of the ClinicalSTS shared task, which consists of a collection of electronic health records from the Mayo Clinic's dataset [4]. In total, 2054 sentence pairs were independently annotated to their degree of semantic textual similarity on a scale from 0 (not similar at all) to 5 (completely similar) by two medical experts. The focus of semantic textual similarity is whether two sentences have similar meaning and content in contrast to, for example, the number of words used in both sentences [3,4]. The created annotations are a mixture of integer and noninteger values, whereby the latter arise when averaging the result of multiple annotations. The training set consists of 1642 sentence pairs and the test set of 412. The performance of the different methods is measured by the Pearson correlation coefficient which aims to measure the linear correlation between the predicted similarity scores and the annotated similarity scores. More detailed information about the creation of the dataset, its properties, and its evaluation can be found elsewhere [1,3,4].

We started by applying ClinicalBERT [23] to the dataset to obtain a baseline. That is, we used the [CLS] token from the last layer of the BERT model and fed its values to an additional linear layer that consists of a single neuron performing the similarity regression task. The whole network, including BERT and the additional layer, were trained on the Mean Square Error.

We use the [CLS] token because it is designed for sentence classification and regression tasks. During training of BERT, [CLS] tokens are used for the next sentence prediction task. The [CLS] token is part of every sentence pair and captures the aggregated attention weights from each token of the sentence pair [13]. Next, we analyzed the predictions of BERT to find sentences with a high deviation from the ground truth. For this, we extracted InferSent embeddings for each sentence pair, as they are suited to cover the semantic representation of sentences [24], clustered them via -means, and calculated the absolute difference between the BERT scores and the ground truth for the whole cluster. The cluster analysis is shown in Figure 2. To make the comparison between the clusters easier, we show an overview of the absolute score differences per cluster in Figure 3. From these visualizations, we see that cluster 3 has the highest difference on average or, in other words, that BERT cannot handle these sentences well. Looking at the sentences, we see that this cluster is dominated by sentences which prescribe medication, for example "ondansetron [ZOFRAN] 4 mg tablet 1 tablet by mouth three times a day as needed" or "furosemide [LASIX] 40 mg tablet 1 tablet by mouth two times a day." This weakness was essentially the motivation for our third approach (medication graph) which focuses solely on the medication sentence type.

In the following section, we describe the approaches shown in our pipeline (Figure 1) in more detail. Information about the preprocessing steps and further implementation details can be found in Multimedia Appendix 1 and Multimedia Appendix 2.

**Figure 2.** Clustering of the sentences to reveal BERT's weaknesses. Each point represents a sentence pair from the training set, and the corresponding absolute score difference is visualized as opacity, or, in other words, the more opaque a point is, the higher the deviation from the ground truth. The points are the t-SNE projected InferSent embeddings of all sentence pairs. For each cluster, the average absolute deviation from the ground truth as well as the distribution of the differences is shown in the legend. Best viewed in colour. BERT: Bidirectional Encoder Representations from Transformers. t-SNE: t-distributed stochastic neighbor embedding.
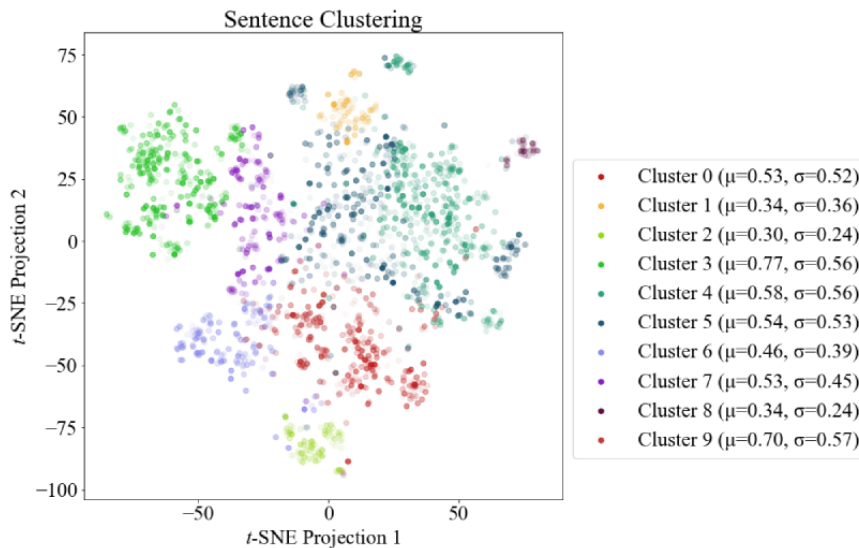


**Figure 3.** Box plot showing the absolute score differences for each cluster emphasizing the opacity information from Figure 2. The number below the bold cluster index is the cluster size. For each box plot, the following information is depicted: the box ranges from the lower to the upper quartiles with the notch at the median position. The whiskers extend up to 1.5 times the interquartile range. Remaining points (outliers) are not shown. The white square denotes the mean value.
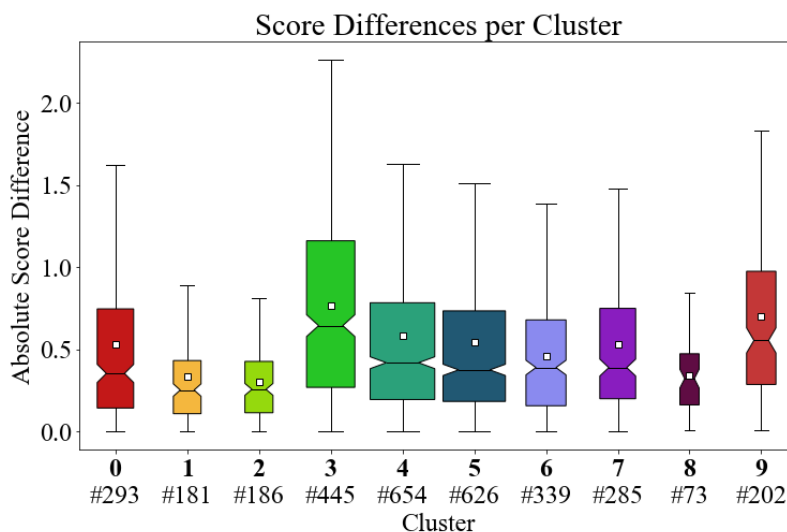


## Approaches

### Approach 1: Enhancing BERT With Features Based on Similarity Measures

The motivation behind this approach is to enhance BERT with additional information that BERT might not be able to capture in its model. On a token level, BERT uses a predefined tokenizer based on a set of rules; however, it might be valuable to compare arbitrary tokens based on character -grams. On a sentence-level, BERT does have a classifier token, [CLS], to compare two sentences. However, the [CLS] token was not designed to be a sentence-embedding [25,26]. Therefore, comparing embeddings like InferSent. which are specifically designed to represent the semantic of a whole sentence, might add additional valuable information to predict the similarity between two sentences.

In this approach, we used two kinds of similarity measures: (1) token-based and (2) sentence embedding–based. For a token-based similarity measure, -grams of characters are created and then compared with each other. For example, Jaccard Similarity compares the proportion between the intersection and the union of -grams in two input sentences. For a sentence embedding-based similarity measure, the embeddings of two sentences are compared, for example by taking the cosine similarity between the two embeddings of the two input sentences. The similarity measures were inspired by Chen et al [12].

We combined BERT with two feature sets of similarity measures at two different positions in our pipeline (Figure 1). In a first step named Enhanced BERT, we enhanced the [CLS] token of BERT with similarity measures from the first feature set (Feature

Set I) before feeding the concatenated vector to the final linear regression layer. In a second step named Voting Regression, we fed the predicted output scores from Enhanced BERT together with a second feature set (Feature Set II) into several estimators (see Multimedia Appendix 3) whose predicted output scores were averaged with the help of a voting regressor [27]. Feature Sets I and II were created by successively trying out different combinations of similarity features in order to gauge the best performance. A breakdown of the two features sets can be found in Multimedia Appendix 3.

### *Approach 2: M-Heads*

Ensembling methods have been very popular in recent machine learning challenges [28,29]. The general approach is to duplicate a model or parts of a model, repeat the prediction for each model, and aggregate the prediction results. The intuition is that different models can focus on various aspects of the input data (eg, different sentence types) so that they produce different predictions. The aggregation over these predictions can help to emphasize the group opinion over the dominance of a single model, thereby mitigating the risk that a model just reacted to noise in the input data [30].

We took up this point and decided to include a simple ensembling method directly into the architecture of BERT. More concretely, we duplicated the final linear layer (the head) which receives the last [CLS] token from the BERT model and which is responsible for calculating the regression (score prediction). We initialized each head layer with different weights to allow the different solutions per head. We employed a loss scaling which enforces specialization of the different heads similar to methods seen in other research [31,32]. A detailed description of our *M*-Heads updated scheme during training and how we performed predictions on new samples can be found in the Multimedia Appendix 4.

### *Approach 3: Medication Graph*

In this approach, we focused on a subset of the sentence pairs which we named "medication sentences," for example "ibuprofen 150 mg tablet 2 tablets by mouth every 7 hours as needed." Further examples are listed in the discussion. These sentences are fairly structured and can be compared by analyzing individual entities. We used the MedEx-UIMA system [33,34] to extract medication related fields from the sentences and decided to use the entity's active agent ("ibuprofen"), strength ("150 mg"), dose ("2 tablets"), and frequency ("7"). We considered the active agent as the major contributing factor in terms of the similarity of medication sentences. Hence, we modeled similarities between active agents that were then further modified by the remaining entities to retrieve a similarity score for each medication sentence pair.

Our general idea was to determine the property of similarity between active agent pairs as compared to unknown active agent pairs. That is, we assumed that the similarity of active agents A and B as well as B and C also contained information about the similarity between A and C. We generalized this process by constructing a graph containing all active agents as nodes with corresponding similarities assigned to the edges, using the shortest path between arbitrary active agents as a foundation to predict a similarity score which could then be further modified by the remaining entities (ie, every entity except the active agent).

In the following section, we describe how we delt with the remaining entities, in which way we constructed the graph of all active agents, and how we used this information to predict similarity scores for new sentence pairs.

### Feature Construction

Even though we considered the active agents as the central part regarding sentence similarity, we still did not want to neglect other influences and, hence, we constructed a set of additional features per sentence pair, which reflect the similarity of everything except the active agents. More concretely, we constructed a set of similarity features $\Delta_k$ and compared the entity value of the first sentence $e_{k,1}$ with the entity value of the second sentence $e_{k,2}$. For nominally scaled entities, we calculated $\Delta_k$ as



and for ratio-scaled entity types, we used the squared difference



For entities like "strength" (eg, "4 mg"), we first separated the unit ("mg") from the number ("4"), used the nominal approach to compare the unit, and applied the squared difference equation on the number part. This differentiation gives us k=1 , … , $N$ features per sentence in total (N=5 in our case, since we used strength with amount and unit, and dose with amount and unit as well as frequency).

### Graph Construction

We used all medication sentences S = ($a_1$, $a_2$, $s$, $\Delta_1$, …, $\Delta_N$) from the training set with the active agents $a_1$ and $a_2$ from the sentence pair, the similarity score $s$, and the remaining entity features $\Delta_k$. We constructed our similarity graph $G$ ($V, E$) by using the possible active agents $A_i$ as nodes $V=\{A_1, A_2, …\}$ and connected all node pairs which occurred together in a sentence pair. More precisely, we constructed an edge set $E=\{(A_i, A_j, w_{ij})_{(i,j)\ P}\}$ with the set of all possible active agent pairs and the edge weight



which models the modified similarity score $w_{ij}$ between the active agents. $C$ denotes the set of all sentence pairs with the same active agents, and $\lambda$ represents weights for the entity differences learned during the training process (see later text). $\lambda_0$ can be interpreted as a bias. The tahn($x$) function limits the change of the similarity score, $s$, and the final result is clipped to stay in the valid range defined by $s_{min} = 0$ and $s_{max} = 5$.

The intuition here is that the weights, $w_{ij}$, should model the similarities between the active agents without factoring in the remaining entities. However, the true similarity value is not available as the similarity score, $s$, is also influenced by the remaining entities. The idea is that the weighted sum between the weights, $w_{ij}$, and the differences, $\Delta_k$, allows us to alter the

similarity score, $s$, in a way so that $w_{ij}$ models the true similarity between the active agents. The outer sum, responsible for averaging the items in the set $C$, is only necessary because it is possible that multiple sentence pairs with the same active agents exist.

## Inference

The goal of the inference phase was to calculate a sentence similarity score, $s$, between two active agents, $A_i$ and $A_j$, based on a similarity score, , obtained from the graph and the entity differences, $\Delta_k$. This consisted of two steps: first, we calculated the active agent similarity via the graph, and then we altered this similarity to account for the remaining entity features $\Delta_k$.

Step 1: In its simplest form, the similarity between two active agents is just the weight of the edge between the two corresponding active agent nodes. For example, $w_{ij} = $ . However, this is only possible when the weight already occurs in the training set and is not applicable in general, as $G(V, E)$ is not a complete graph, and it may be the case that an edge between the two active agent nodes does not exist. As we still wanted to make a prediction for these cases, we proposed to find the shortest path between these two active agent nodes and aggregate all edge weights along the way. This assumes a transitive relationship between the nodes or, for example, when there is a connection between $A_1$ and $A_2$ as well as $A_2$ and $A_3$, we can still say something about the nonexisting connection between $A_1$ and $A_3$. More concretely, we aggregate the information  along the shortest path



where $p_{ij}(1)$, $p_{ij}(2)$, …, $p_{ij}(M)$ denote the indices of the nodes on the shortest path between $A_i$ and $A_j$. This equation resembles the formula for calculating the resistance of parallel circuits



with the final resistance, $R_{eq}$, of the circuit and the resistances, $R_i$, of the individual flows. We chose this formula because the final resistance, $R_{eq}$, is always smaller than the individual resistances, $R_i$. For instance, $R_{eq} \leq \min(R_1, R_2, …)$

[35]. In our case, this implies that the score  obtained from the graph is always lower than any of the scores along the shortest path. This relies on our assumption that it is not possible to restore dissimilarities; for example, if there is already a score of 1 (low similarity) on an edge, we do not want to increase this value further by adding more connections, as we already know that at least two active agents are dissimilar.
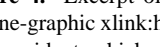
Step 2: The weight, , is the prediction for the similarity of an active agent pair. The final goal was to retrieve a prediction score, $s$, for a sentence pair which is also influenced by the remaining entity features, $\Delta_k$. We accounted for this by altering the predicted score again by



to retrieve a similarity, $s$, for the sentence pair.

Figure 4 shows an excerpt of the graph, which uses all the sentence pairs in the training set. The shortest path between the active agents "calcium" and "prednisone" is highlighted to visualize the prediction steps. Detailed calculations are available in the online version of the graph [36].

It may be possible that the sentences contain additional information that we do not cover in our approach, such as additional words, the relation between words, etc. For this reason, we combined the similarity score, $s_g$, from the graph with the BERT scores, $s_b$, in a Support Vector Regressor trained on all sentences in the training set to retrieve a final prediction score. We used Radial Basis Functions as kernel and optimized the regularization parameter $C$ as well as $\varepsilon$ ($\varepsilon$-tube without penalty) during the learning process of $\lambda_k$.

**Figure 4.** Excerpt of the medication graph, which models similarities between active agent pairs. On the edges, the modified similarity score, <inline-graphic xlink:href="medinform_v9i1e22795_fig10.png" mimetype="image" xlink:type="simple"/>, is shown. The full graph is available as an online widget, which provides further information and shows the graph calculations between arbitrary active agent nodes.

XSL•FO
**RenderX**
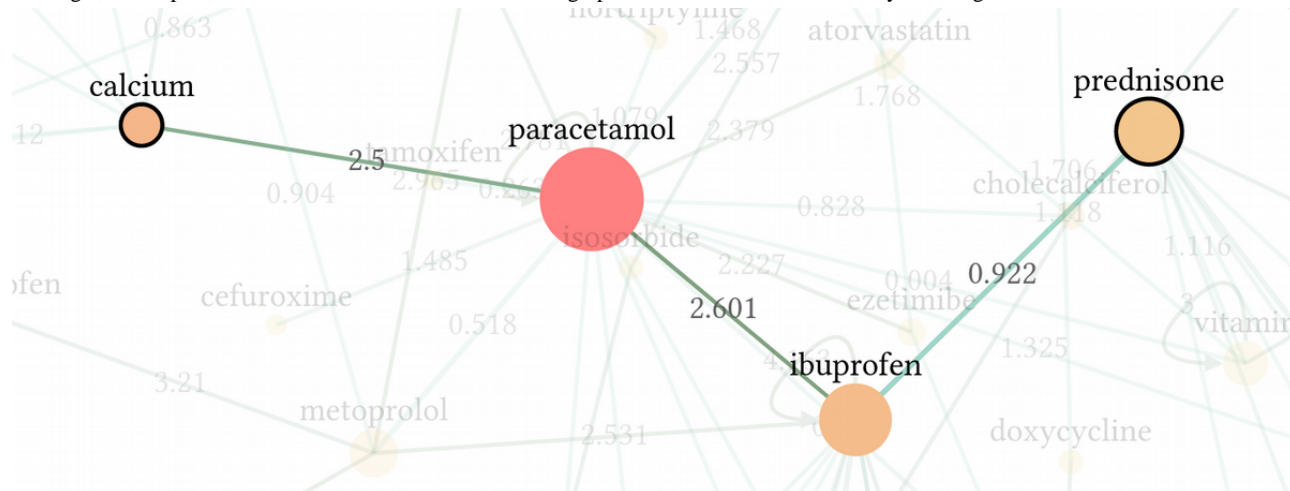
## Learning $\lambda_k$

The parameters $\lambda_k$ are responsible for the transformation between the active agent and sentence similarities as they reflect the importance of each entity feature, which also includes scaling differences. We did not manually craft these weights but learned them in a random walk process instead. The general idea was to randomly change the weights and see whether this improved the graph performance and, only if it did, we kept the change.

For a more stable evaluation, we split the training data into 10 folds, built a graph based on each training set, and evaluated the graph performance based on the corresponding test set. For evaluation, we calculated the mean squared error between the prediction scores and the ground truth. We did not use the Pearson correlation coefficient here because the correlation on a subset may not be as helpful for the correlation on the complete dataset as a measure which directly enforces a closeness with the ground truth.

Let $\lambda = (\lambda_0, \lambda_1, \ldots, \lambda_N)$ denote the vector with the current value of the weights (randomly initialized in the beginning) and let $\text{MES}(\lambda)$ denote the error when using these weights with the predictions from all folds. Then, we randomly selected an index, $k$, and altered the corresponding weight

$$\lambda'_k = \lambda_k + N(0,1)$$

via a sample from a standard normal distribution so that we obtained a new weight vector

$$\lambda' = (\lambda_0, \lambda_1, \ldots, \lambda'_k, \ldots, \lambda_N)$$

which we evaluated again on the graphs from all folds, keeping the change if

$$\text{MES}(\lambda') < \text{MES}(\lambda)$$

We repeated this process in two iterations, alternating with the process of hyperparameter tuning of the SVR model, until we observed no further improvements. For each random walk process, we applied 50 update steps. During development, we found that this setting was already sufficient and that the resulting weights tended to remain unchanged after these updates. For the SVR model, we applied a grid search to find values for the hyperparameters $C$ and ε-tube. We used the final weights to construct a new graph (based on all training data) used to predict the similarity of new sentences.

## Results

### Dataset Evaluation

In order to help with the interpretation of our results in the next section, we applied some basic statistical analysis on the training and test set, which revealed some imbalances. On average, the similarity score of the sentences in the training set (approximately 2.79) was higher than in the test set (approximately 1.76), whereas the standard deviation was slightly higher in the test set (approximately 1.52) than in the training set (approximately 1.39). This is also indicated by the left histogram chart of Figure 5, which, for example, reveals that sentence pairs with a score of approximately 1 are the most prominent ones in the test set, whereas in the training set, sentence pairs with a score of approximately 3 occur most often.

**Figure 5.** Histogram of the label distribution and word lengths of the training and test set.
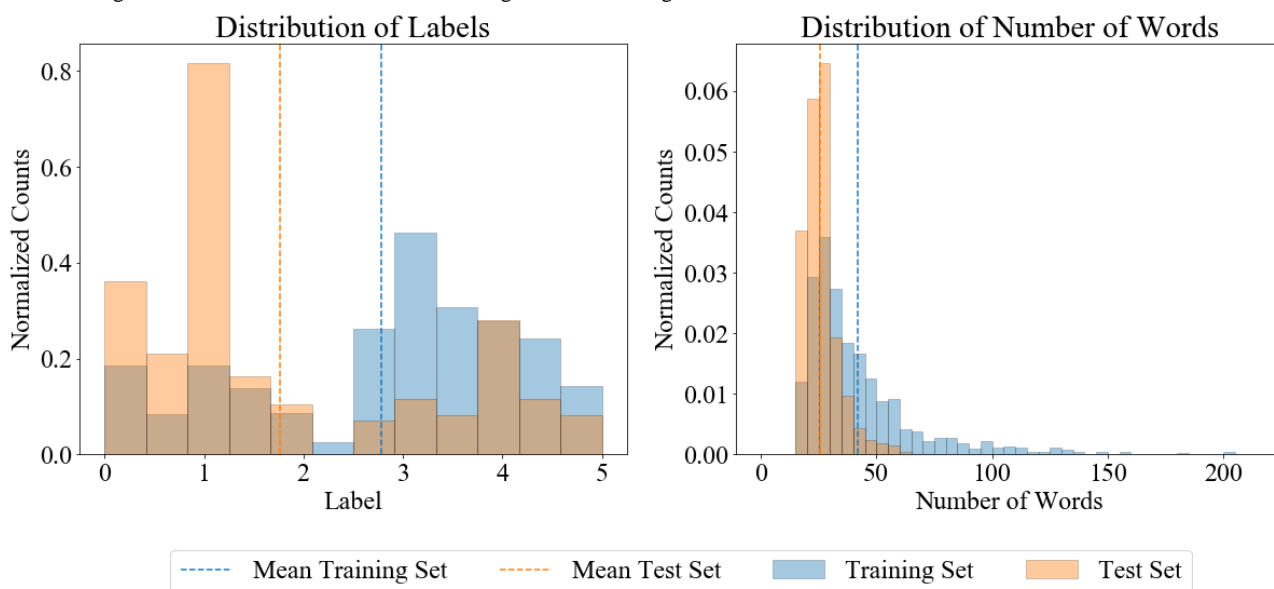


The right histogram chart of Figure 5 shows the distribution of the number of words of the training and test data. On average, the sentences in the test set tended to be shorter than in the training set, with an average sentence length of approximately 26 words per sentence pair (SD of approximately 7 words) in the former and approximately 42 words per sentence pair (SD of approximately 26 words) in the latter.

Finally, we calculated InferSent embeddings of the sentences in the training and test dataset and visualized them in a t-SNE (t-distributed stochastic neighbor embedding) plot (Figure 6). This shows that the sentence types occurring in the test dataset represent only a subset of those occurring in the training dataset, with many clusters of the training set being unoccupied by the

test set, such as the blue cluster in the bottom of Figure 6 without sentences of the test set in the neighborhood.

**Figure 6.** t-SNE projected InferSent embeddings of the sentences in the training and test dataset. Different groups of points correspond to different sentence types. For example, the group on the left upper side corresponds to the medication sentences. t-SNE: t-distributed stochastic neighbor embedding.
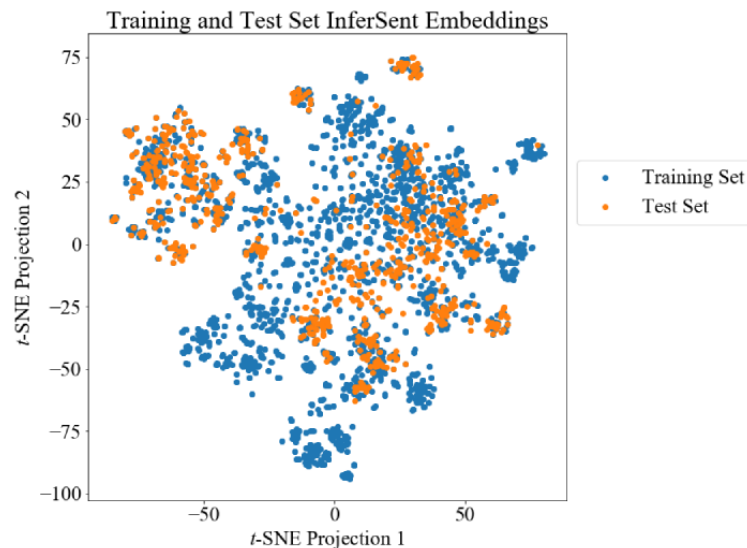


## Evaluation Results

We evaluated all runs on 3 different sets. Firstly, we used the training set with $k = 150$ folds to reduce the influence of noise in the data, to increase the comparability of our models, and to easily employ another ensembling technique for the test set. We wanted to measure the correlation of the training set and not of one of the folds to get comparable results. For this, we concatenated the predictions from each fold together and then calculated the Pearson correlation coefficient only once based on all scores. That is, we did not calculate a Pearson correlation coefficient for each fold, but rather collected the scores from all folds first. The consequence of this approach is that we cannot provide information about the variance, because only one Pearson correlation coefficient value is available.

Secondly, for the evaluation of the test set, we employed an additional ensembling technique by using the model for each fold to calculate a prediction for a sentence pair and then averaged all predictions.

Table 1 gives an overview of our results for the different datasets and approaches. Our best result with a Pearson correlation coefficient of 0.883 was achieved by combining enhanced BERT with $M$-Heads and the medication graph. For comparison, the winner of track 1 from IBM Research reached a Pearson correlation coefficient of 0.901 in their best submission [4].

**Table 1.** Summarization of the different approaches and their results. Training and test Pearson correlation coefficient scores are rounded to 3 decimal places.

| Approach | Training set | Test set |
| --- | --- | --- |
| **Approach 0: baseline** | | |
| ClinicalBERT | 0.850 | 0.859 |
| **Approach 1: voting regression** | | |
| Enhanced BERT | 0.851 | 0.859 |
| Voting Regression | 0.860 | 0.849[a] |
| **Approach 2: $M$-Heads** | | |
| Enhanced BERT with $M$-Heads | 0.853 | 0.876[a,b] |
| Enhanced BERT with $M$-Heads + Med. Graph | 0.853 | 0.883[c] |
| **Approach 3: medication graph** | | |
| Voting Regression + Med. Graph | 0.862 | 0.862[a] |

[a]Our submissions.

[b]We submitted a score of 0.869 for this setting because we were able to use only 10 $k$-folds due to a shortage of time.

[c]Our best result of the test set.

## Discussion

Our 3 approaches performed differently on the two datasets. In the following sections, we discuss the results in more detail and give our thoughts.

### Approach 1: Voting Regression

Evaluating the pure ClinicalBERT model, we see that the Pearson correlation coefficient is slightly higher for the test set as compared to the training set. The Enhanced BERT architecture led to an almost neglectable improvement on the training set and, in the test set, to no improvement at all. This indicates that, in this case, the additional features do not provide more information than what is already contained in the [CLS] token from the last hidden layer of BERT.

The Voting Regression approach showed an improvement of the Pearson correlation coefficient of the training set; however, for the test set, the performance decreased. These results might be traced back to overfitting of the training set. However, the decrease in the test set might also be explained by the imbalances between the training and test set.

### Approach 2: M-Heads

Adding $M = 4$ heads to BERT increased the Pearson correlation coefficient of both the training and test set as compared to ClinicalBERT. Especially for the test set, this indicates that the combination of the different heads improves BERT's performance.

### Approach 3: Medication Graph

Replacing the scores of the sentence subset which prescribes medications (cluster 3) with the medication graph scores led, in both cases (approaches 1 and 2), to an improvement for the test set. For the training set, however, we saw only marginal improvements, such as 0.860 to 0.862 from approach 1 to approach 3. This might be due to the Pearson correlation coefficient metric. In our experiments, we also evaluated our approaches with the Mean Squared Error between the predictions and the ground truth only on the subset of medication sentences. Without applying the medication graph (approach 1), we obtained a Mean Squared Error of 0.70, and with the medication graph (approach 3) a Mean Squared Error of 0.58. Combining the M-Heads approach with the medication graph yielded our best results for the test set, which indicates that BERT does have problems handling this domain-specific knowledge and therefore cannot cope well with these specific types of sentences.

Why did the medication graph perform better for the test set than for the training set? First, we observe that the test set contained more low-ranked sentences (see Figure 5); in particular, the medication sentences had lower scores. For the training set, the mean and standard deviation of the scores was 2.03 and 1.05, respectively, whereas the scores for the test set had only a mean and standard deviation of 1.10 and 0.50, respectively. We also noticed that the medication graph tended to dampen the prediction or, in other words, it led to lower scores. For example, the mean prediction score was 2.58 before and 1.78 after score replacement of the 94 medication sentences in the test set (see Table 2), which shows some example sentences of how the medication graph altered the scores). This could be due to two reasons: (1) the scores on the edges in the graph tended to be low (1.87 on average), and (2) the weight combination enforced low scores when there was at least one edge with a low score, which could explain why the medication graph achieved better predictions. This effect is facilitated by the fact that the training dataset contained only 147 out of 1642 (8.95%) of sentences that prescribed medication, whereas the test set contained 94 out of 412 (22.82%) medication sentences.

**Table 2.** Comparison of the Pearson correlation coefficient scores predicted by the Voting Regression (approach 1, A1) and the medication graph (approach 3, A3) via randomly selected example sentences. T denotes the ground truth score for the corresponding sentence pair. This table shows only the relevant entities from the original example sentences.

| Sentence a | Sentence b | T | A1 | A3 | Set |
|---|---|---|---|---|---|
| Ondansetron, 4 mg, 1 tablet, three times a day | Amoxicillin, 500 mg, 2 capsules, three times a day | 3.0 | 1.68 | 1.70 | Training |
| Prozac, 20 mg, 3 capsules, one time daily | Aleve, 220 mg, 1 tablet, two times a day | 0.5 | 2.02 | 1.68 | Training |
| Hydrochlorothiazide, 25 mg, one-half tablet, every morning | Ibuprofen, 600 mg, 1 tablet, four times a day | 1.5 | 1.59 | 1.70 | Training |
| Aleve, 220 mg, 1 tablet, two times a day | Acetaminophen, 500 mg, 2 tablets, three times a day | 1.5 | 2.74 | 1.68 | Test |
| Lisinopril, 10 mg, 2 tablets, one time daily | Naproxen, 500 mg, 1 tablet, two times a day | 1.0 | 2.29 | 1.69 | Test |

### Conclusions

To tackle the problem of semantic textual similarity of medical data, we developed 3 different approaches. We proposed to add additional features to BERT and to weigh different regression models based on the BERT result and other features. Moreover, we proposed the application of M-Heads and an attempt to automatically extrapolate medical knowledge from the training data. We observed that the success of the different methods strongly depended on the underlying dataset. In future work, it might be interesting to evaluate the methods on different and bigger datasets from other domains. The medication graph could be a powerful method with the possibility to be applied to other domains where it is necessary to extrapolate information from known entities and where it is not possible to calculate this information directly. It may also be used to model other concepts which exist in the medical domain, such as ontologies.

XSL·FO

RenderX

## Acknowledgments

## Conflicts of Interest

None declared.

Multimedia Appendix 1
Preprocessing.
[DOCX File , 11 KB - medinform_v9i2e22795_app1.docx ]

Multimedia Appendix 2
Implementation Details.
[DOCX File , 552 KB - medinform_v9i2e22795_app2.docx ]

Multimedia Appendix 3
Detailed description of Feature Set I and Feature Set II.
[DOCX File , 552 KB - medinform_v9i2e22795_app3.docx ]

Multimedia Appendix 4
M-Heads: training and prediction.
[DOCX File , 14 KB - medinform_v9i2e22795_app4.docx ]

## References

1. Wang Y, Afzal N, Fu S, Wang L, Shen F, Rastegar-Mojarad M, et al. MedSTS: a resource for clinical semantic textual similarity. Lang Resources & Evaluation 2018 Oct 24;54(1):57-72. [doi: 10.1007/s10579-018-9431-1]
2. Liu S, Wang Y, Liu H. Selected articles from the BioCreative/OHNLP challenge 2018. BMC Med Inform Decis Mak 2019 Dec 27;19(Suppl 10):262 [FREE Full text] [doi: 10.1186/s12911-019-0994-6] [Medline: 31882003]
3. Wang Y, Afzal N, Liu S, Rastegar-Mojarad M, Wang L, Shen F. Overview of BioCreative/OHNLP Challenge 2018 Task 2: Clinical Semantic Textual Similarity. In: Proceedings of the BioCreative/OHNLP Challenge. 2018. 2018 Presented at: BioCreative/OHNLP Challenge. 2018; August, 2018; Washington DC, USA. [doi: 10.1145/3233547.3233672]
4. Wang Y, Fu S, Shen F, Henry S, Uzuner O, Liu H. The 2019 n2c2/OHNLP Track on Clinical Semantic Textual Similarity: Overview. JMIR Med Inform 2020 Nov 27;8(11):e23375 [FREE Full text] [doi: 10.2196/23375] [Medline: 33245291]
5. Cer D, Diab M, Agirre E, Lopez-Gazpio I, Specia L. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017).: Association for Computational Linguistics; 2017 Presented at: 11th International Workshop on Semantic Evaluation (SemEval-2017); 2017; Vancouver, Canada p. 1-14. [doi: 10.18653/v1/s17-2001]
6. Fan H, Ma Z, Li H, Wang D, Liu J. Enhanced answer selection in CQA using multi-dimensional features combination. Tinshhua Sci. Technol 2019 Jun;24(3):346-359. [doi: 10.26599/tst.2018.9010050]
7. Majumder G, Pakray P, Gelbukh A, Pinto D. Semantic Textual Similarity Methods, Tools, and Applications: A Survey. CyS 2016 Dec 26;20(4). [doi: 10.13053/cys-20-4-2506]
8. Zhang S, Zheng X, Hu C. A survey of semantic similarity and its application to social network analysis. In: 2015 IEEE International Conference on Big Data (Big Data).: IEEE Computer Society; 2015 Presented at: IEEE International Conference on Big Data (Big Data); 2015; Santa Clara, CA, USA. [doi: 10.1109/bigdata.2015.7364028]
9. H.Gomaa W, A. Fahmy A. A Survey of Text Similarity Approaches. IJCA 2013 Apr 18;68(13):13-18. [doi: 10.5120/11638-7118]
10. Embi PJ, Weir C, Efthimiadis EN, Thielke SM, Hedeen AN, Hammond KW. Computerized provider documentation: findings and implications of a multisite study of clinicians and administrators. J Am Med Inform Assoc 2013;20(4):718-726 [FREE Full text] [doi: 10.1136/amiajnl-2012-000946] [Medline: 23355462]
11. Zhang R, Pakhomov SV, Lee JT, Melton GB. Using language models to identify relevant new information in inpatient clinical notes. AMIA Annu Symp Proc 2014;2014:1268-1276 [FREE Full text] [Medline: 25954438]
12. Chen Q, Du J, Kim S, Wilbur W, lu Z. Combining rich features and deep learning for finding similar sentences in electronic medical records. In: Proceedings of the BioCreative/OHNLP Challenge. 2018. 2018 Sep Presented at: BioCreative/OHNLP Challenge. 2018; August 2018; Washington DC, USA.

XSL•FO
RenderX

13. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: 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). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT 2019; 2019 Presented at: Conference of the North American Chapter of the Association for Computational Linguistics; 2019; Minneapolis, Minnesota p. 4171-4186.

14. Wang W, Bi B, Yan M, Wu C, Bao Z, Xia J. StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding. eprint arXiv:1908.04577. 2019 Aug. URL: https://arxiv.org/abs/1908.04577 [accessed 2021-01-07] [WebCite Cache ID 2019arXiv190804577W]

15. Sun Y, Wang S, Li Y, Feng S, Tian H, Wu H, et al. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding. In: AAAI Technical Track: Natural Language Processing. 2020 Apr 03 Presented at: AAAI Conference on Artificial Intelligence, 34(05); February, 2020; New York, USA p. 8968-8975. [doi: 10.1609/aaai.v34i05.6428]

16. Clark K, Luong MT, Le QV, Manning CD. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. 2020 Presented at: International Conference on Learning Representations; April 2020; Addis Ababa, Ethiopia p. 23.

17. Liu X, He P, Chen W, Gao J. Multi-Task Deep Neural Networks for Natural Language Understanding. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.: Association for Computational Linguistics; 2019 Presented at: 57th Annual Meeting of the Association for Computational Linguistics; July, 2019; Florence, Italy p. A. [doi: 10.18653/v1/P19-1441]

18. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research 2010;21:1-67.

19. Liu H, Perl Y, Geller J. Transfer Learning from BERT to Support Insertion of New Concepts into SNOMED CT. AMIA Annu Symp Proc 2019;2019:1129-1138 [FREE Full text] [Medline: 32308910]

20. Sun C, Yang Z. Transfer Learning in Biomedical Named Entity Recognition: An Evaluation of BERT in the PharmaCoNER task. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks.: Association for Computational Linguistics; 2019 Presented at: The 5th Workshop on BioNLP Open Shared Tasks; November 2019; Hong Kong, China p. 100-104. [doi: 10.18653/v1/d19-5715]

21. Lin C, Bethard S, Dligach D, Sadeque F, Savova G, Miller T. Does BERT need domain adaptation for clinical negation detection? J Am Med Inform Assoc 2020 Apr 01;27(4):584-591. [doi: 10.1093/jamia/ocaa001] [Medline: 32044989]

22. MIC-DKFZ/n2c2 Challenge 2019 Source-Code. MIC-DKFZ GitHub. URL: https://github.com/MIC-DKFZ/n2c2-challenge-2019 [accessed 2021-07-01]

23. Alsentzer E, Murphy J, Boag W, Weng W, Jindi D, Naumann T. Publicly Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop.: Association for Computational Linguistics; 2019 Presented at: 2nd Clinical Natural Language Processing Workshop; June, 2019; Minneapolis, Minnesota, USA p. 72-78. [doi: 10.18653/v1/w19-1909]

24. Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017 Presented at: Conference on Empirical Methods in Natural Language Processing; September, 2017; Copenhagen, Denmark p. 670-680. [doi: 10.18653/v1/d17-1070]

25. Wang B, Kuo CJ. SBERT-WK: A Sentence Embedding Method by Dissecting BERT-Based Word Models. IEEE/ACM Trans. Audio Speech Lang. Process 2020;28:2146-2157. [doi: 10.1109/taslp.2020.3008390]

26. Reimers N, Gurevych I. Sentence-BERTntence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).: Association for Computational Linguistics; 2019 Presented at: 2019 Conference on Empirical Methods in Natural Language Processing; November, 2019; Hong Kong, China p. A. [doi: 10.18653/v1/d19-1410]

27. Varoquaux G, Buitinck L, Louppe G, Grisel O, Pedregosa F, Mueller A. Scikit-learn. GetMobile: Mobile Comp. and Comm 2015 Jun;19(1):29-33. [doi: 10.1145/2786984.2786995]

28. Opitz D, Maclin R. Popular Ensemble Methods: An Empirical Study. jair 1999 Aug 01;11:169-198. [doi: 10.1613/jair.614]

29. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis 2015 Apr 11;115(3):211-252. [doi: 10.1007/s11263-015-0816-y]

30. Lee S, Purushwalkam S, Cogswell M, Crandall D, Batra D. Why M Heads are Better than One: Training a Diverse Ensemble of Deep Networks. eprint arXiv:1511.06314. 2015 Nov. URL: https://arxiv.org/abs/1511.06314 [accessed 2020-01-07]

31. Ilg E, Çiçek , Galesso S, Klein A, Makansi O, Hutter F. Uncertainty Estimates and Multi-hypotheses Networks for Optical Flow. In: European Conference on Computer Vision (ECCV). 2018 Presented at: European Conference on Computer Vision (ECCV); 2018; Munich, Germany. [doi: 10.1007/978-3-030-01234-2_40]

32. Rupprecht C, Laina I, DiPietro R, Baust M, Tombari F, Navab N. Learning in an Uncertain World: Representing Ambiguity Through Multiple Hypotheses. In: 2017 IEEE International Conference on Computer Vision (ICCV).: IEEE; 2017 Presented at: IEEE International Conference on Computer Vision (ICCV); 2017; Venice, Italy p. 3611-3620. [doi: 10.1109/iccv.2017.388]

33. Jiang M, Wu Y, Shah A, Priyanka P, Denny JC, Xu H. Extracting and standardizing medication information in clinical text - the MedEx-UIMA system. AMIA Jt Summits Transl Sci Proc 2014;2014:37-42 [FREE Full text] [Medline: 25954575]

34. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc 2010;17(1):19-24 [FREE Full text] [doi: 10.1197/jamia.M3378] [Medline: 20064797]

35. Meschede D. Gerthsen Physik. In: Gerthsen Physik. Heidelberg, Germany: Springer Spektrum; 2015.

36. Medication Graph Visualization. URL: https://med-graph.jansellner.net/ [accessed 2021-07-01]

## Abbreviations

**BERT:** Bidirectional Encoder Representations from Transformers
**NLP:** Natural language Processing
**n2c2:** National NLP Clinical Challenges
**OHNLP:** Open Health Natural Language Processing

XSL•FO
**RenderX**

<u>Original Paper</u>

# Measuring Adoption of Patient Priorities–Aligned Care Using Natural Language Processing of Electronic Health Records: Development and Validation of the Model

Javad Razjouyan[1,2,3], PhD; Jennifer Freytag[1], JD, PhD; Lilian Dindo[1,2], PhD; Lea Kiefer[1], MPH; Edward Odom[1], BS; Jaime Halaszynski[4,5,6], LCSW; Jennifer W Silva[5,6], LCSW; Aanand D Naik[1,2,3,7], MD

[1]VA Health Services Research and Development Service, Center for Innovations in Quality, Effectiveness and Safety, Michael E DeBakey VA Medical Center, Houston, TX, United States

[2]Department of Medicine, Baylor College of Medicine, Houston, TX, United States

[3]Big Data Scientist Training Enhancement Program (BD-STEP), VA Office of Research and Development, Washington, DC, United States

[4]Social Work Service, Butler VA Health Care System, Butler, PA, United States

[5]VA National Social Work Program Office, Care Management and Social Work, Patient Care Services, Department of Veterans Affairs, Washington, DC, United States

[6]VA Tennessee Valley Healthcare System, Nashville, TN, United States

[7]VA Quality Scholars Coordinating Center, IQuESt, Michael E DeBakey VA Medical Center, Houston, TX, United States

**Corresponding Author:**
Javad Razjouyan, PhD
VA Health Services Research and Development Service
Center for Innovations in Quality, Effectiveness and Safety
Michael E DeBakey VA Medical Center
2002 Holcombe Blvd
Houston, TX, 77030
United States
Phone: 1 5203424525
Email: javad.razjouyan@bcm.edu

## *Abstract*

**Background:**  Patient Priorities Care (PPC) is a model of care that aligns health care recommendations with priorities of older adults who have multiple chronic conditions. Following identification of patient priorities, this information is documented in the patient's electronic health record (EHR).

**Objective:**  Our goal is to develop and validate a natural language processing (NLP) model that reliably documents when clinicians identify patient priorities (ie, values, outcome goals, and care preferences) within the EHR as a measure of PPC adoption.

**Methods:**  This is a retrospective analysis of unstructured National Veteran Health Administration EHR free-text notes using an NLP model. The data were sourced from 778 patient notes of 658 patients from encounters with 144 social workers in the primary care setting. Each patient's free-text clinical note was reviewed by 2 independent reviewers for the presence of PPC language such as priorities, values, and goals. We developed an NLP model that utilized statistical machine learning approaches. The performance of the NLP model in training and validation with 10-fold cross-validation is reported via accuracy, recall, and precision in comparison to the chart review.

**Results:**  Of 778 notes, 589 (75.7%) were identified as containing PPC language (kappa=0.82, *P*<.001). The NLP model in the training stage had an accuracy of 0.98 (95% CI 0.98-0.99), a recall of 0.98 (95% CI 0.98-0.99), and precision of 0.98 (95% CI 0.97-1.00). The NLP model in the validation stage had an accuracy of 0.92 (95% CI 0.90-0.94), recall of 0.84 (95% CI 0.79-0.89), and precision of 0.84 (95% CI 0.77-0.91). In contrast, an approach using simple search terms for PPC only had a precision of 0.757.

**Conclusions:**  An automated NLP model can reliably measure with high precision, recall, and accuracy when clinicians document patient priorities as a key step in the adoption of PPC.

XSL•FO
**RenderX**

## Introduction

Older adults with multiple chronic conditions (MCC) frequently receive some of the most intensive and expensive health care, much of which is of uncertain benefit [1-3]. Care for these patients is often inconsistent and fragmented because the multiple specialists they see provide care based on single disease guidelines that do not take into account the complexities of MCC. Moreover, it often does not address what matters most to patients [1,2]. With input from patients, caregivers, clinicians, health system leaders, payers, and health care design experts, we developed Patient Priorities Care (PPC), which involves identifying, documenting, and providing care consistent with patients' health priorities [4,5]. PPC is an intervention that aims to help clinicians provide health care to older patients with MCC that aligns with their priorities, which include their individual values (what matters most), desired outcome goals, and health care preferences [1-6].

In the PPC process, there are 3 key steps. First, a trained facilitator works with each patient to identify the patient's health care priorities: their values, outcome goals, and health care preferences [1-3,5]. Values are the people, things, activities, and capabilities *that matter most* to a person, such as connecting with others, independence, enjoying life, and balancing quality and quantity of life [7,8]. Patient values provide the basis for setting specific, measurable, attainable, realistic, and timely outcome goals (SMART goals). Health care preferences are the activities a patient is willing (or not willing) to do to reach their outcome goals. The second key step in PPC involves documenting patient priorities in the patient's electronic health record (EHR) [4,5]. This makes information about patient priorities available to different clinicians working with adults with MCC. In the final step of PPC, the clinician uses documented patient priorities to make treatment decisions that align with patient priorities [3].

Social workers frequently serve as facilitators in the PPC approach. Social workers use a holistic approach and encourage self-determination while promoting dignity and worth of the individual. They are also uniquely qualified to provide a wide array of quality social work interventions, including care coordination, case management, individual and group therapy, and supportive counseling for older adults with MCC. Social workers address social determinants of health and provide proactive interventions for social conditions that contribute to poor health outcomes. Social workers facilitate communication between the patient and health care provider and act as an advocate on behalf of the patient.

The PPC process thus relies on identifying and documenting patient priorities in the EHR so facilitators can share priorities with other clinicians. This documentation is done in the free text of individual visit notes. To measure adoption of PPC, the free text of clinician notes must be analyzed. Analyzing the free-text notes as yes (PPC present) or no (PPC not present) is

entitled as text classification [9]. Although the gold standard for text classification is human chart review, it is tedious, time consuming, expensive, and potentially risks patient privacy [10]. To overcome these limitations, natural language processing (NLP) has been introduced as an alternative or adjunct to manual chart abstraction [10]. NLP refers to the computational methods of concepts, entities, events, and relations extraction from free text [11]. NLP algorithms are used in a wide spectrum of health care–related purposes such as identifying disease risk factors, evaluating efficiency of care and costs, and extracting information from clinical notes [12].

The first aim of this study was to develop an NLP algorithm capable of confirming social worker use of PPC in the free-text notes of patients' EHRs. We hypothesized that the NLP model could confirm documentation of patient priorities in the free text of facilitator notes. We compared the performance of the NLP model with a chart‑reviewed reference standard by reporting performance measures (ie, accuracy, recall, specificity, precision, negative predictive value). To further evaluate the performance of the NLP algorithm, we compared its accuracy to a simple search terms approach. We also used the NLP algorithm to monitor the uptake of PPC among trained social workers following dedicated training sessions and track the persistence of PPC adoption over time.

## Methods

### Overview

To create the NLP model to assess the adoption of PPC, we used the notes of a cohort of social workers who were trained to be PPC clinicians over the course of 9 months. This cohort of social workers was hired through the VA Social Work in Patient Aligned Care Teams (PACT) Staffing Program. This program was developed in May 2016 and is the result of a partnership between the Veterans Health Administration (VHA) National Social Work Program Office and the Office of Rural Health. The goal of the program is to embed social workers in rural and highly rural areas to increase Veteran access to high-quality social work services. The innovative program supports the Office of Rural Health's mission of breaking down barriers separating rural Veterans from high-quality care. Social workers are often one of the first clinicians a Veteran has contact with in the VHA. Social workers in this program provide interventions at numerous rural health VHA medical centers and outpatient clinics across the nation. Two cohorts of social workers were trained: 56 social workers from 5 VA health care systems in 2018 and 88 social workers from 12 VA health care systems in 2019. Social workers were trained to assess, evaluate, and select patients who were good candidates for PPC. Patients had to be older than 60 years of age and have at least two primary care encounters in a prior year. Social workers were trained to facilitate PPC conversations using patient and facilitator workbooks. Social workers were instructed to document all PPC-related interactions in the patient free-text

note including all components of PPC. A PPC rubric was provided to social workers to document components of PPC: patient values, outcome goals, and health care preferences (see Multimedia Appendix 1). After reviewing the PPC template and related notes, we recognized that all notes must have 2 phrases: "goal" and "value." Notes without these 2 expressions are not considered as PPC.

## Data Collection (Search Term Approach)

Notes of eligible patients were retrieved from the VHA Corporate Data Warehouse [13]. First, notes of all patients who visited with a PPC-trained social worker were retrieved from the Corporate Data Warehouse and included in a database. Notes were obtained from the first facilitated PPC visit through the end of the 2019 calendar year. The Baylor College of Medicine Institutional Review Board and the Veterans Affairs research service (IRB# H-41886) approved this study. Access to patients' data was granted through the Veterans Informatics and Computing Infrastructure.

## The Reference Standard

To develop the NLP algorithm, we extracted 778 notes from 658 patients who were seen by the social workers. We developed a reference standard for identifying PPC using a formal chart review. We provided complete copies of the notes from the 2018 and 2019 cohorts to 2 independent reviewers (EO and JF). The reviewers ensured the existence of language for values, outcome goals, and care preferences for each free-text note. Charts were labeled as 0 or 1; 0 refers to absence of documented patient priorities for the patient's note, and 1 refers to documentation of priorities in free-text notes of patients. Any disagreement between reviewers was resolved by a third reviewer, a geriatrician involved in the development of PPC (AN). The agreement among reviewers is reported using the kappa statistic, and the significance level was set at $P<.05$; statistical analysis was conducted using SPSS software.

## NLP Algorithm

The NLP algorithm was developed using 3 steps: preprocessing, processing, and postprocessing (Figure 1). In the preprocessing step, we cleaned the text, performed features extraction, and reduced the number of features to reduce the size of the feature space (dimensionality reduction). In the processing step, we developed a classification model to identify the yes-PPC from no-PPC notes. In postprocessing, we reported the performance models developed in the processing step and assessed the generalizability of the model.

**Figure 1.** Process of developing the natural language processing (NLP) algorithm: (A) retrieving 106,505 notes from trained social workers after the first training workshop; (B) searching notes for existence of 2 expressions "goal" and "value," which resulted into 778 notes; (C) review and resolution of disagreements; (D) labeling of notes as yes-Patient Priorities Care (PPC; 1) or no-PPC (0); (E) development of NLP algorithm.
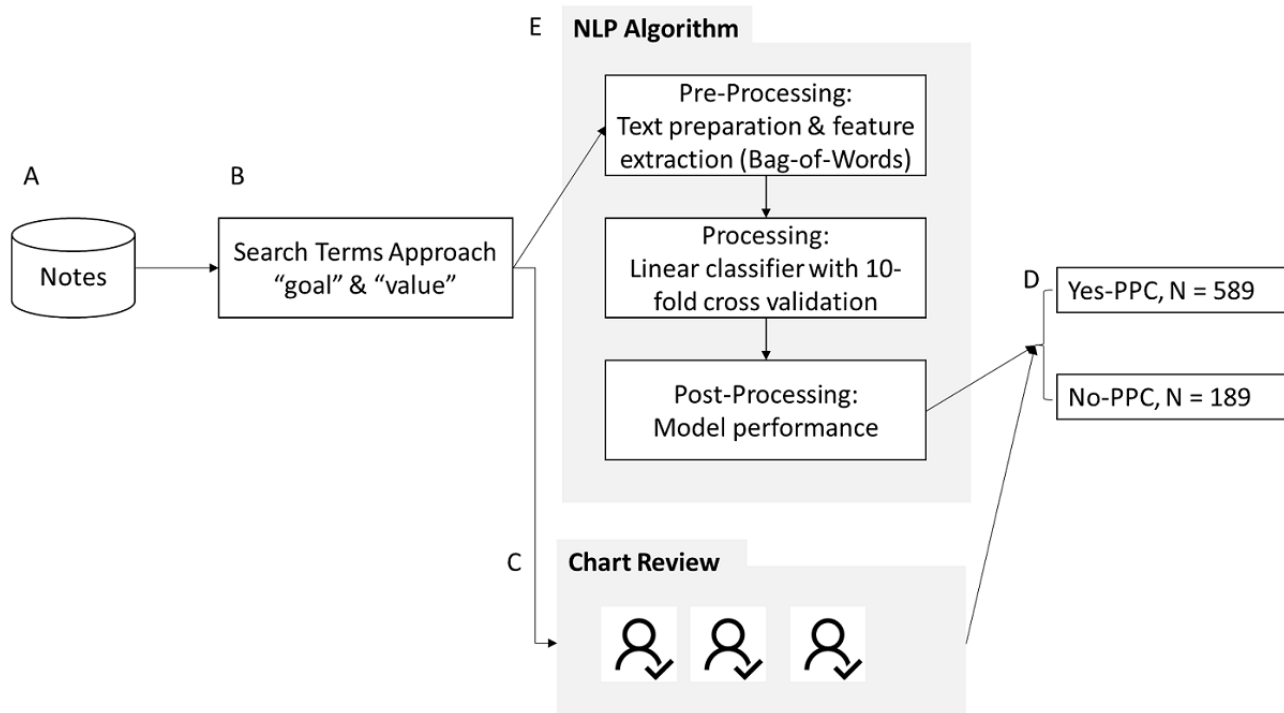


### Preprocessing, Cleaning the Text

We converted the stream of text into meaningful elements such as words, phrases, and symbols, called tokenization. We added the part of speech (eg, noun, verb, and pronoun) to each word. We removed stop words (eg, articles, conjunctions, and prepositions) and punctuation as they do not contain informational content. Words were normalized by reducing words to their word stem (stemming technique). For example,

the words caregiver, caregivers, and caregiving would be reduced to "caregiv."

### Cleaning, Features Extraction, and Feature Reduction

To extract features from the text, we used "bag-of-words" (BoW). BoW (also known as a term-frequency counter) records the number of times that each individual word (unigram) appears in a document. We excluded unigrams with a frequency ≤2. These words must appear in the entire sentence related to goals

and values <2 times (Multimedia Appendix 2 visualizes the BoW).

### Processing Step, Developing a Classification Model

We fit a linear regression model to the BoW as an independent variable and the document label of 0 versus 1 as a dependent variable. To develop the model, we used K-fold cross-validation, with K=10. The goal of cross-validation is to test an approach's ability to predict new notes not used in training to avoid problems such as overfitting or selection bias [14]. In the 10-fold cross-validation, we randomly split our dataset into 10 equal sections [15]. We trained a model with 9 splits, and we validated the performance of the model on the remaining split. The process was repeated 10 times, and the average performance is reported.

### Postprocessing Step

We tested each model using a validation cohort, using the model with the highest performance as compared to the chart review (validation of the model). To evaluate the performance of the NLP model in training, validation, and test phases, we used sensitivity or recall, specificity, accuracy, positive predictive value or precision, negative predictive value, and area under the curve [16,17]. To further validate the NLP approach, we compared its performance to an approach that used a simple search term assessment for the 2 expressions: "goal" and "value."

### PPC Adoption

As a second objective, we evaluated the adoption of PPC among trained social workers. We counted the number of yes-PPC notes in each month, and we created a time series of notes identified by the PPC algorithm. The x axis shows the month,

and the y axis shows the number of documented notes with PPC language in the EHR. This evaluation helped to understand the implementation of the PPC approach among social workers who received formal training and the persistence of PPC adoption over time.

## Results

### Data

We collected 778 notes that had "goal" and "value" within their text. These notes were collected from 658 patients from 17 VA facilities by 82 social workers. Mean patient age was 70.4 years (SD 15.2 years), 91.6% (603/658) were male, 23.4% (154/658) had all-cause-of-mortality recorded, 91.8% (604/658) were of "Not Hispanic or Latino" ethnicity, 71.9% (473/658) were White, 6.7% (44/658) were "Black or African American," and 47.4% (312/658) were married.

The 2 reviewers disagreed on 54 notes and agreed on 724 (724/778, 93.1%) notes. The level of agreement after removing the chance effect was reduced to 82% (kappa=0.820, standard error 0.023, $P<.001$).

### Model Performance

The performance of the NLP model is shown in Table 1. The NLP model performance was calculated over notes (N=778) that had "goal" and "value" in their text. In the 10-fold cross-validation, the model validation reached an average of 0.92 (95% CI 0.90-0.94) for accuracy, 0.84 (95% CI 0.79-0.89) for sensitivity, 0.95 (95% CI 0.92-0.97) for specificity, 0.84 (95% CI 0.77-0.91) for the positive predictive value, and 0.95 (95% CI 0.93-0.96) for the negative predictive value.

**Table 1.** Comparison of chart-reviewed, free-text notes and natural language processing (NLP) model for the training and validation.

| Parameters | Training, mean (95% CI)[a] | Validation, mean (95% CI)[a] |
|---|---|---|
| Accuracy | 0.98 (0.98-0.99) | 0.92 (0.90-0.94) |
| Sensitivity or recall | 0.98 (0.98-0.99) | 0.84 (0.79-0.89) |
| Specificity | 0.99 (0.98-0.99) | 0.95 (0.92-0.97) |
| Positive predictive values or precision | 0.98 (0.97-1.00) | 0.84 (0.77-0.91) |
| Negative predictive value | 0.99 (0.98-0.99) | 0.95 (0.93-0.96) |

[a]Reported for 10-fold cross validation.

In contrast, an approach using simple search terms ("goal" and "value") applied to large text performed less well. The total number of notes that the social workers produced was 106,505. Table 2 describes the precision of this model. The precision of

the NLP model was >0.99, while the precision of the simple search terms was 0.7571. In the other words, the simple search terms method falsely recognized 1 out of 4 notes as PPC.

**Table 2.** Comparison of the natural language processing (NLP) model and simple search terms on all 106,505 notes.

| Parameters | Search terms | NLP model |
|---|---|---|
| Accuracy | 0.9982 | 0.9999 |
| Sensitivity or recall | 1.0000 | 0.9999 |
| Specificity | 0.9982 | 0.9983 |
| Positive predictive values or precision | 0.7571 | 1.0000 |
| Negative predictive value | 1.0000 | 0.9849 |

Figure 2 describes the adoption of PPC within the notes of trained social workers over time. The 2 peaks represent the times immediately following training of social workers to use PPC. Figure 2 also demonstrates that several months following the cessation of training, the persistence of PPC adoption declined. The NLP algorithm was useful in tracking PPC adoption over time.

**Figure 2.** Monthly report of the number of Patient Priorities Care (PPC) encounters documented in the patients' free-text notes in the Veterans Administration (VA) electronic health records; the 2 peaks are associated with training of social workers to use PPC.
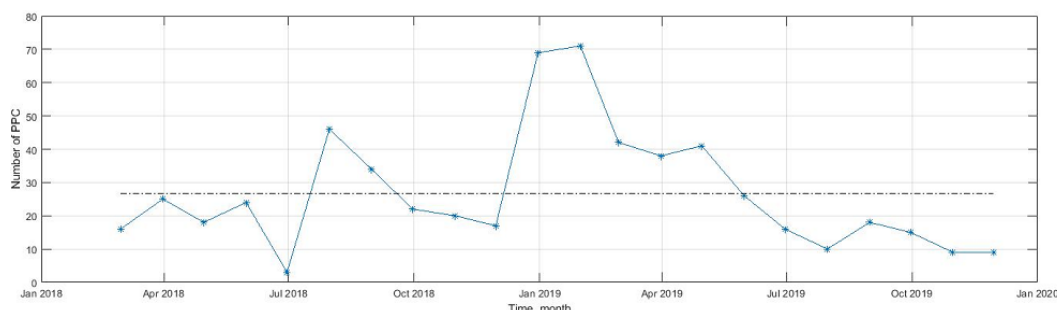


## Discussion

### Principal Findings

Our results demonstrate that an NLP algorithm was highly accurate in identifying patients' free-text notes with PPC language. Such a design has shown promise in previous work identifying risk factors for heart disease [18]. The NLP model had a high performance of 0.84 in validation, and its precision was higher than an approach using only simple search terms. The NLP algorithm was also able to capture peaks in PPC adoption following social worker training and the eventual decline in PPC adoption over time. Such a tool may be used to monitor the overall implementation of interventions like PPC among trained social workers and to adjust the training process.

While this NLP model performed well, model performance (precision and recall) could improve with the use of more complex strategies such as utilizing bigram or n-gram BoW, other sophisticated feature extractions (eg, graph mining [19]), and nonlinear models (eg, deep learning [20]). To reduce the computational cost of developing and implementing a model to test the use of PPC by trained facilitators, we propose the NLP model as a method that reduces the time and cost with other methods of assessment. Further development of this model will offer assessment with more precision. Despite these limitations, the NLP model performed better than the simple search term method, which had a 1-in-4 notes rate of misclassification.

This study opens 3 new avenues for further investigation. By identifying PPC notes, we can longitudinally assess care for patients throughout VA facilities and understand changes in health care delivered after PPC encounters [6,21]. We can identify patterns and consistency in how social workers and other clinicians document patient priorities over the course of multiple interactions with patients [3]. By further expanding the NLP algorithm, we can also examine common priorities among our patient population and increase clinician, patient, and caregiver awareness of priorities recorded in the EHR. For successful implementation of PPC, we must assess other factors that play roles in the successful implementation of PPC, such as resource availability, adoption of PPC on a provider and facility level, and readiness of clinicians for implementation [4]. We can also use sociodemographic data to pinpoint the ways individual characteristics of patients might influence the implementation of PPC by clinicians. We can also provide individualized feedback to each clinician to enhance their patient engagement process. Using this algorithm, we can measure long-term outcomes of interest, such as mortality and independent living.

### Implications

This study achieved its overall objective of developing a simple NLP model to identify free-text notes that show evidence of documented patient priorities as part of the PPC process. Our NLP model complements PPC by measuring the extent to which social workers trained in PPC identify and document patient priorities. The NLP model can enable trainers to provide feedback on the implementation of PPC in the EHR. Additionally, this method can analyze the free text of any note for documentation of patient priorities generally in an effort to link those documented priorities with many patient outcomes without requiring labor-intensive methods for individual chart review.

XSL•FO

**RenderX**

## Authors' Contributions

AN and JR had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. AN and JR conceived and designed the study. JR, AN, and LD acquired, analyzed, and interpreted the data. JR and JF drafted the manuscript. JV performed the statistical analysis. AN obtained funding. JR, AN, LD, JH, and JS critically revised the manuscript for important intellectual content. JF, LK, EO, JH, and JS provided administrative, technical, or material support. AN and LD supervised the study.

## Conflicts of Interest

None declared.

Multimedia Appendix 1
Note template provided to the social workers during online training.
[PNG File , 607 KB - medinform_v9i2e18756_app1.png ]

Multimedia Appendix 2
Visualization of each word in the Patient Priorities Care (PPC) notes, with the importance based on frequency.
[PNG File , 389 KB - medinform_v9i2e18756_app2.png ]

## References

1. Ferris R, Blaum C, Kiwak E, Austin J, Esterson J, Harkless G, et al. Perspectives of Patients, Clinicians, and Health System Leaders on Changes Needed to Improve the Health Care and Outcomes of Older Adults With Multiple Chronic Conditions. J Aging Health 2018 Jun 01;30(5):778-799. [doi: 10.1177/0898264317691166] [Medline: 28553806]
2. Tinetti ME, Esterson J, Ferris R, Posner P, Blaum CS. Patient Priority-Directed Decision Making and Care for Older Adults with Multiple Chronic Conditions. Clin Geriatr Med 2016 May;32(2):261-275. [doi: 10.1016/j.cger.2016.01.012] [Medline: 27113145]
3. Tinetti ME, Naik AD, Dindo L, Costello DM, Esterson J, Geda M, et al. Association of Patient Priorities-Aligned Decision-Making With Patient Outcomes and Ambulatory Health Care Burden Among Older Adults With Multiple Chronic Conditions: A Nonrandomized Clinical Trial. JAMA Intern Med 2019 Oct 07;179(12):1688 [FREE Full text] [doi: 10.1001/jamainternmed.2019.4235] [Medline: 31589281]
4. Blaum CS, Rosen J, Naik AD, Smith CD, Dindo L, Vo L, et al. Feasibility of Implementing Patient Priorities Care for Older Adults with Multiple Chronic Conditions. J Am Geriatr Soc 2018 Oct 03;66(10):2009-2016 [FREE Full text] [doi: 10.1111/jgs.15465] [Medline: 30281777]
5. Naik AD, Dindo LN, Van Liew JR, Hundt NE, Vo L, Hernandez-Bigos K, et al. Development of a Clinically Feasible Process for Identifying Individual Health Priorities. J Am Geriatr Soc 2018 Oct 03;66(10):1872-1879. [doi: 10.1111/jgs.15437] [Medline: 30281794]
6. Freytag J, Dindo L, Catic A, Johnson AL, Bush Amspoker A, Gravier A, et al. Feasibility of Clinicians Aligning Health Care with Patient Priorities in Geriatrics Ambulatory Care. J Am Geriatr Soc 2020 Sep 20;68(9):2112-2116. [doi: 10.1111/jgs.16662] [Medline: 32687218]
7. Naik AD, Martin LA, Moye J, Karel MJ. Health Values and Treatment Goals of Older, Multimorbid Adults Facing Life-Threatening Illness. J Am Geriatr Soc 2016 Mar 21;64(3):625-631 [FREE Full text] [doi: 10.1111/jgs.14027] [Medline: 27000335]
8. Karel MJ, Mulligan EA, Walder A, Martin LA, Moye J, Naik AD. Valued life abilities among veteran cancer survivors. Health Expect 2016 Jun 29;19(3):679-690 [FREE Full text] [doi: 10.1111/hex.12343] [Medline: 25645124]
9. Kowsari K, Jafari Meimandi K, Heidarysafa M, Mendu S, Barnes LE, Brown DE. Text Classification Algorithms: A Survey. Information 2019 Apr 23;10(4):150. [doi: 10.3390/info10040150]
10. Carrell D, Halgrim S, Tran DT, Buist DSM, Chubak J, Chapman WW, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. Am J Epidemiol 2014 Mar 15;179(6):749-758 [FREE Full text] [doi: 10.1093/aje/kwt441] [Medline: 24488511]
11. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: A literature review. J Biomed Inform 2018 Jan;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
12. Koleck T, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc 2019 Apr 01;26(4):364-379 [FREE Full text] [doi: 10.1093/jamia/ocy173] [Medline: 30726935]
13. Begoli E, Kistler D, Bates J. Towards a heterogeneous, polystore-like data architecture for the US Department of Veteran Affairs (VA) enterprise analytics. 2017 Presented at: 2016 IEEE International Conference on Big Data; December 5-8, 2016; Washington, DC. [doi: 10.1109/bigdata.2016.7840896]
14. Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research 2010;11:2079-2107.

15. Chicco D. Ten quick tips for machine learning in computational biology. BioData Min 2017 Dec 8;10(1):35 [FREE Full text] [doi: 10.1186/s13040-017-0155-3] [Medline: 29234465]

16. Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. J Biomed Inform 2006 Dec;39(6):589-599 [FREE Full text] [doi: 10.1016/j.jbi.2005.11.004] [Medline: 16359928]

17. Ly T, Pamer C, Dang O, Brajovic S, Haider S, Botsis T, et al. Evaluation of Natural Language Processing (NLP) systems to annotate drug product labeling with MedDRA terminology. J Biomed Inform 2018 Jul;83:73-86 [FREE Full text] [doi: 10.1016/j.jbi.2018.05.019] [Medline: 29860093]

18. Yang H, Garibaldi JM. A hybrid model for automatic identification of risk factors for heart disease. J Biomed Inform 2015 Dec;58 Suppl:S171-S182 [FREE Full text] [doi: 10.1016/j.jbi.2015.09.006] [Medline: 26375492]

19. Madan S, Szostak J, Komandur Elayavilli R, Tsai RTH, Ali M, Qian L, et al. The extraction of complex relationships and their conversion to biological expression language (BEL) overview of the BioCreative VI (2017) BEL track. Database (Oxford) 2019 Jan 01;2019 [FREE Full text] [doi: 10.1093/database/baz084] [Medline: 31603193]

20. Weng W, Wagholikar KB, McCray AT, Szolovits P, Chueh HC. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Med Inform Decis Mak 2017 Dec 01;17(1):155 [FREE Full text] [doi: 10.1186/s12911-017-0556-8] [Medline: 29191207]

21. Feder S, Kiwak E, Costello D, Dindo L, Hernandez-Bigos K, Vo L, et al. Perspectives of Patients in Identifying Their Values-Based Health Priorities. J Am Geriatr Soc 2019 Jul;67(7):1379-1385 [FREE Full text] [doi: 10.1111/jgs.15850] [Medline: 30844080]

## Abbreviations

**BoW:** bag-of-words
**EHR:** electronic health record
**MCC:** multiple chronic conditions
**NLP:** natural language processing
**PACT:** Patient Aligned Care Teams
**PPC:** Patient Priorities Care
**VHA:** Veterans Health Administration

XSL•FO
**RenderX**

Original Paper

# Driving Digital Transformation During a Pandemic: Case Study of Virtual Collaboration in a German Hospital

Nicholas R J Frick[1*], BSc, MSc; Henriette L Möllmann[2*], Dr med, Dr med dent; Milad Mirbabaie[3], Prof Dr rer pol; Stefan Stieglitz[4], Prof Dr rer pol

[1]Department of IT-Strategy, MAINGAU Energie GmbH, Obertshausen, Germany

[2]Department of Oral-, Maxillo- and Plastic Facial Surgery, Heinrich-Heine-University Duesseldorf, Duesseldorf, Germany

[3]Department of Business Studies and Economics, University of Bremen, Bremen, Germany

[4]Professional Communication in Electronic Media/Social Media, University of Duisburg-Essen, Duisburg, Germany

[*]these authors contributed equally

**Corresponding Author:**
Nicholas R J Frick, BSc, MSc
Department of IT-Strategy
MAINGAU Energie GmbH
Ringstraße 4
Obertshausen, 63179
Germany
Phone: 49 1711904454
Email: nicholas.frick@maingau-energie.de

## Abstract

**Background:** The COVID-19 pandemic has not only changed the private lives of millions of people but has significantly affected the collaboration of medical specialists throughout health care systems worldwide. Hospitals are making changes to their regular operations to slow the spread of SARS-CoV-2 while ensuring the treatment of emergency patients. These substantial changes affect the typical work setting of clinicians and require the implementation of organizational arrangements.

**Objective:** In this study, we aim to increase our understanding of how digital transformation drives virtual collaboration among clinicians in hospitals in times of crisis, such as the COVID-19 pandemic.

**Methods:** We present the lessons learned from an exploratory case study in which we observed the introduction of an information technology (IT) system for enhancing collaboration among clinicians in a German hospital. The results are based on 16 semistructured interviews with physicians from various departments and disciplines; the interviews were generalized to better understand and interpret the meaning of the statements.

**Results:** Three key lessons and recommendations explain how digital transformation ensures goal-driven collaboration among clinicians. First, we found that implementing a disruptive change requires alignment of the mindsets of the stakeholders. Second, IT-enabled collaboration presupposes behavioral rules that must be followed. Third, transforming antiquated processes demands a suitable technological infrastructure.

**Conclusions:** Digital transformation is being driven by the COVID-19 pandemic. However, the rapid introduction of IT-enabled collaboration reveals grievances concerning the digital dissemination of medical information along the patient treatment path. To avoid being caught unprepared by future crises, digital transformation must be further driven to ensure collaboration, and the diagnostic and therapeutic process must be opened to disruptive strategies.

**KEYWORDS**

## Introduction

The impact of COVID-19, the disease caused by SARS-CoV-2, has not only changed our lives and interactions as human beings but has had substantial consequences for health care systems worldwide [1]. Above all, hospitals must slow the spread of SARS-CoV-2 and further ensure a certain level of daily routine to maintain regular care and treatment of emergency patients. In addition to changes to clinical routines [2,3], this process involves adjustments in the way that clinicians work together.

Collaboration among hospital employees, particularly between physicians and interconnected medical departments, is essential for the overall diagnosis and treatment process [4-7]. Because the health care system is interdisciplinary in nature, goal-driven collaboration enhances the performance of services, quality of care, and patient outcomes [8-11]. Collaboration in hospitals can be defined as a professional alliance between health care specialists from multiple medical disciplines with diverse backgrounds and varying expertise, who jointly provide benefits for patients [12,13]. Although collaboration is considered to be a top priority in hospitals [14], some common issues are frequently experienced by clinicians [15]. Insufficient communication, including improper exchange of medical information, is a prevalent concern [16]. Clinicians manage their tasks in isolation with a lack of mutual understanding, which prevents the sharing of knowledge across departmental boundaries and inhibits collaboration between disciplines [17,18]. Furthermore, hospitals operate in complex clinical infrastructures, with a wide range of information systems containing different medical information; these systems are disconnected from each other and operated by multiple professionals [19,20]. Consistent presentation of medical data is lacking, and the retrieval of information stored across diversified systems is a time-consuming procedure [21,22].

In addition to existing collaborative challenges in hospitals [15], clinicians are being confronted with issues emerging from the COVID-19 pandemic [23]. Medical workers must not only examine individuals who are infected with the novel virus but must also respond to normal emergencies and patient cases. The required response involves substantial dedication and demands the implementation of diverse organizational arrangements as well as safety precautions [24]. Changes in surgery include decreasing the number of elective procedures to gain more capacity for emergency interventions [25]. Staff caring for patients with COVID-19 are separated from those caring for other patients, and postoperative visits are suspended and replaced by telephone calls to prevent in-hospital spread [3]. Furthermore, clinics are divided into multiple teams to prevent quarantine of an entire department, thus providing patient care in the case of infection. The underlying objective of reducing interpersonal communication compensates for any shortfall and maintains clinical operations. However, due to these necessary adjustments, collaboration in hospitals has changed significantly and has rapidly shifted toward virtual environments. Clinicians have started exchanging information using social media or instant messaging, and meetings are increasingly being conducted using applicable and available technology [3,26].

The integration of technologies into existing processes is a key component of the digital transformation [27,28] that improves cross-functional collaboration and coordination among individuals [29,30]. Digital transformation is defined as "a process that aims to improve an entity by triggering significant changes to its properties through combinations of information, computing, communication, and connectivity technologies" [30]. Therefore, digital transformation addresses changes associated with the introduction of new information technology (IT) in current organizational structures [31,32]. Because digital transformation is concerned with improving collaboration between individuals [30], it is closely linked to the computer-supported cooperative work (CSCW) research stream. CSCW determines how technology can increase group communication and work efforts [33]; thus, it is a vital component for understanding the dynamics of digital transformation [34] and associated changes of collaboration in practice [30].

In health care, digital transformation deliberately seeks to answer the question of how the quality of care and its related services can be improved with technology, as both rely on accurate, relevant, integrated and quickly accessible information [35,36]. For example, recent research examined how processes and digital infrastructure must be aligned to drive IT-enabled innovations [37] and whether advanced technologies, such as the Internet of Things (IoT) and artificial intelligence (AI) [38,39], can be integrated to empower caregivers to make evidence-based clinical decisions [28]. Extant research is further concerned with the potential of health information technology (HIT) to improve patient outcomes while reducing costs [40]. One example of HIT that is fundamentally transforming health services is electronic health records (EHRs), which digitally capture patients' retrospective, concurrent, and prospective information to guide medical treatment [41]. In this realm, scholars are exploring how technology is creating smart hospitals [42], that is, clinical environments with optimized and automated processes based on technological advancements and intelligent facilities to introduce new capabilities and provide an ideal surrounding for patients [42-44].

However, the increasing application of technology also serves as a basis for enhancing information exchange and collaboration in hospitals [15,45-47]. Collaborative visioning promotes joint decision-making and helps to overcome related issues frequently experienced by clinicians [48]. Communication is improved as information is presented transparently for involved experts. For example, central platforms coordinate the diffusion of reports and information exchange among participants [29,49]. The free transfer of medical information across departmental boundaries further establishes a closer alignment with collaborators from other departments, which dissolves silo thinking [50,51]. In addition, the introduction of standards simplifies complex clinical infrastructures and contributes to a feasible exchange of medical data between detached HIT systems; it also assists clinicians in gathering information more quickly [28].

In summary, the technological capacity for enhancing information exchange and collaboration focusing on patient outcomes is widely available [45-47]. Digital transformation and associated technologies are essential to facilitate the

exchange of information among clinicians across organizational boundaries, as new forms of collaborative work drive operational performance [15,30,31]. However, in their complex clinical environments, hospitals traditionally struggle to adopt new technologies [19,20], especially when aligning IT-enabled transformations with existing infrastructure [52]. Agarwal et al [17] stated that "an IT-enabled transformation of health care is just beginning, and it cannot happen too fast." However, the present situation of the COVID-19 pandemic is confronting hospitals with the urgent need to identify valid solutions while maintaining a sufficient level of collaboration and complying with the legal requirements of social distancing. Radical actions and restrictions are unavoidable for hospitals; however, they involve profound implications. Habitual collaboration is altered instantly, with no or only partial consideration of possible influencing factors such as available resources or security and data protection. Furthermore, the entire process of integrating technologies is likely to be faster during a crisis than under normal conditions, as people are conducting sense-making on a different level [53,54]. Disruptive technologies that are being applied to combat the COVID-19 pandemic are replacing traditional communication in health care teams [55] and are fundamental for sustained collaboration in hospitals. The ongoing pandemic is therefore creating a pressing need to rethink collaboration between clinicians; however, research on how digital transformation advocates collaboration in hospitals during a crisis is still lacking. We argue that this topic is of great interest to researchers and practitioners because virtual collaboration in hospitals will increase even further in the near future and persist for a long time. Thus, our research is guided by the following question: *How is digital transformation driving collaboration among clinicians in a time of crisis?*

We report the unique findings of an exploratory case study of a German hospital, in which we observed the introduction of an IT system to enhance collaboration among physicians during the COVID-19 pandemic. Our results are based and on 16 generalized semistructured interviews with physicians from various departments. We explain how clinicians worked together before the ongoing crisis, what has changed as a result of the pandemic, how the introduction of an IT system ensured effective collaboration, and how the digital transformation should continue to transform hospital operations.

We aim to increase our understanding of how digital transformation can improve collaboration among clinicians in the subject hospital and other hospitals. Moreover, it will demonstrate how the proposed IT improvements can ensure accurate, safe, and effective patient care not just for future crises but for daily operations. Researchers will be able to use our key lessons to understand the difference between a regular installation and one made out of necessity, along with their advantages and disadvantages. Practitioners will be able to understand how our recommendations help to ensure goal-driven collaboration and how hospitals can benefit from them. We hope to guide decision makers who want to introduce IT to improve collaboration between clinicians and stimulate additional research in this important field by expanding the body of knowledge.

## Methods

The German health care system is a highly developed sector in which the health of the population and life expectancy has continuously risen [56]. In the fight against SARS-CoV-2, wide-ranging countermeasures were backed by the German government and introduced at an early stage. A crisis taskforce was convened on February 27, 2020, and the first nationwide restrictions on public access were adopted on March 22 [57]. There is no evidence that the system was overburdened, as only 60% of the total of 33,051 intensive care beds were in use [58]. However, the new restrictions undoubtedly influenced hospitals and the ways in which their employees interacted with each other. To determine how collaboration between clinicians changed during the COVID-19 pandemic, we observed the introduction of an IT system at a large German hospital during the period from March 24 to April 24, 2020.

The hospital decided to introduce and provide Microsoft Teams (Microsoft Corporation) throughout the clinic. It was necessary to maintain collaboration between physicians, including those who were working from home, during the pandemic. Furthermore, restricted personal contact was needed to slow the spread of the virus. Microsoft Teams is a collaboration platform with features such as chatting, video calls, and file sharing. It can be used by many devices, such as personal computers, notebooks, tablets, and smartphones [59]. In response to the COVID-19 outbreak, Microsoft decided to offer the Teams software for free to help individuals stay connected [60].

Additionally, several organizational changes were initiated by the hospital. The departments were split into two different teams and worked in a weekly rotation to ensure that if one team was contaminated with the virus, the second team would be able to continue working and providing appropriate patient care. Furthermore, all elective procedures were reduced to a minimum, and only urgent surgeries were performed. However, the most crucial action was the minimization of any face-to-face exchanges between clinicians.

As presented in Figure 1, the case study preceded the implementation and test phase conducted by the hospital's IT department. The decision to implement the corresponding emergency concept, including the rollout of Microsoft Teams, was made on March 17 by the hospital management. On March 24, the software was installed on the employees' computers. Installation was followed by a week-long test phase during which additional hardware, such as cameras, microphones, and speakers, was installed. Because Microsoft Teams had not been used in the clinic before, clinicians had little or no knowledge of how to use the system. We provided advice and tips exclusively remotely, as visits to the hospital were forbidden to prevent the virus from infecting other patients and spreading. Finally, the first formal meeting and the official start of our support took place on April 1.

After the new system was introduced, we conducted 16 semistructured interviews with physicians. The interviews were designed to provide "questioning guided by identified themes in a consistent and systematic manner" [61]. We developed a guide containing relevant questions in advance. Our goal was

to make the guidelines as comprehensive as possible because physicians are typically available for a limited time, especially during a crisis. The guide was divided into chapters, with initial questions and subsequent follow-up questions (see Table 1).

Because we were interested in the meaning of the experts' substantive statements and not in their linguistic habitus or expressions, which are not necessary for understanding the context, nor in their physical gestures or facial expressions, we decided to paraphrase the interviews instead of conducting verbatim transcriptions. The analysis of the interviews was based on the recommendations in the qualitative assessment of content analysis by Schilling [28]. Paraphrasing the data reduced the volume by removing unnecessary words to form short, concise sentences. We listened carefully to the interview recordings and paraphrased the content of the physicians' statements. Next, we generalized and reduced the content to better understand and interpret the meaning of the statements [62]. The subsequent categorization was guided by the recommendations of Mayring [29] for an organization of data

derived from the material itself. This inductive category formation approach described the data without predefined criteria, leading to an unbiased "understanding of the material in terms of the material" [63].

The physicians were aged between 25 and 42 years (mean 32.2 years, SD 4.4), with 9 female and 7 male experts, and their tenures ranged from 0.5-17 years (mean 4.3 years, SD 4.4). Of the participants, 3 were senior physicians and 13 were resident physicians from 3 different clinics. The interviews were conducted using Microsoft Teams, with which the physicians were by then familiar. The interviews lasted between 9:04 and 28:25 minutes and were recorded, analyzed, and deleted once the evaluation was finished to protect the privacy of the participants. An overview of our sample is outlined in Table 2. The results yielded novel insights on how physicians collaborate with the assistance of technology during a pandemic. They broadened our view on collaboration among physicians in hospitals.

**Figure 1.** Process of the introduction of Microsoft Teams in the hospital.
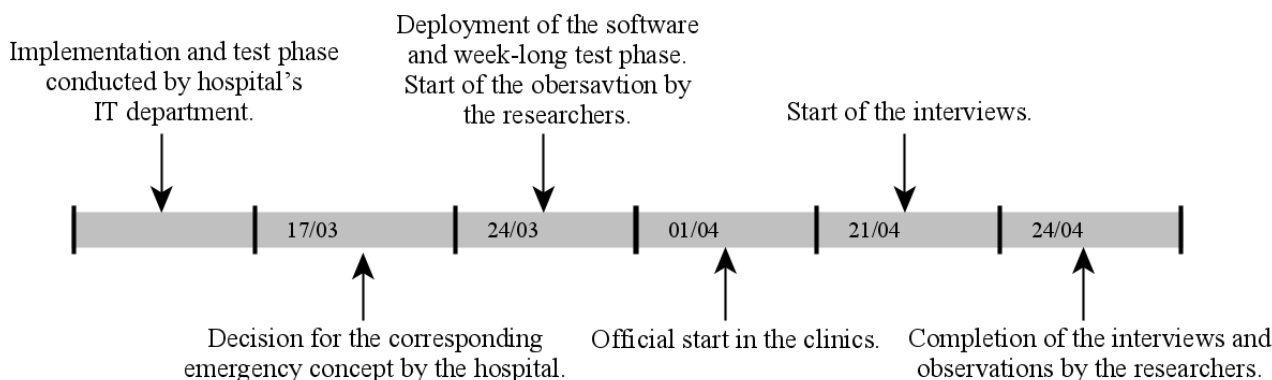
**Table 1.** Interview guideline.

| Chapter | Content or questions | Follow-up questions |
| --- | --- | --- |
| Before the interview | Introduction and summary of the purpose of this research, including an explanation of the interviewee's rights and verbal consent for the interview to be recorded | None |
| Demographic data | 1. What is your name and how old are you? | None |
| | 2. What is your current position at the hospital? | |
| | 3. What tasks does your current position involve? | |
| The conditions before the crisis/pandemic | 4. How was the collaboration with each other (with your colleagues) before the crisis/pandemic? | a. With whom did you primarily collaborate? |
| | | b. How did you collaborate? |
| | | c. Which IT[a] systems did you use for this purpose? |
| | 5. Where have been problems in the collaboration? | None |
| The conditions during the crisis/pandemic | 6. What was the goal of introducing Microsoft Teams in your hospital or clinic during the crisis/pandemic? | None |
| | 7. How has the introduction changed collaboration during the crisis/pandemic? | a. With whom do you collaborate? |
| | | b. How has your collaboration behavior changed? |
| | | c. What types of devices do you use for collaborating? |
| | | d. Have your tasks or responsibilities changed? |
| | | e. Has your culture within the clinic changed? Do you treat each other differently? |
| | 8. What advantages have resulted from the introduction in your collaboration? | a. Benefits for the team and/or individuals? |
| | | b. Benefits for other hospitals employees? |
| | | c. Benefits for patients? |
| | 9. What disadvantages have arisen from the introduction in your collaboration? | a. Surveillance? |
| | | b. Data protection? |
| | | c. Ethical issues or legal questions? |
| | 10. What challenges have been identified during the introduction? | a. Technical problems? |
| | | b. Acceptance, handling, resistance? |
| The conditions after the crisis/pandemic | 11. Would you like to receive further technical support for collaboration? | a. Further use of Microsoft Teams? |
| | 12. In which other areas would you like to get more support? | a. Artificial intelligence? |
| | | b. Intelligent systems? |
| | 13. What recommendations would you give other hospitals or their staff to improve collaboration in their clinic? | None |
| After the interview | Conclusion of the interview and the possibility for the expert to ask further questions or give closing remarks | None |

[a]IT: information technology.

**Table 2.** Sample overview of the expert interviews.

| Number | Age | Gender | Tenure (years) | Position | Clinic | Length (minutes:seconds) |
|--------|-----|--------|----------------|----------|--------|--------------------------|
| E1 | 31 | Female | 1.5 | Resident physician | Craniomaxillofacial surgery | 28:25 |
| E2 | 33 | Female | 0.5 | Resident physician | Craniomaxillofacial surgery | 19:21 |
| E3 | 42 | Male | 17 | Senior physician | Craniomaxillofacial surgery | 10:05 |
| E4 | 32 | Male | 5 | Resident physician | Craniomaxillofacial surgery | 17:31 |
| E5 | 26 | Male | 0.75 | Resident physician | Orthopedics and trauma surgery | 12:36 |
| E6 | 34 | Female | 6 | Resident physician | Craniomaxillofacial surgery | 09:04 |
| E7 | 25 | Female | 0.3 | Resident physician | Otorhinolaryngology | 10:57 |
| E8 | 33 | Female | 1 | Resident physician | Otorhinolaryngology | 09:34 |
| E9 | 29 | Male | 2 | Resident physician | Orthopedics and trauma surgery | 14:41 |
| E10 | 28 | Female | 2.5 | Resident physician | Orthopedics and trauma surgery | 14:41 |
| E11 | 37 | Male | 7 | Senior physician | Craniomaxillofacial surgery | 16:33 |
| E12 | 38 | Female | 11 | Senior physician | Craniomaxillofacial surgery | 14:40 |
| E13 | 29 | Female | 2.5 | Resident physician | Orthopedics and trauma surgery | 14:00 |
| E14 | 32 | Female | 3.5 | Resident physician | Otorhinolaryngology | 13:38 |
| E15 | 31 | Male | 3.5 | Resident physician | Otorhinolaryngology | 22:13 |
| E16 | 36 | Male | 4 | Resident physician | Orthopedics and trauma surgery | 15:17 |

## Results

This section describes the level of collaboration between clinicians before the crisis, what has changed because of the pandemic, and how it should be organized in the future. Excerpts from the German interviews have been translated into English for the reader's convenience. For the purpose of anonymization, the numbering of the interviewees does not correspond to the order of the interviewees as depicted in Table 2.

Figure 2 summarizes the results.

**Figure 2.** Summary of case study results.



Predeployment
- Collaboration primarily via personal contact followed by telephone and pager.
- Physical presence at meetings is mandatory.

During the COVID-19 Pandemic
- Collaboration assisted by Microsoft Teams.
- Reduction of personal contacts to help contain the virus.
- More effective and less stressful collaboration.
- Better patient-centered decision-making.
- Limited technological infrastructure and diverse stakeholders.

Future Vision
- Continue usage of Microsoft Teams.
- Expand digital support.
- Integrate in clinical processes.

XSL•FO
RenderX

## Predeployment Conditions

In every clinic in the hospital, an early meeting is held in which all physicians meet to discuss patients collaboratively and to derive diagnoses and therapies (eg, E1, E2, and E5). Physicians from other disciplines frequently participate by offering their medical expertise. Afterward, physicians conduct ward rounds, perform surgeries, or treat patients. In many disciplines, additional meetings are held in the afternoon to discuss treatment procedures for specific patients:

> *It depends a bit on the department, but it is generally discussed who is assigned where, the night shift is discussed, X-rays are viewed together with a radiologist, at noon there is another meeting in which current cases are discussed with the senior and chief physician.* [E5]

Collaboration between physicians in hospitals is characterized by face-to-face or telephone communication (eg, E4, E10, E12, E14). In many cases, personal communication is conducted on the same hierarchical level and within departmental boundaries. For example, information exchange among resident physicians takes place in person, while senior physicians or head physicians are usually approached via telephone:

> *Much usually happens verbally, with colleagues from the same level.* [E3]

> *Most personal communication with other resident physicians is according to our schedule. If I need to talk to a senior physician, I usually call him [or her].* [E2]

However, not every physician has their own phone; some have pagers (eg, E1, E6). If it is necessary to contact a physician, the number of their pager is called from any telephone in the clinic. The pager will then display the number of the caller with a request for a callback. The physician must find a telephone to respond to the call. Up to this point, the identity of the caller and the reason for the call remain uncertain. In our study, the physicians explained that no official IT system facilitates collaboration (eg, E1, E5, E12). It was also described that existing IT systems lack ease of use and are rather difficult to handle. Furthermore, the velocity of the current infrastructure needs improvement, and retrieving information is not straightforward because medical data are stored across systems (eg, E3, E6, E12):

> *In the rarest cases communication is done by email, but only when I am sitting in my office. There is no use of any IT system; we are rather old-school.* [E6]

> *Systems are frequently just slow. Handling is not intuitive and programs crash.* [E3]

The physicians reported that communication and collaboration in the hospital are sufficient but are dependent on collaborators adhering to certain organizational structures and communication channels (eg, E1, E7). Problems occasionally arise, such as misunderstandings in shift plans; colleagues can be difficult or impossible to reach, such as when a physician is in the operating theatre; or a faulty transfer from aftercare to day care can be made when information is captured in handwriting (eg, E2, E8, E9):

> *There are problems when changing shifts if a list has only been filled out by hand and you do not see each other.* [E8]

However, the physicians explained that no serious issues arise during habitual collaboration (eg, E9, E13, E14) but likewise admitted that the current mode of communication is not being questioned and certainly could benefit from adjustments (eg, E1, E11, E12).

## Collaboration During the COVID-19 Pandemic

Prior to the pandemic, Microsoft Teams was not used in the clinic, nor did any other standardized system exist for collaboration between clinicians in the hospital.

The technological support for the collaboration was equally accepted, but some behaviors changed. Due to the reduction of personal contact between clinicians, the communication was more professional and less social (eg, E2, E6, E10, E14, E16). Communication over the telephone decreased significantly, and many issues were resolved directly via chat. In addition, response times were shorter, and information was shared digitally (eg, E3, E8, E10, E14):

> *It is already a somewhat different culture which is much more impersonal. There is less personal interaction.* [E5]

The experts saw various advantages in technical assistance for collaboration. Foremost was compliance with the current rules and hygiene regulations (eg, E1-E6). Collaboration between physicians became easier. Data could be shared directly with everyone, and information was immediately captured digitally. Physicians could access data from anywhere, and even a physician who had not attended a meeting could obtain information via digital protocols. Colleagues could be reached much more quickly, and interdepartmental communication was simplified (eg, E5, E9, E12):

> *My colleagues are easier to reach, there is a closer connection between us. Medical information is available for everyone and decisions can be made faster and better.* [E8]

In addition, the IT system simplified group meetings. The experts often complained that in large meeting rooms, the presentation screen was simply too far away (eg, E1, E2, E16). Relevant information was more visible; for example, patient x-rays could be interpreted more easily. Resident physicians reported a learning effect from the joint discussions of pictures (eg, E9, E13, E16):

> *When you look at the pictures, you get more opinions, they can be used as a basis for a decision, someone can also intervene, if you can see the pictures well, there is also a learning effect.* [E4]

Meeting via Microsoft Teams was also perceived by the physicians as less stressful and more efficient (eg, E5, E7). Information could be looked up quickly during an appointment. Subsequent discussions could be kept within limits, and there was no longer a need to wait for participants on their way to a meeting room (eg, E8, E13). In addition, there was no need to use unknown hardware from an unfamiliar room for a

presentation. Furthermore, experts reported that one could better prepare for transferring patient information:

> *As a physician, you can prepare yourself better, the information is already there. Everyone can prepare themselves and you don't have to do it together.* [E4]

In addition, the physicians recognized indirect advantages for patients. The more effective and efficient appointments saved time, which could be used for more relevant tasks such as diagnosis or planning treatment. Through the joint discussion of patient reports, several professional opinions could be obtained, possibly leading to better decision-making (eg, E6, E12, E14). Finally, in the time of the pandemic, the potential risk of infection was minimized:

> *I believe we physicians are the greatest threat to everyone.* [E9]

This new type of collaboration requires cognitive effort and a high level of self-discipline. Physicians have a highly interactive job. During a pandemic, they are not always allowed to meet a colleague and must encourage themselves to participate in digital communication. There is a feeling of a need to be constantly available, which creates stress and the perception of being controlled (eg, E4, E8, E9). A physician is always expected to be up to date, even when working from home or when on vacation:

> *But it is expected that you always look in and up to date, even if you are at home you still look in there, you are afraid to miss something out.* [E7]

Finally, challenges of the collaborative adjustment were described. In addition to the partially slow internet connection, the new hardware had to be purchased quickly. Very few computers were equipped with proper cameras, microphones, or loudspeakers; therefore, collaboration with colleagues was more difficult (eg, E1, E2, E12):

> *Everybody had to install the software by himself and hardware was only available little by little. I couldn't log in from the office. The internet connection was bad so that people couldn't understand each other.* [E2]

The system had many diverse users, such as resident and senior physicians. The system was used with varying degrees of intensity and not to its full extent. The full range of functions of the software was not manageable from the beginning, and the users were required to learn the functions themselves (eg, E2, E7, E12):

> *Everyone used it the way they thought was right. We just got it.* [E11]

## A Vision of the Future

In the last part of the interviews, we asked the physicians how a future collaboration supported by technology would look. Some stated that Microsoft Teams should continue to be used, as it is broadly accepted and adopted (eg, E1, E2, E7). In addition, the change from classical messaging apps such as WhatsApp toward Microsoft Teams for professional purposes was already visible (eg, E7, E8):

> *Organizational things are shared different. It just works better.* [E7]

However, the different types and levels of use must be considered. On the one hand, the technological support for collaboration should be further integrated. For example, physicians tend to continue virtual meetings in the future, as those are particularly helpful when individuals from different clinical disciplines need to participate and the physical attendance of participants is not required (eg, E3, E4, E11). In addition, information exchange among physicians is likely to change, as medical data are easier to retrieve (eg, E6, E13):

> *I would like to keep the tumor board meetings going like this. Everyone sits at their own workplace; x-ray images are easier to see and disease patterns better to recognize; ad-hoc information for a specific case can be obtained quickly.* [E5]

> *I think we should continue to use this especially for appointments that are attended by many people from different disciplines or in the future for such things as e-learning on surgery techniques.* [E3]

> *There are some useful areas where we can use [Microsoft] Teams. A lot of our digital communication is conducted via email. I think this might be limitable or even replaceable for internal information exchange.* [E6]

On the other hand, personal contact is still fruitful and frequently the more favored means of collaboration. Especially for younger physicians, who are at the beginning of their medical training and rather inexperienced, personal contact is indispensable for conveying medical knowledge and demonstrating practical techniques on patients (eg, E1, E15, E16). Furthermore, mere virtual communication may result in negative consequences for teamwork and team spirit in the hospital (eg, E2, E5, E13):

> *When someone explains something to me, it's easier to understand when we meet in person. I can't imagine how virtual meetings look like when treating or operating patients where it is about the practical execution.* [E15]

> *We spend so much time together, I sometimes want personal interaction with my colleagues. Unfortunately, this is currently somewhat lost.* [E13]

> *I think [virtual collaboration] is especially beneficial when you already know each other and each other's work.* [E5]

Therefore, not every type of collaboration should be supported or replaced by technology. However, the physicians explained that the current pandemic is forcing hospitals to undertake collaborative alignments (eg, E4-E8) as necessary endeavors to limit personal contact to a minimum, some of which should have been undertaken even sooner (eg, E4, E9):

> *We should continue this kind of virtual collaboration. I think we need to cope with the restrictions [due to the pandemic] for some time.* [E10]

> *We need the implementation as quickly as possible during a pandemic.* [E7]

Finally, technology may improve not only collaborations among physicians (and other clinicians) but those with external partners. For example, some physicians reported that scientific collaborations have been initiated for research purposes (eg, E1, E2) with participants across departmental boundaries, that is, from internal clinics, external laboratories, and independent institutions.

## Discussion

In this section, we present the most relevant insights gained from our case study. We explain general key lessons learned and offer recommendations on how digital transformation can improve collaboration. Our goal is to guide hospitals and their decision makers who want to embrace digital transformation for collaborative purposes.

### Lesson 1: Organizational Change Requires Alignment of the Mindset

Digital transformation can change the way clinicians collaborate in hospitals. Achieving organizational change is not easy in surroundings that are characterized by hierarchical and authoritarian individuals [36]. IT-enabled transformation does not always proceed as desired [40], as working practices in hospitals have not greatly changed in the last few decades. Although the technological possibilities have matured and investments continue to rise [64], collaboration is still conducted with common, proven instruments [65].

In our study, the physicians' experience was that the assistance of an IT system simplified tasks in many areas of their daily working life. Simplification enhanced adaptations in the types of collaboration, moving from personal contact to a virtual environment. Aligning one's mindset required a high degree of self-control to avoid falling back into habitual patterns. One expert described this as "a different kind of communication that requires effort and self-discipline in a profession that is strongly characterized by personal interaction" [E5]. However, physicians observed that this alignment is necessary because "otherwise everyone wants to do it like we did before" [E2].

Recommendation: Hospitals should develop an overarching adoption strategy to meet the varying expectations of involved stakeholders. The introduction of technology today differs from that of 20 years ago. IT is simply a means to an end, and it should not be the center of attention. Instead, it provides a new way of working and of transforming certain processes. The entrenched mindset can only be changed piece by piece, accompanied by a variety of formats. This deliberate process might include different training sessions or workshops, such as how clinicians can collaborate to jointly develop an IT system to support, improve, or even completely replace certain processes.

### Lesson 2: Develop and Adhere to Behavioral Rules for Collaboration

Enabling new forms of collaborative work drives operational performance, not only within the clinic but across departmental boundaries [30,31]. The introduction of IT systems in hospitals is transformative and may be disruptive [66]. Furthermore, technological capabilities are broadly available to improve information exchange and collaboration [45-47]. The existence of technology does not necessarily describe how to handle that technology. Systems may be used in various ways; however, there is no "right" or "wrong" when it comes to the individual handling of a system.

In our work, physicians described a system that is used for different purposes with various functions. Some clinicians conducted video calls, while others used the chat function or shared files. Many questions remained unanswered: Which functions should be used at all? How quickly should one react to chat messages? Is a reaction also expected outside regular working hours? Which files are shared, and in what type of structure are they stored? Questions should be clarified in advance to ensure a structured collaboration. One expert explained that "there were communication problems, cameras were bought without anyone knowing. We need rules on how to act and how to behave" [E2].

Recommendation: Before or during the introduction of new technology, each clinic should define rules for its behavior and IT-enabled collaboration with stakeholders. Changes can never be initiated and sustained alone. The goal of improved collaboration and thus enhanced patient care can only be achieved if everyone participates and follows certain rules. The change should be embraced collaboratively.

### Lesson 3: Antiquated Technical Infrastructure Hinders a Shift

Achieving cross-functional collaboration among hospital physicians presupposes the integration of suitable technologies to improve the overall treatment of patients [27,28]. To enhance the overall diagnostic and treatment process further requires IT systems with stored information to be accurate, relevant and integrated as well as quickly accessible when needed [35,36]. However, hospitals still struggle to adopt new technologies in their complex clinical environments and to integrate them into their underlying infrastructure [52].

Our results revealed that physicians appreciated the IT-enabled support and wanted to continue to use it; however, they felt restricted due to technical problems. The participants reported that the internet connection in the hospital regularly failed or was only adequate for transmitting sound without a video signal. In addition, the computers were all stationary and frequently outdated. The lack of devices, in turn, led to many physicians using private devices, on which data protection and security were problematic. Physicians simply observed that "the infrastructure must be available" [E1] and "the technology must be upgraded and ideally interlinked in all parts of the health care system" [E4].

Recommendation: Hospitals must provide technical fundamentals to promote IT-enabled collaboration between physicians. This training applies not only to the infrastructure itself, such as network availability, but to the technical equipment. Every physician could be equipped with appropriate devices such as notebooks, tablets, or smartphones. A positive outcome of this would be that the equipment managed by the

XSL•FO

**RenderX**

clinic's IT department would comply with data protection regulations.

## Conclusions

Digital transformation, with all of its disruptive approaches, improves collaboration between clinicians, not only during crises such as the COVID-19 pandemic. Hospitals are slowly turning into smart hospitals that incorporate innovative technologies with optimized and automated processes to provide better patient care [42-44]. Technical advancements had been available before [45-47]; however, a crisis forced the introduction of IT-enabled assistance [55]. The reduction of personal contact among physicians to contain the virus and lower the infection rate was the ultimate objective.

In our study, physicians positively reviewed the introduction of the IT system to enhance collaboration in their hospital. However, it is necessary to ensure that the technical basis for the use of the system is available. Furthermore, it takes time for an organizational change to be embraced. Finally, rules for virtual collaboration should be defined and should be followed by every stakeholder.

The rapid and unavoidable introduction of IT-enabled collaboration revealed profound grievances concerning the technological requirements of hospitals and the digital expertise of clinicians. Digital dissemination of medical information along the patient treatment path is not entirely complete. Medical data are exchanged verbally or in handwriting and are often not available throughout the hospital [15,22]. Technology has found its way into our everyday lives, but physicians are still not fully aware of the possibilities that digital transformation reveals [65]. There is a lack of practice and actual use within hospitals, as IT-enabled collaboration is not necessarily needed, and the health care sector traditionally lags in adopting new technologies [30,67].

To avoid being caught unprepared by future crises, digital transformation must be further driven to ensure effective and efficient collaboration even without personal contact. Such virtual collaboration requires the diagnostic process and subsequent treatment to be opened to disruptive strategies. Openness will not only contribute to faster change but will embrace patient-oriented behavior. Moreover, a rapid rollout contributes to quicker adoption. The adoption of pragmatic and innovative solutions will increase their acceptance for future use.

## Conflicts of Interest

None declared.

## References

1. Lipsitch M, Swerdlow DL, Finelli L. Defining the Epidemiology of Covid-19 — Studies Needed. N Engl J Med 2020 Mar 26;382(13):1194-1196. [doi: 10.1056/nejmp2002125]

2. Shen Y, Cui Y, Li N, Tian C, Chen M, Zhang Y, et al. Emergency Responses to Covid-19 Outbreak: Experiences and Lessons from a General Hospital in Nanjing, China. Cardiovasc Intervent Radiol 2020 Jun;43(6):810-819 [FREE Full text] [doi: 10.1007/s00270-020-02474-w] [Medline: 32342164]

3. Wong J, Goh QY, Tan Z, Lie SA, Tay YC, Ng SY, et al. Preparing for a COVID-19 pandemic: a review of operating room outbreak response measures in a large tertiary hospital in Singapore. Can J Anaesth 2020 Jun 11;67(6):732-745 [FREE Full text] [doi: 10.1007/s12630-020-01620-9] [Medline: 32162212]

4. Gilbert J, Yan J, Hoffman S. A WHO report: framework for action on interprofessional education and collaborative practice. J Allied Health 2010;39 Suppl 1:196-197. [Medline: 21174039]

5. McInnes S, Peters K, Bonney A, Halcomb E. An integrative review of facilitators and barriers influencing collaboration and teamwork between general practitioners and nurses working in general practice. J Adv Nurs 2015 Sep;71(9):1973-1985. [doi: 10.1111/jan.12647] [Medline: 25731727]

6. Wagner EH, Austin BT, Davis C, Hindmarsh M, Schaefer J, Bonomi A. Improving chronic illness care: translating evidence into action. Health Aff (Millwood) 2001;20(6):64-78. [doi: 10.1377/hlthaff.20.6.64] [Medline: 11816692]

7. Rosen MA, DiazGranados D, Dietz AS, Benishek LE, Thompson D, Pronovost PJ, et al. Teamwork in healthcare: Key discoveries enabling safer, high-quality care. Am Psychol 2018 May;73(4):433-450 [FREE Full text] [doi: 10.1037/amp0000298] [Medline: 29792459]

8. D'Amour D, Goulet L, Labadie J, Martín-Rodriguez LS, Pineault R. A model and typology of collaboration between professionals in healthcare organizations. BMC Health Serv Res 2008 Sep 21;8:188 [FREE Full text] [doi: 10.1186/1472-6963-8-188] [Medline: 18803881]

9. Lenz R, Reichert M. IT support for healthcare processes – premises, challenges, perspectives. Data Knowl Eng 2007 Apr;61(1):39-58 [FREE Full text] [doi: 10.1016/j.datak.2006.04.007]

10. McCaffrey RG, Hayes R, Stuart W, Cassell A, Farrell C, Miller-Reyes C, et al. A program to improve communication and collaboration between nurses and medical residents. J Contin Educ Nurs 2010 Apr 01;41(4):172-178. [doi: 10.3928/00220124-20100326-04] [Medline: 20411889]

11. Zwarenstein M, Goldman J, Reeves S. Interprofessional collaboration: effects of practice‐based interventions on professional practice and healthcare outcomes. Cochrane Database Syst Rev 2009(3). [doi: 10.1002/14651858.cd000072.pub2]

12.  Schot E, Tummers L, Noordegraaf M. Working on working together. A systematic review on how healthcare professionals contribute to interprofessional collaboration. J Interprof Care 2020;34(3):332-342. [doi: 10.1080/13561820.2019.1636007] [Medline: 31329469]

13.  Morgan S, Pullon S, McKinlay E. Observation of interprofessional collaborative practice in primary care teams: An integrative literature review. Int J Nurs Stud 2015 Jul;52(7):1217-1230. [doi: 10.1016/j.ijnurstu.2015.03.008] [Medline: 25862411]

14.  Leggat SG. Effective healthcare teams require effective team members: defining teamwork competencies. BMC Health Serv Res 2007 Feb 07;7:17 [FREE Full text] [doi: 10.1186/1472-6963-7-17] [Medline: 17284324]

15.  Mirbabaie M, Stieglitz S, Frick N. Hybrid Intelligence in Hospitals - Towards a Research Agenda for Collaboration. Electron Mark 2021;forthcoming 2021.

16.  Kripalani S, LeFevre F, Phillips CO, Williams MV, Basaviah P, Baker DW. Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. JAMA 2007 Feb 28;297(8):831-841. [doi: 10.1001/jama.297.8.831] [Medline: 17327525]

17.  McCracken SS, Edwards JS. Implementing a knowledge management system within an NHS hospital: a case study exploring the roll-out of an electronic patient record (EPR). Knowl Manag Res Pract 2017 Dec 19;15(1):1-11. [doi: 10.1057/kmrp.2015.7]

18.  Driscoll M, Tobis K, Gurka D, Serafin F, Carlson E. Breaking down the silos to decrease internal diversions and patient flow delays. Nurs Adm Q 2015;39(1):E1-E8. [doi: 10.1097/NAQ.0000000000000080] [Medline: 25474675]

19.  Angst CM, Agarwal R, Sambamurthy V, Kelley K. Social Contagion and Information Technology Diffusion: The Adoption of Electronic Medical Records in U.S. Hospitals. Manage Sci 2010 Aug;56(8):1219-1241. [doi: 10.1287/mnsc.1100.1183]

20.  Hanseth O, Bygstad B. Flexible generification: ICT standardization strategies and service innovation in health care. Eur J Inf Syst 2017 Dec 19;24(6):645-663. [doi: 10.1057/ejis.2015.1]

21.  Meier J, Boehm A, Kielhorn A, Dietz A, Bohn S, Neumuth T. Design and evaluation of a multimedia electronic patient record "oncoflow" with clinical workflow assistance for head and neck tumor therapy. Int J Comput Assist Radiol Surg 2014 Nov 27;9(6):949-965. [doi: 10.1007/s11548-014-0988-x] [Medline: 24573598]

22.  Driving the digital transformation of Germany's healthcare system for the good of patients. Federal Ministry of Health, Germany. URL: https://www.bundesgesundheitsministerium.de/digital-healthcare-act.html [accessed 2021-01-04]

23.  Baughman AW, Hirschberg RE, Lucas LJ, Suarez ED, Stockmann D, Hutton Johnson S, et al. Pandemic Care Through Collaboration: Lessons From a COVID-19 Field Hospital. J Am Med Dir Assoc 2020 Nov;21(11):1563-1567. [doi: 10.1016/j.jamda.2020.09.003] [Medline: 33138938]

24.  Sun P, Lu X, Xu C, Sun W, Pan B. Understanding of COVID-19 based on current evidence. J Med Virol 2020 Jun;92(6):548-551 [FREE Full text] [doi: 10.1002/jmv.25722] [Medline: 32096567]

25.  Burgansky A, Coletta-Lucas J, Garcia MS. COVID-19 in a community hospital. Semin Perinatol 2020 Oct;44(6):151292 [FREE Full text] [doi: 10.1016/j.semperi.2020.151292] [Medline: 32800406]

26.  Archbald-Pannone LR, Harris DA, Albero K, Steele RL, Pannone AF, Mutter JB. COVID-19 Collaborative Model for an Academic Hospital and Long-Term Care Facilities. J Am Med Dir Assoc 2020 Jul;21(7):939-942 [FREE Full text] [doi: 10.1016/j.jamda.2020.05.044] [Medline: 32563752]

27.  Klinker K, Wiesche M, Krcmar H. Digital Transformation in Health Care: Augmented Reality for Hands-Free Service Innovation. Inf Syst Front 2019 Jun 25;22(6):1419-1431. [doi: 10.1007/s10796-019-09937-7]

28.  Gopal G, Suter-Crazzolara C, Toldo L, Eberhardt W. Digital transformation in healthcare - architectures of present and future information technologies. Clin Chem Lab Med 2019 Feb 25;57(3):328-335 [FREE Full text] [doi: 10.1515/cclm-2018-0658] [Medline: 30530878]

29.  Klötzer C, Pflaum A. Toward the Development of a Maturity Model for Digitalization within the Manufacturing Industry's Supply Chain. In: Proceedings of the 50th Hawaii International Conference on System Sciences. 2017 Presented at: 50th Hawaii International Conference on System Sciences; January 4-7, 2017; Waikoloa Village, HI p. 4210-4219. [doi: 10.24251/hicss.2017.509]

30.  Vial G. Understanding digital transformation: A review and a research agenda. J Strateg Inf Syst 2019 Jun;28(2):118-144. [doi: 10.1016/j.jsis.2019.01.003]

31.  Hess T, Benlian A, Matt C, Wiesböck F. Options for Formulating a Digital Transformation Strategy. MIS Q Executive 2016;15(2).

32.  Frick NRJ, Mirbabaie M, Stieglitz S, Salomon J. Maneuvering through the stormy seas of digital transformation: the impact of empowering leadership on the AI readiness of enterprises. Journal of Decision Systems 2021 Jan 06:1-24. [doi: 10.1080/12460125.2020.1870065]

33.  Koch M, Schwabe G, Briggs RO. CSCW and Social Computing. Bus Inf Syst Eng 2015 Mar 3;57(3):149-153. [doi: 10.1007/s12599-015-0376-2]

34.  Lewkowicz M, Liron R. The Missing "Turn to Practice" in the Digital Transformation of Industry. Comput Supported Coop Work 2019 Apr 24;28(3-4):655-683. [doi: 10.1007/s10606-019-09347-y]

35. Mihailescu M, Mihailescu D. The emergence of digitalisation in the context of health care. In: Proceedings of the 51st Hawaii International Conference on System Sciences. 2018 Presented at: 51st Hawaii International Conference on System Sciences; January 3-6, 2018; Maui, HI. [doi: 10.24251/hicss.2018.387]

36. Goh JM, Gao G, Agarwal R. Evolving Work Routines: Adaptive Routinization of Information Technology in Healthcare. Inf Sys Res 2011 Sep;22(3):565-585. [doi: 10.1287/isre.1110.0365]

37. Bygstad B, Øvrelid E, Lie T, Bergquist M. Developing and Organizing an Analytics Capability for Patient Flow in a General Hospital. Inf Syst Front 2019 May 4;22(2):353-364. [doi: 10.1007/s10796-019-09920-2]

38. Frick N, Brünker F, Ross B, Stieglitz S. Der Einsatz von künstlicher Intelligenz zur Verbesserung des Incident Managements. HMD 2019 Feb 15;56(2):357-369. [doi: 10.1365/s40702-019-00505-w]

39. Frick N, Brünker F, Ross B, Stieglitz S. Design requirements for AI-based services enriching legacy information systems in enterprises: A managerial perspective. In: Proceedings of the 31st Australasian Conference on Information Systems (ACIS). 2020 Dec 03 Presented at: 31st Australasian Conference on Information Systems (ACIS); December 1-4, 2020; Wellington, New Zealand.

40. Agarwal R, Gao G, DesRoches C, Jha AK. Research Commentary—The Digital Transformation of Healthcare: Current Status and the Road Ahead. Inf Sys Res 2010 Dec;21(4):796-809. [doi: 10.1287/isre.1100.0327]

41. Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. Int J Med Inform 2008 May;77(5):291-304. [doi: 10.1016/j.ijmedinf.2007.09.001] [Medline: 17951106]

42. Moro Visconti R, Martiniello L. Smart hospitals and patient-centered governance. COC 2019 Feb 27;16(2):83-96. [doi: 10.22495/cocv16i2art9]

43. Moro Visconti R, Morea D. Healthcare Digitalization and Pay-For-Performance Incentives in Smart Hospital Project Financing. Int J Environ Res Public Health 2020 Mar 30;17(7) [FREE Full text] [doi: 10.3390/ijerph17072318] [Medline: 32235517]

44. Bullen M, Hughes T, Marshall J. The Evolution of Nice Medtech Innovation Briefings And Their Associated Technologies. Value in Health 2017 Oct;20(9):A595. [doi: 10.1016/j.jval.2017.08.1114]

45. Ayal M, Seidman A. An Empirical Investigation of the Value of Integrating Enterprise Information Systems: The Case of Medical Imaging Informatics. J Manag Inf Sys 2014 Dec 08;26(2):43-68. [doi: 10.2753/mis0742-1222260203]

46. Jack E, Powers T. Volume Flexible Strategies in Health Services: A Research Framework. Prod Oper Manag 2009;13(3):244. [doi: 10.1111/j.1937-5956.2004.tb00508.x]

47. Khoumbati K, Themistocleous M, Irani Z. Evaluating the Adoption of Enterprise Application Integration in Health-Care Organizations. J Manag Inf Sys 2014 Dec 08;22(4):69-108. [doi: 10.2753/mis0742-1222220404]

48. Eden R, Burton-Jones A, Casey V, Draheim M. Digital Transformation Requires Workforce Transformation. MISQE 2019 Mar 1;18(1). [doi: 10.17705/2msqe.00005]

49. Kretzer M, Gaß O, Mädche A. Design Principles for Diffusion of Reports and Innovative Use of Business Intelligence Platforms. 2015 Presented at: WI 2015, the 12th International Conference on Wirtschaftsinformatik; March 4-6, 2015; Osnabrück, Germany p. 46.

50. Baiyere A, Salmela H, Tapanainen T. Digital transformation and the new logics of business process management. Eur J Inf Syst 2020 Mar 01;29(3):238-259. [doi: 10.1080/0960085x.2020.1718007]

51. Broadbent M, Weill P, Clair DS. The Implications of Information Technology Infrastructure for Business Process Redesign. MIS Q 1999 Jun;23(2):159. [doi: 10.2307/249750]

52. Bygstad B, Øvrelid E. Architectural alignment of process innovation and digital infrastructure in a high-tech hospital. Eur J Inf Syst 2020 Feb 27;29(3):220-237. [doi: 10.1080/0960085x.2020.1728201]

53. Stieglitz S, Mirbabaie M, Kroll T, Marx J. "Silence" as a strategy during a corporate crisis – the case of Volkswagen's "Dieselgate". INTR 2019 Aug 05;29(4):921-939. [doi: 10.1108/intr-05-2018-0197]

54. Mirbabaie M, Bunker D, Stieglitz S, Marx J, Ehnis C. Social media in times of crisis: Learning from Hurricane Harvey for the coronavirus disease 2019 pandemic response. J Inf Technol 2020 Jun 09;35(3):195-213. [doi: 10.1177/0268396220929258]

55. Abdel-Basset M, Chang V, Nabeeh NA. An intelligent framework using disruptive technologies for COVID-19 analysis. Technol Forecast Soc Change 2021 Feb;163:120431 [FREE Full text] [doi: 10.1016/j.techfore.2020.120431] [Medline: 33162617]

56. Rechel B, Maresso A, Sagan A, Hernández-Quevedo C, Williams G, Richardson E, et al. Organization and financing of public health services in Europe: Country reports. Copenhagen, Denmark: European Observatory on Health Systems and Policies; 2018. URL: https://www.ncbi.nlm.nih.gov/books/NBK507325/pdf/Bookshelf_NBK507325.pdf [accessed 2021-01-20]

57. Coronavirus SARS-CoV-2: Chronik der bisherigen Maßnahmen. Bundesministerium für Gesundheit. 2020. URL: https://www.bundesgesundheitsministerium.de/coronavirus/chronik-coronavirus.html [accessed 2020-04-30]

58. Deutsche Interdisziplinäre Vereinigung für Intensiv- und Notfallmedizin Intensivregister. DIVI. URL: https://www.intensivregister.de/#/intensivregister [accessed 2020-04-30]

59. Microsoft Teams. Microsoft. URL: https://www.microsoft.com/en-us/microsoft-365/microsoft-teams/group-chat-software [accessed 2020-04-26]

60.    Responding to COVID-19 together. Microsoft. URL: https://news.microsoft.com/covid-19-response/ [accessed 2020-04-26]
61.    Qu SQ, Dumay J. The qualitative research interview. Qualitative Res Acc & Man 2011 Aug 30;8(3):238-264. [doi: 10.1108/11766091111162070]
62.    Schilling J. On the Pragmatics of Qualitative Assessment. Eur J Psychol Assess 2006 Jan;22(1):28-37. [doi: 10.1027/1015-5759.22.1.28]
63.    Mayring P. Qualitative Content Analysis: Theoretical Background and Procedures. In: Bikner-Ahsbahs A, Knipping C, Presmeg N, editors. Approaches to Qualitative Research in Mathematics Education. Advances in Mathematics Education. Dordrecht, the Netherlands: Springer; 2015:365.
64.    Mettler T, Sprenger M, Winter R. Service robots in hospitals: new perspectives on niche evolution and technology affordances. Eur J Inf Syst 2017 Dec 19;26(5):451-468. [doi: 10.1057/s41303-017-0046-1]
65.    Győrffy Z, Radó N, Mesko B. Digitally engaged physicians about the digital health transition. PLoS One 2020;15(9):e0238658 [FREE Full text] [doi: 10.1371/journal.pone.0238658] [Medline: 32986733]
66.    Aanestad M, Vassilakopoulou P. Collaborative Innovation in Healthcare: Boundary Resources for Peripheral Actors. In: ICIS 2019 Proceedings. 2019 Presented at: International Conference on Information Systems (ICIS) 2019; December 15-18, 2019; Munich, Germany.
67.    Lucas, Jr. HC, Agarwal R, Clemons EK, El Sawy OA, Weber B. Impactful Research on Transformational Information Technology: An Opportunity to Inform New Audiences. MISQ 2013 Feb 2;37(2):371-382. [doi: 10.25300/misq/2013/37.2.03]

## Abbreviations

**AI:** artificial intelligence
**CSCW:** computer-supported cooperative work
**EHR:** electronic health record
**HIT:** health information technology
**IoT:** Internet of Things
**IT:** information technology

Viewpoint

# Blockchain-Based Digital Contact Tracing Apps for COVID-19 Pandemic Management: Issues, Challenges, Solutions, and Future Directions

Sheikh Mohammad Idrees[1]; Mariusz Nowostawski[1]; Roshan Jameel[2]

[1]Department of Computer Science, Norwegian University of Science and Technology, Gjovik, Norway

[2]Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India

**Corresponding Author:**
Sheikh Mohammad Idrees
Department of Computer Science
Norwegian University of Science and Technology
Teknologivegen 22
Gjovik, 2815
Norway
Phone: 47 46248610
Email: sheikh.idrees99@gmail.com

## Abstract

The COVID-19 pandemic has caused substantial global disturbance by affecting more than 42 million people (as of the end of October 2020). Since there is no medication or vaccine available, the only way to combat it is to minimize transmission. Digital contact tracing is an effective technique that can be utilized for this purpose, as it eliminates the manual contact tracing process and could help in identifying and isolating affected people. However, users are reluctant to share their location and contact details due to concerns related to the privacy and security of their personal information, which affects its implementation and extensive adoption. Blockchain technology has been applied in various domains and has been proven to be an effective approach for handling data transactions securely, which makes it an ideal choice for digital contact tracing apps. The properties of blockchain such as time stamping and immutability of data may facilitate the retrieval of accurate information on the trail of the virus in a transparent manner, while data encryption assures the integrity of the information being provided. Furthermore, the anonymity of the user's identity alleviates some of the risks related to privacy and confidentiality concerns. In this paper, we provide readers with a detailed discussion on the digital contact tracing mechanism and outline the apps developed so far to combat the COVID-19 pandemic. Moreover, we present the possible risks, issues, and challenges associated with the available contact tracing apps and analyze how the adoption of a blockchain-based decentralized network for handling the app could provide users with privacy-preserving contact tracing without compromising performance and efficiency.

## Introduction

COVID-19, caused by the novel coronavirus SARS-CoV-2, was unknown prior to its spread in late 2019 in Wuhan, China [1]. The infection began with 27 people in Wuhan, which has now grown to more than 42 million people globally [2]. Within a month of the initial outbreak, COVID-19 was declared a global public health emergency by the World Health Organization (WHO) in January 2020. SARS-CoV-2 is usually transmitted from infected person to healthy person via physical contact [3]. Several studies also claim that the disease is highly contagious and can be transmitted through air particles, which contributes to its uncontrollable spread. Infected people can be symptomatic or asymptomatic, but both types of patients can transmit the virus. Therefore, the only way to control the spread is to keep infected individuals in isolation for 14 days, as the incubation period ranges between 1-14 days, per research carried out in Wuhan [4]. Since COVID-19 appeared suddenly, there was no information about treatment and prevention. Although clinical trials and research are being carried out globally to deal with this pandemic, the availability of a vaccine to the public may not be possible in the near future. Therefore, as of now,

everybody must follow some simple practices every day like avoiding gatherings, maintaining social distance, washing hands, using hand sanitizers, and wearing masks and gloves when going out.

To combat this coronavirus, governments throughout the world are implementing several measures to avoid social contact among people to minimize spread. Complete lockdowns were imposed in several countries in which national and international borders were sealed, schools and universities were shut down, employees were asked to continue their work from home, malls and markets were closed, and gatherings and functions were suspended; all these efforts are being made to restrict human contact as much as possible. Nevertheless, the impact of these actions on the economy has raised global concerns, which makes it important to have a balance between prevention mechanisms and economic activities. Health organizations globally are keeping manual track of people who might have come into contact with COVID-19–positive persons, but this process is time consuming, prone to errors, and inefficient. Therefore, a digital contact tracing system is needed to identify, assess, and manage the people and locations that have been exposed to COVID-19–infected patients in order to prevent the spread of the virus and break the chain [5]. Digital contact tracing is a mechanism that utilizes the technological concepts as a tool to collect data to identify contacts and prevent transmission [6].

However, contact tracing means the continuous monitoring of data such as personal details and location of users and infected patients, which creates a sense of fear in society that their movements are being captured and broadcasted to the app managers. Furthermore, the privacy of individuals' data becomes a concern in a pandemic situation, as people become unsure how the government is going to use this collected information. In this paper, an outline is provided on how containment in pandemic situations such as COVID-19 can be achieved using digital contact tracing and elaborate on the issues and challenges that this mechanism imposes on society. A brief description is also provided on the contact tracing systems implemented specifically for COVID-19, followed by a discussion on blockchain-based decentralized contact tracing, which can handle users' data in a transparent and immutable manner, thus providing security as well as ensuring privacy.

## Digital Contact Tracing

It is necessary to have containment in order to limit the spread of communicable diseases and prevent them from reaching the level of a pandemic like COVID-19. Contact tracing is a mechanism of identifying infected people, keeping track of others who might have come in contact with them, and collecting relevant information about these contacts to take timely measures to minimize the spread of contagious diseases [6,7].

Traditional contact tracing has been employed in the past for controlling the spread of several communicable diseases. The process of traditional contact tracing is based on keeping track of people an individual has been in touch with, as well as places they have visited in the past few days or weeks. Traditionally, these details were kept on paper files, and people with a chance of infection were informed via mail or phone call. This is a tedious and error-prone method that demands too much labor. With time, everything has been digitized, including the process of contact tracing. Digital contact tracing is a mechanism that can keep track of these details through mobile phones, which can be used to identify people who might have come into contact with an infected patient [6]. The WHO has divided the mechanism of contact tracing into three steps as follows [8]:

1. Identification of contacts: when a person is diagnosed with a communicable viral disease, their contacts are identified by asking about the routine and activities of the infected person. These contacts could be anyone who had come in contact with the patient, such as family members, friends, colleagues, or health care practitioners.
2. Listing of contacts: after identifying the individuals who have come in contact with the infected person in the past few days, a list of these contacts is compiled. These contacts are then informed of the situation and guided about the practices that they must follow if they develop any symptoms. The contacts are also advised to remain under isolation or quarantine, if required.
3. Follow-up with contacts: periodic and regular follow-ups with the contacts are mandatory as they help in the timely monitoring of health symptoms and test results and prevent the further spread of the disease.

Nowadays, contact tracing is being implemented through smartphones. Development and deployment of several mobile apps are being done to resolve the problems associated with the traditional, manual contact tracing approaches. Bluetooth technology is one of the mainstream approaches for contact tracing apps to uncover the individuals who might have been in proximity of infected people. The Bluetooth-based approaches do not store the location or data of users; they only notify infected persons' contacts with minimal interference to the user's privacy. Such techniques could be centralized or decentralized. In case of decentralized techniques, the details are stored on the user's device only, which gives users both power and control over their personal data. The analysis and processing of the data for tracking contacts are also done on the user's device, which promotes transparency, privacy, and user consent [9]. In centralized apps, the analysis and processing are done on a central server, and users are alerted if required. Figure 1 depicts how digital contact tracing works in decentralized and centralized setups.

XSL•FO

**RenderX**

**Figure 1.** Digital contact tracing mechanism.
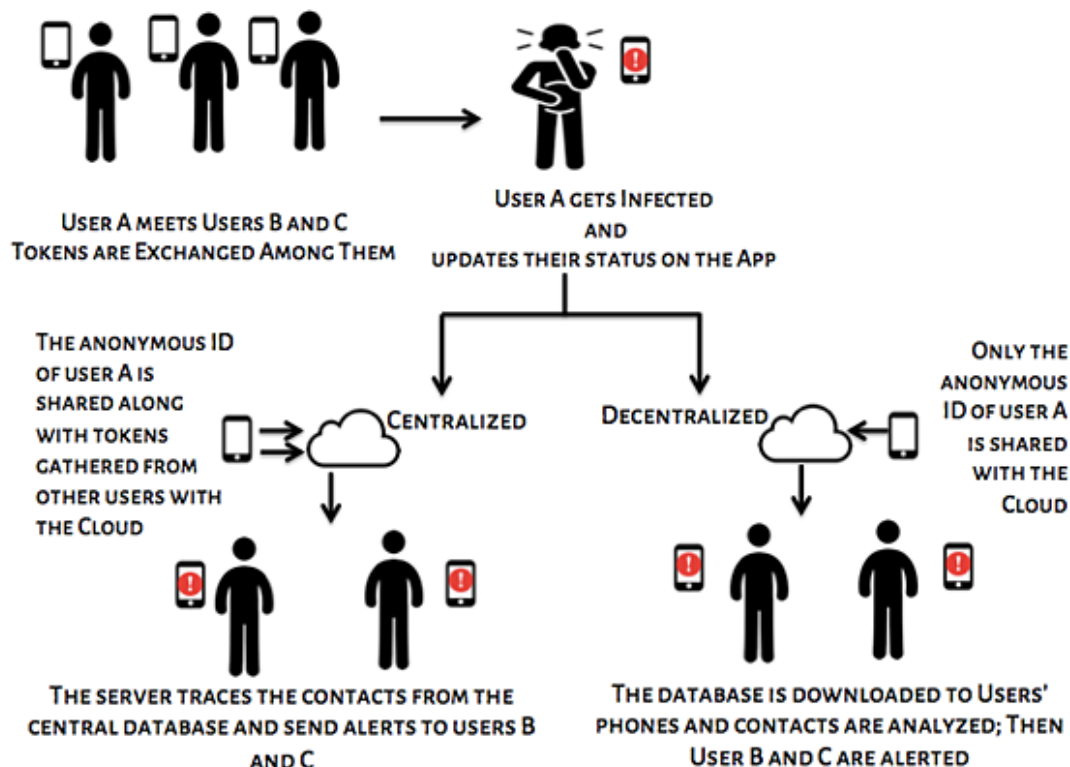


## Existing Digital Contact Tracing Techniques for COVID-19

There are several apps developed recently to effectively handle the COVID-19 pandemic. These apps help health care workers as well as the public to gain insight into the situation so as to make suitable decisions. Some of these apps are discussed below, followed by a comparison depicted in Table 1.

### TraceTogether

TraceTogether is an app based on BlueTrace protocols that utilizes Bluetooth low-energy technology to determine and record the proximity details of users [10]. In order to run this app, users need to keep their mobile devices in the active state, which allows for the broadcasting of their location at all times. This, however, affects battery usage. Bluetooth has always been vulnerable to security attacks (eg, sniffers), and the chances of replay attacks are also high, which might generate false information leading to panic among the public. Furthermore, concealing the hardware device associated with Bluetooth technology is difficult and can lead to the exposure of user identity. Moreover, the proximity limit of Bluetooth is limited. The TraceTogether app works on a centralized network, which means that the identities of users might not be revealed to other users, but this identity is known to a central authority and an attacker might gain control of the data by hacking the central node of the network (single point of failure).

### COVID Trace

Developed by Apple and Google, COVID Trace is an app that is based on Bluetooth technology that only collects necessary information like location with respect to time and does not retain any recognizable user data (eg, identity), thereby keeping the data anonymous [11]. In this app, if 2 individuals stay in proximity to one another for more than 10 minutes, tokens are exchanged and stored in the app. Whenever a person is diagnosed with COVID-19, this information is shared with the app server, which is referred to as escrow. These generated tokens are only shared to the public database with the permission of the user. This app keeps the users' location records for the past 3 weeks, which is recorded only when the app user leaves their home. On the basis of the collected data, the symptoms, the area of exposure, and the number of days a person has been sick are stored. App managers are responsible for the anonymization of user location while uploading. Moreover, the exact time is not disclosed and is rounded off for privacy reasons. One of the drawbacks of this app is that the presence or absence of the disease is not verified. Nevertheless, it offers improved privacy properties as compared to storing raw Bluetooth device IDs such as in the previous example.

### HowWeFeel

This app was developed specially for researchers, health care professionals, and doctors. It collects and aggregates the symptoms and personal information of users for the purpose of sharing it with health care practitioners who are working to have a better understanding of COVID-19 and ways to combat it [12]. The collected data are utilized to better understand the nature of its spread and to identify people at risk of infection, in order to take the necessary steps on time. Data are collected for both healthy and unhealthy individuals. The app asks users certain questions in terms of how they are feeling and whether or not they have gotten a COVID-19 test, as well as the result of the test. Moreover, a unique token is issued for the identification of every user. Firewalls, antiviruses, and

cryptographic algorithms are also applied to ensure authorized access.

## NOVID

NOVID is another app developed for contact tracing that utilizes the concepts of Bluetooth technology and ultrasonic sound waves for accurately counting the number of contacts to provide exact distance measurement [13]. The microphones of mobile phones are utilized in this app to listen to inaudible sounds by exploiting Bluetooth technology whenever a device passes another. Then the proximity among users is calculated without demanding any personal information or location. Whenever a new user is registered on the app, a unique user ID is generated randomly to identify the device. The user information gathered by this app does not link it with the users' personal details such as name, email, or device without the user's permission. The IDs of the tracked users are gathered and stored on a server, and if a person is diagnosed with COVID-19, then the user ID is encrypted and then sent to the other users of the app.

## COVID Shield

The COVID Shield app collects the IDs of individuals using Bluetooth technology in a random manner and shares it with nearby users [14]. If a person is found to be COVID-19 positive, then their ID is shared anonymously with other users with a chance of exposure to the disease. The app downloads these randomly generated IDs from the server after an interval of time, and these IDs are mapped with user information extracted from personal devices to determine one's chances of exposure to the virus in the past few days. COVID Shield encompasses a web portal, a server, and a mobile app. No personal information is directly collected on the app; instead, a random ID is assigned along with temporary keys and logs. The app uses the Google and Apple Exposure Notification framework to maintain the privacy of users' data. The collected data are stored on AWS (Amazon Web Services) cloud, and users have complete jurisdiction over their data. Users are able to turn their notifications off and delete their exposure log.

## Zero

The Zero app is installed on the user's mobile device, through which data and user ID are collected and then stored on a centralized cloud storage provided by Google Cloud [9]. This app supports API (application programming interface) and TCN (Temporary Contact Numbers) protocols provided by both Apple and Google [15]. App managers can share the collected data with other entities for lawful purposes such as law enforcement or clinical research. The data remain anonymous while being shared. In order to maintain the security of the data, SSL (Secure Socket Layer) protocols are used. The users have the right to file a complaint with the app managers if their data are compromised in any way. However, the app providers do not assure the privacy of the collected information.

## ShareTrace

When a new user registers on the ShareTrace app, a random ID is generated and assigned. Personal information, such as email, name, and location, is kept safe on the user's device, and the mobile phone of the user communicates with the app server using the assigned ID only [16]. Contact tracing is done using Bluetooth technology, which shares users' packets (ie, Bluetooth ID of the device, symptoms, diagnosis, and contact history associated with the device) when 2 individuals come into physical contact with one another. The mobile devices of the users keep track of contact tracing along with the symptoms and infected people. The information is updated on the app server periodically, and the analysis on the data is performed while maintaining the privacy of user identities.

## Safe2

This app uses the integration of both Bluetooth technology and GPS for efficient contact tracing. Bluetooth is utilized to detect droplets of coughs and sneezes, while GPS is used to trace the surfaces in contact with the infected individuals. Self-assessment results are stored and never leave the user's device until a positive symptom is detected [17]. When the user gets infected, all the data (anonymized close contacts, location history, and exposure status) are uploaded to the alerting system of the app, which keeps user identity anonymous and notifies other users who have been in close proximity with the infected person. The Safe2 app managers have the right to release the user's personal data to the government or law enforcement agencies if required.

## Aarogya Setu App

This app was developed by the Government of India for the purpose of COVID-19 contact tracing. However, it remained controversial for a long time because people were worried about their privacy. This forced the Indian government to come out with a privacy policy for the app, elaborating on the type of data that is collected and how the data are used. Prominent features include automatic contact tracing via Bluetooth, self-assessment tests, travel advisories, risk status, location data, nationwide COVID-19 status, emergency helpline numbers, and more [18,19]. When an Aarogya Setu app user comes within the Bluetooth proximity of another user, the two devices securely exchange digital signatures related to this interaction. It contains the location data, time, proximity, and duration. These data are stored on the devices of all individuals. When a person tests positive, an alert notification is sent to all users who have been in proximity to that individual within the last 14 days and recommends suitable action. The updated risk of infection is closely monitored by the Government of India.

## Exposure Notification

Google and Apple have recently announced a joint effort to fight COVID-19 by using Bluetooth technology for contact tracing to help governments and health care organizations to contain the spread of this virus. Called Exposure Notification, this system has been designed keeping in mind the importance of user privacy and security. A random ID is generated when a user installs the Exposure Notifications system on their mobile phone. To preserve the user's privacy, these random IDs change every 10-20 minutes so that the identity of the user or the geographical location cannot be detected. The user's mobile phone and those nearby will then continue to exchange these random IDs via Bluetooth. Data are collected, stored, and processed on the user's mobile phone only. If, at any point in time, a user is diagnosed as positive for the virus, he or she updates their status in the app. Other users' devices concurrently

and occasionally match all the random IDs with positive COVID-19 cases against its own random IDs. During the whole process, the user's identity is not shared with anyone—not even with Google and Apple. Figure 2 outlines how the Exposure Notification system works [20].

**Figure 2.** The Exposure Notification system.
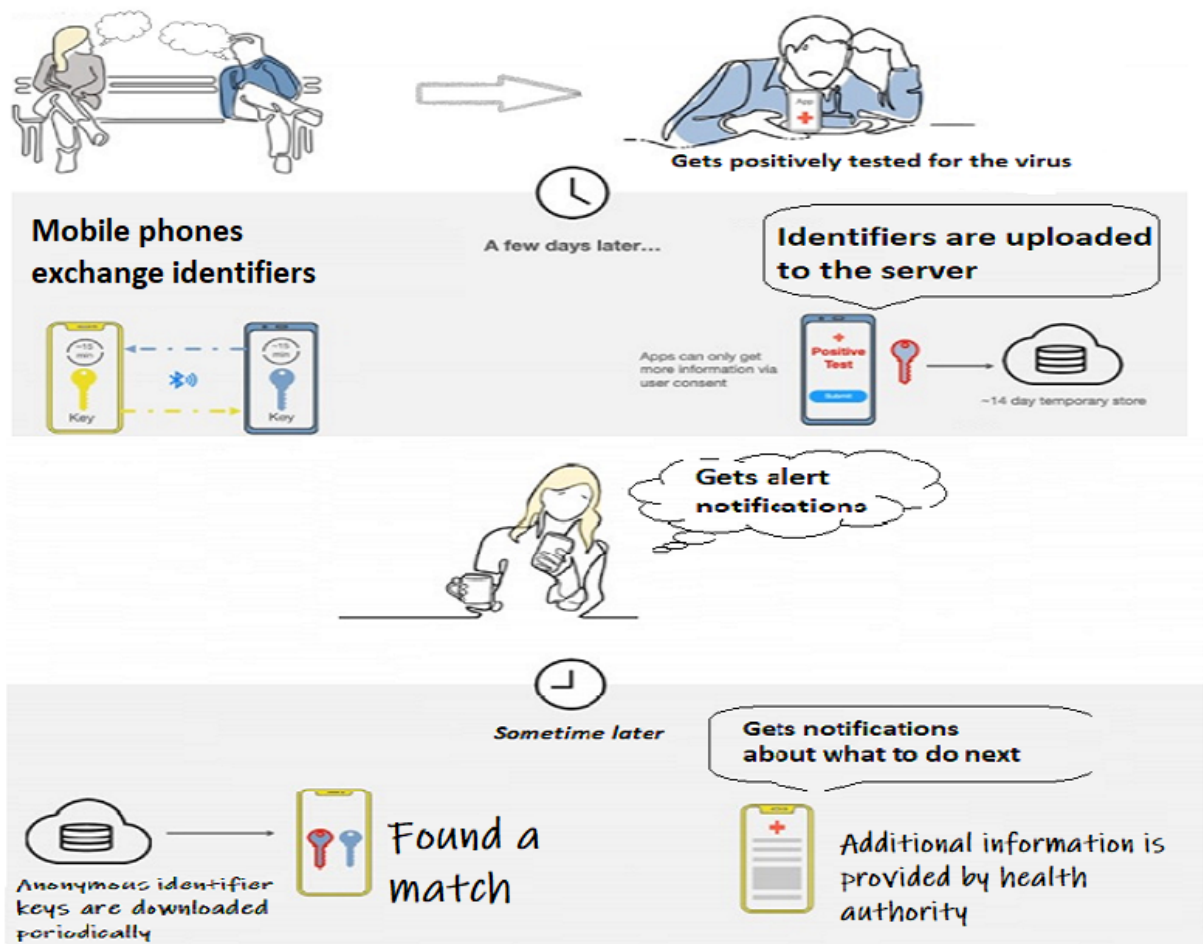


Users' privacy is ensured as follows:

- Users have full control over the Exposure Notification system. It is up to the user when to turn the app on or off;
- The geographical location of the user is not shared with anyone—not even with the government, Google, or Apple;
- To preserve user privacy, the random Bluetooth IDs change every 10-20 minutes;

- Notifications are carried out on the user's mobile phone only. To preserve the privacy of the user, if a person tests positive, their identity is not shared with other users or with Apple or Google;
- The objective of this system is to help the health care system fight COVID-19, which is why Google and Apple expect to deactivate this system when it is no longer needed.

**Table 1.** Existing contact tracing techniques for COVID-19.

| App name | Information-gathering technique | Type of data collected | Data storage and management |
|---|---|---|---|
| TraceTogether [10,21] | Bluetooth | Anonymous IDs, proximity of contacts | Data corresponding to anonymous IDs are generated from mobile devices and stored on a central server; the real identity of the users is hidden from other users but known to the server |
| COVID Trace [11] | Bluetooth | Geographic location, temporal details, self-assessed symptoms | Location and temporal details are stored on the users' personal device; data used only with permission from users |
| HowWeFeel [12] | Manual | Age, sex, postal code, self-assessed symptoms | Self-reported data are integrated and stored along with user IDs; storage and management via a central app server |
| NOVID [13] | Bluetooth and ultrasonic sound waves | Device details, OS version, time, language, Bluetooth, sonic signals specifications, proximity of contacts | Generates user ID and corresponding password; utilization and management by a central server |
| COVID Shield [14,22] | Bluetooth | Unique random ID, app logs, temporary keys for exposure | Data are not directly uploaded to a server; the generated user ID is uploaded to the central server in a secure manner |
| Zero (Safemap) [9] | Manual and automatic | Mobile number, email ID, GPS information, IP[a] address | Data are stored on users' individual server, managed by a central server |
| ShareTrace [16] | Bluetooth | Proximity of contacts, symptoms of users, diagnosis results | Data are stored on users' individual server, managed by a central server |
| Safe2 [17] | Bluetooth and GPS | Proximity of contacts, locations, self-assessed symptoms, lab test results | Generated random user IDs and data are stored on user devices; federated servers are used for handling data |

[a]IP: Internet Protocol.

There are several other apps and protocols that have been developed as a solution to handle the COVID-19 pandemic via contact tracing, such as NHS COVID-19, HealthCodeSystem, Pan-European Privacy-Preserving Proximity Tracing [23], Decentralized Privacy-Preserving Proximity Tracing apps, and others [24]. These apps work in a similar manner to the above-mentioned ones, with some minute differences. However, all of these apps face some issues related to security and privacy of the user's data. Most have their own central server that stores and manages data, which creates vulnerabilities such as data theft, data manipulation, data leakage, single point of failure, etc. Moreover, there is no verification mechanism provided by these apps that assures whether a particular individual is actually infected or not, as there are chances that a malicious user might put fake information on the app to create panic among other users, which raises issues related to the trust and reliability of the app. Furthermore, the personal data being collected by the apps still remain the main concern of app users. Apart from these apps, there are certain other apps that demand permission to access users' mobile phone contacts, media, location, files, phone ID, etc. Since these apps are constantly gathering user information for contact tracing, the privacy and security of the user's data continues to be the biggest concern and needs to be addressed [25].

## Issues and Challenges Posed by Digital Contact Tracing

Contact tracing accelerates the process of identifying the individuals who could have come into contact with infected cases. However, there are certain issues, challenges, and concerns associated with this mechanism that need to be addressed.

### Data Privacy and Compliance With the General Data Protection Regulation

One of the biggest concerns is that the people are hesitant in sharing their information, as they are not sure how their data will be used, who is going to be in charge of their data, and for how long. Contact tracing apps must gather data in compliance with the General Data Protection Regulation (GDPR), which states that any information related to a person that could make him or her directly or indirectly recognizable is said to be his or her personal information. Some important terms related to the GDPR are discussed below [26-29]:

- Data subject: refers to any individual who can possibly be identified directly or indirectly by means of an identifier such as a name, location data, personal ID number, or through some specific factors related to an individual's physical, genetic, economic, cultural, or social identity;
- Personal data: according to the GDPR, data become personal data when any kind of identification of the subject is possible through that data;
- Data controllers: refers to the main policy and decision makers who control the purpose behind the data collection and method of data processing;
- Data processors: work according to the instructions provided by data controllers to process the data.

In addition, there are various rights given by the GDPR to users regarding data and privacy, which include [30]:

- Right to access: the individual has the right to access personal data. Additionally, they have the right to know how their data are being used, processed, stored, and transferred to other organizations;
- Right to be informed: the individual has the right to be informed before data are gathered and processed;
- Right to data portability: the individual has the right to transfer their data from one service provider to another at any time;
- Right to be forgotten: the individual has the right to have their data deleted if they are no longer customers, or withdraw their consent of data usage;
- Right to object: the individual has the right to object the use or processing of their data;
- Right to restrict processing: the individual can ask to stop processing of certain kinds of data;
- Right to be notified: the individual has the right to be notified within 72 hours in case of a breach of their personal data;
- Right to rectification: the individual has the right to request the data controller to update, modify, or correct their data.

In terms of health care data, the GDPR requires more protection; in the context of digital contact tracing apps, the wireless technology and app managers receive the data along with user identifiers, making it their responsibility to keep the data anonymous. All the strategies for contact tracing available today employ one simple and common mechanism that tracks the movement of the carrier of the virus and does not protect the privacy of users. Therefore, the people who have been diagnosed with the disease are afraid of losing confidentiality since details about their location is broadcasted publicly. Although the actual identity of the individual is not disclosed, they can be mapped easily due to the limited number of carriers on the same route. Once identified, people can start making speculations about their personal lives, develop incorrect perceptions, or generate rumors.

## Data Quality and Transparency

COVID-19 is generating considerable amounts of data such as information on infected people, hospitalization and death rates, transmission details, etc. However, it is necessary to analyze these data so as to gain insights and make better and efficient decisions since the fight against the spread of COVID-19 is dependent on these decisions and the research being carried out. Therefore, it is essential that the data being analyzed are accurate and the quality of the data is assured. It must consist of metadata and context to avoid any misjudgment or discrepancy since lives are at stake. Moreover, the transparency of the data is another critical aspect; the origin and the generator of the data must be known to the researchers and/or users in order to increase trust and prevent the spread of false information or panic.

## Lack of Medical Understanding

COVID-19 is a highly communicable disease that is not only caused by physical contact but can also spread through air,

which makes tracking difficult since an individual can become infected without even coming into contact with an infected person [31]. Another issue that hurdles the implementation of contact tracing is that scientific data on COVID-19 are limited at present. For example, it affects each individual differently. Some patients are facing difficulties in breathing and dying while others experience no symptoms but are found to be COVID-19 positive. Additionally, asymptomatic and presymptomatic patients can also infect other people, which makes it difficult to track the source of contamination, as there is a chance that person A, who was asymptomatic and unaware of their infection, came into contact with person B, who became infected and experienced a severe outcome (eg, complications or death) because of the infection.

## Availability of Testing

The accuracy and efficiency of contact tracing is completely dependent on the amount of testing being performed. As of now, the government is only conducting tests on those who have a history of contact with an infected person or is experiencing symptoms related to COVID-19. Moreover, fees for the diagnostic tests are very high, making it difficult for the public in poor countries to get tested on their own.

## Trust Issues Between Governments and Citizens

Several contact tracing apps have been developed in the past few months by governments to combat the spread of the COVID-19 pandemic; however, many people criticize this as they believe this to be an attempt by their governments to regulate their lives and breach their fundamental rights of privacy. For example, Israeli legislation has allowed government officials to monitor the mobile phone data of people suspected of infection [32]. Similarly, a publicly available database has been created by the government of South Korea that consists of the personal information of infected individuals such as their job, travel routes, gender, age, etc [33]. Furthermore, GPS-based apps capture and broadcast the location details of users, which invades individuals' privacy [34]. Thus, it is necessary that the information being shared by users must be handled lawfully and with user consent.

## Technical Inefficiencies

Various digital contact tracing techniques based on GPS tracking and Bluetooth technology have been employed in the past. However, there are certain issues with these technologies relating to surveillance, malicious users, snoopers, etc. One such technique based on Bluetooth was proposed in 2018 [35], which was used to detect whether or not certain individuals came into proximity with infected people. This can be estimated on the basis of the strength of Bluetooth signals, which becomes low if there are any kinds of obstacles in between (like walls). Thus, considering the current scenario where there is an abundance of large buildings, the effectiveness of such a technical solution is highly affected. Moreover, the adoption rate of such systems is not very high, which is also a reason behind its limited effectiveness. Although GPS has been used to retrieve the location details of users in a digital contact tracing app, it is not very safe and efficient since it is very easy for malicious attackers to create false information on GPS networks

[36]. In addition, there is a chance that an app might provide users with false notions of safety. It is possible that a person might have come into contact with an infected person but failed to be notified by the app, thus the user remains unaware of the exposure and does not follow preventive measures.

## Probable Attacks on Digital Contact Tracing Apps

Figure 3 displays some of the most common ways digital contact tracing apps can be compromised by a malicious user/attacker. Some of these attacks are generic while others are specific to the wireless technology being used [37].

Figure 3. Probable attacks on contact tracing apps.
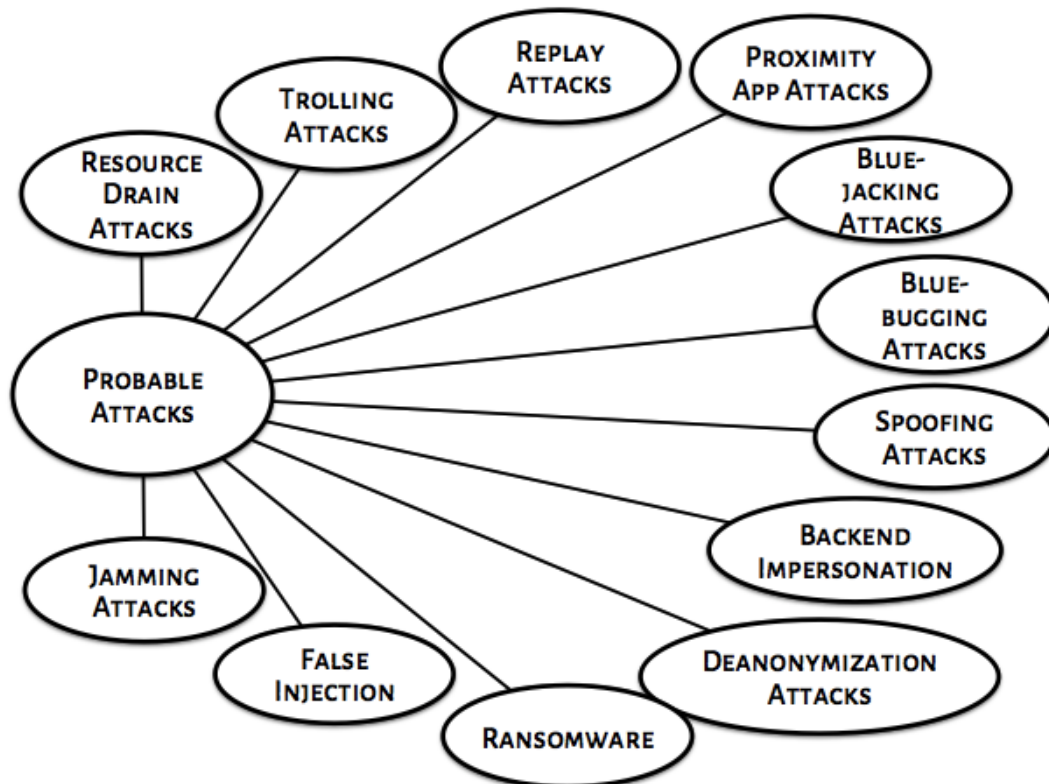


### Resource Drain Attacks

This is one of the most common types of attack that leads to denial of service. In such attacks, the attacker sends a massive number of trash messages, either valid or invalid, from their device to the server that forces the drainage of the other device's resources (eg, battery life) [38]. These attacks have no impact on the digital contact tracing app but might lead to poor performance of the mobile device.

### Trolling Attacks

In a trolling attack of a digital contact tracing app, the attacker spreads false information, such as lying about being exposed to the virus or being in close proximity with someone who has been diagnosed with the virus [34]. This misinformation creates panic among users, leading them to conduct diagnostic tests and inducing loss of trust in the system.

### Replay Attacks

In replay attacks, the attacker uses one or more mobile device(s) at different instances to broadcast the same message multiple times. In the context of digital contact tracing apps, replay attacks might spread anxiety among the public as a malicious user who is diagnosed positive with the virus pretends to be

more than one individual and sends alerts to even those users who were not near them [39].

### Proximity App Attacks

The proximity information of the user is supposed to be confidential, but an attacker can hack into a user's device and leak the information on users' daily life events to others. This attack creates fear among users and makes them reluctant in using the app and sharing their proximity details.

### Blue-Jacking Attacks

These attacks are specific to Bluetooth-based digital contact tracing apps, in which the attacker exploits Bluetooth and sends spam messages (containing the name and model of the sender device) to other devices that have their Bluetooth turned on. The receiving user has no idea about the sending user. The aim of such attacks is to make the receiving user behave in a certain manner and access the receiver's device.

### Blue-Bugging Attacks

Blue-bugging attacks are also specific to Bluetooth technology. In such attacks, the attacker tries to gain unauthorized access to a user's device and attempts to give commands. It is the most

dangerous attack, as the attacker might acquire total control over the user's device and misuse personal information.

### Spoofing Attacks

These attacks are usually common in digital contact tracing systems based on GPS technology, where the attacker exploits GPS signals near a mobile device in order to transmit incorrect location data that affects the overall performance of the tracing apps and results in incorrect judgments and decisions [40].

### Backend Impersonation Attacks

In such attacks, the attacker pretends to be someone else by replacing their own identity with the identity of another device. Then false information is broadcasted to other devices within the network.

### Deanonymization Attacks

Deanonymization attacks disrupt the privacy of users. The main focus of such attacks is the disclosure of the identity of the user to the entire network, which could lead to social stigma or discrimination if the user's diagnosis is found to be positive.

### Ransomware

In ransomware, the attacker develops fake contact tracing apps to gain access to users' mobile phones. This type of attack misuses the pandemic situation to gain information and data of users through their phones such as bank details, photos, contacts, etc [41]. Moreover, the attacker might lock the phone and can demand money to unlock it.

### False Injection Attacks

In false injection attacks, the attacker introduces deceptive information and tries to compromise information transactions among users, which lead to poor performance in tracing apps. If the attacker becomes capable of getting control over multiple devices within the network, the attacker may be able to create fraudulent reports.

### Jamming Attacks

Jamming attacks are specific to GPS-based contact tracing apps, in which signal interference (intentional or unintentional) prevents genuine signals from being received by users. The main objective of such attacks is to disrupt tracking via GPS signals, hence affecting the overall performance of the contact tracing system.

## Blockchain-Based Digital Contact Tracing for COVID-19

Digital contact tracing produces certain issues and challenges related to the privacy and security of users' personal information as discussed in the sections above. Since users' information is collected, traced, tallied, and broadcasted to the network, it becomes necessary to maintain confidentiality and prevent the divulgement of the user's identity. However, in certain apps, consent is asked of users, which provides some control to users over their information. However, verification of the authenticity of the information being shared is still lacking, as well as assurances to the users that only data relevant for tracing COVID-19 spread is being captured and nothing else. Blockchain technology can play a significant role in contact tracing as it supports the distributed peer-to-peer connectivity of the network nodes that bridges the gap between the users and app managers. The technical features offered by blockchain allow for sharing of information while preserving the privacy of users [42]. Table 2 provides a brief description of how the default properties provided by blockchain technology can be utilized for contact tracing apps.

**Table 2.** Features of blockchain technology and the corresponding application in the digital contact tracing process.

| Feature | Application |
| --- | --- |
| Decentralized network | The management of the data is user centric, which gives the power of data ownership to the users |
| Data security | The data within the blockchain is kept after applying encryption, which can only be decrypted by an authorized user |
| Data provenance | The information being entered in the blockchain is stamped with the digital signature of the source, which proves the legitimacy of the source as well as the data |
| Data availability | The data are distributed among all the nodes within the network, which makes them available all the time to every user |
| Data immutability | The information in the blockchain is immutable, which means once a detail is entered it can never be modified. This provides reliability and transparency to all users |
| Time stamping | The data within the blockchain network is time stamped, which eliminates the chances of discrepancies being present |

The privacy provided to the user and the performance of the app in terms of tracing contacts effectively should be the main focus of these apps. As most of the available apps lack the capability to protect user privacy, the main concern today is to provide effective tracing without compromising the users' identity and privacy [43,44]. Furthermore, the sharing of the data for decision-making processes in a secure manner is another challenge that is difficult in a centralized network due to the risks of data manipulation. This could be handled in blockchain-based systems as the network is totally distributed and the identities of the users are made anonymous in the beginning. Apart from preserving users' privacy, the app should also perform effectively in terms of tracing possible contacts, network coverage, infection prevention, etc. Currently, the apps that support decentralization are available for small geographical networks, which would not be beneficial for those who travel on a daily basis for work. Therefore, a blockchain-based decentralized network can provide accessibility and traceability at a global level and connect a larger number of users from different geographical locations without compromising their

privacy. Moreover, the information shared on blockchain can be collected from any means of technology (Bluetooth, GPS, etc) and can provide better and richer interactions.

The spreading of false information and rumors that might lead to panic among people is very dangerous and should be prevented. The reason behind the false information could be attributed to inaccurate details or a lack of transparency. It is necessary for health care organizations to have trustworthy authorities to verify and validate the information received to avoid inaccuracy or discrepancy in the data. Therefore, the blockchain network would be the best option, as it provides transparent tracing of contacts while maintaining privacy in a verifiable manner. Moreover, the privacy of the data should be ensured for the entire lifecycle of the data from its generation

to its disposal, which is usually 14 days in the case of COVID-19. Blockchain technology provides users with full control over the management of their data from start to finish and allows them to share and withdraw their data any time they want. Figure 4 depicts a typical blockchain-based contact tracing app framework that shows how the information regarding proximity and health status of the user are collected, analyzed, and utilized. Any activity within a blockchain-supported architecture takes place in the form of transactions, and these transactions are requested and represented in the form of blocks. These blocks of transactions are broadcasted to every node of the network and validated only if the nodes within the network verify them. Figure 5 illustrates the flow of possible transactions in a blockchain-based contact tracing app.

**Figure 4.** Blockchain-based contact tracing app framework.
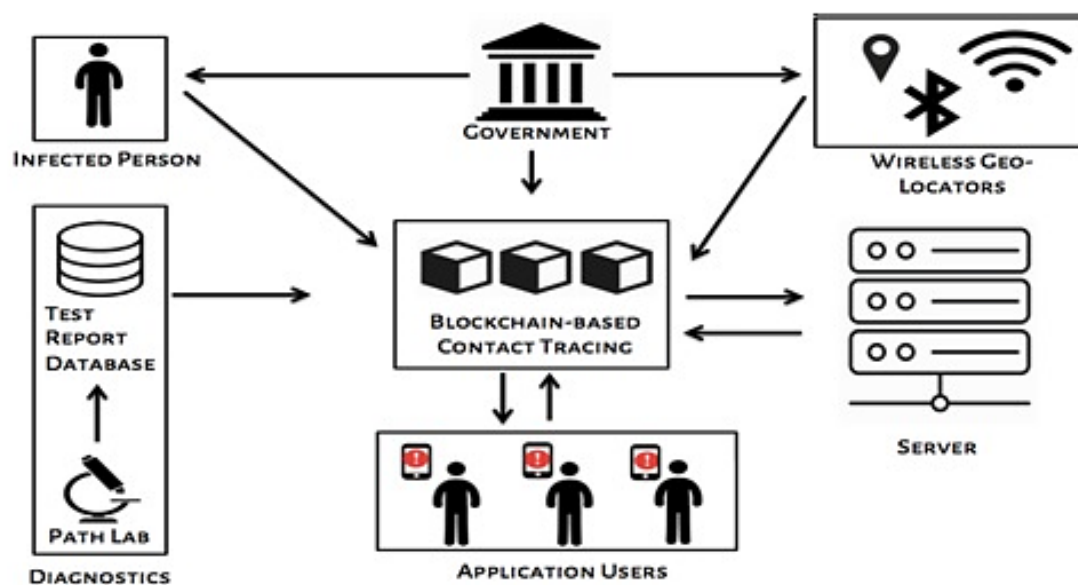
**Figure 5.** Data transaction flow in a blockchain-based contact tracing app.
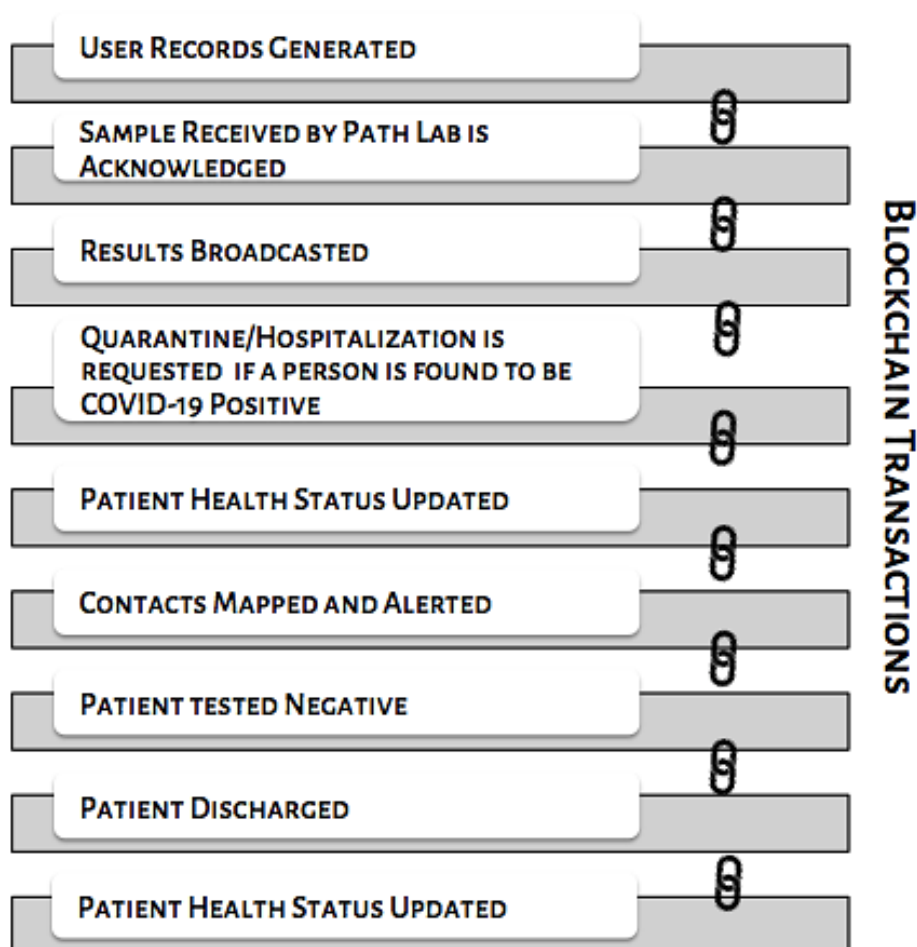


All users of the app, including the infected ones, will upload their data, that is, their unique ID and geographical information after applying encryption, to the blockchain network and perform matching on their own devices. When a user goes to the diagnostic center for a test, the results would be uploaded to the blockchain contact tracing app, which would then map the infected person to retrieve the details of their contacts with the help of the server. The servers basically read the data from the blockchain, and the network does the mapping using the geographical data available and provides the results back to the blockchain. The geographical data provided to the servers are collected using wireless technologies, such as Wi-Fi, Bluetooth, or GPS, that capture the location details of the users. The government monitors all data transactions occurring between users, path labs, server, geo-locators, and the blockchain.

## *The GDPR and Privacy Preservation in Contact Tracing Apps*

Privacy experts fear that the widespread use of contact tracing apps could disrupt the development of privacy-preserving regulations. In this context, some of the concerns addressed by the GDPR include [45]:

- Emphasis is on Bluetooth low-energy technology rather than the collection of location data from users through contact tracing apps;

- When it comes to using location data, preference should always be given to the processing of anonymized data rather than personal data;
- Once the user is diagnosed as infected with the virus, only the persons with whom the user has been in immediate proximity and within a specified time period should be informed;
- Information related to proximity between users of contact tracing apps can be obtained without locating them. The main motive is that the apps should not use the location data;
- Given the complexity of anonymization processes, transparency regarding the anonymization methodology is highly encouraged;
- If the app is based on a centralized type of architecture, then in accordance with the ethics of data minimization and data protection by design, the data processing capability of a centralized server should be restricted to minimum;
- Information kept on a central server should not allow the controller to recognize infected users or those people who were in proximity of an infected person. Additionally, it should not allow for the interpretation of contact patterns that could lead to the identification of contacts;
- To preserve privacy, it is adequate to exchange pseudonymous identifiers through Bluetooth low-energy technology across the users' device (mobile phone, tablet, watches, etc);

- These pseudonymous identifiers must be generated using various state-of-the-art cryptographic processes;
- The identifiers must be renewed at regular intervals of time to reduce the risk of physical tracking and linkage attacks;
- The app should not convey to the users information that allows them to infer the identity or the diagnosis of others. The central server must neither identify users, nor infer information about them.

## Discussion

The prevention of subsequent waves of COVID-19 transmission is the top priority in many countries around the world. Consequently, contact tracing is gaining much attention. Contact tracing is a highly effective cornerstone in the fight against infectious diseases. A similar version can be used to identify various other infectious diseases, but this cannot be generalized to COVID-19. Alternatively, COVID-19 tracing involves observation or surveillance to detect the early outbreak of this infectious disease, protect the population, and monitor testing resources properly [46].

Other issues include:

- Testing resources need to be enhanced. In addition, the various protocols and guidelines for state-level testing need to be evaluated;
- Research suggests that contact tracing has been most effective in places where the number of circulating cases is low;
- COVID-19 essentially has both a long period of infectiousness spread and asymptomatic spread. This makes manual contact tracing difficult because its success relies on an individual's ability to recall contacts while having been contagious, including prior to feeling ill;
- Research suggests that contact tracing and associated quarantine policies are not 100% effective, which burdens individuals in underresourced communities.

These challenges are real, but contact tracing needs to be done regardless because we have yet to come up with an effective treatment for COVID-19 [47]. Despite all these challenges, we need to adjust our strategies for handling this virus. Possible measure can be:

- Secure and confidential public health surveillance systems that are integrated with standardized reporting infrastructure;
- Monitoring of health system reports of patients feeling unwell and observe for concerning spikes, if any;
- Implementation of an enhanced surveillance model;
- Direct outreach to underresourced communities to monitor their health conditions to help expand the efforts against COVID-19;
- Prioritizing testing in high-risk groups to maximize the benefits of testing and consequent contact tracing.

Recently, a few entrepreneurs from India devised clever artificial intelligence (AI)–enabled technologies to keep workers safe from COVID-19. In a factory, for example, as soon as someone becomes infected, transmission is quick. As a result, it can be difficult to plan for production in such a work setting when it is unknown who is coming to work, who is well, and who is unwell. In order to curb this, one organization named BLP Group devised three AI solutions involving contact tracing to handle these challenges:

- Smart cameras monitored workers and detected whether or not they were wearing a mask via a model that uses computer vision to scan the workspace. This allowed for the detection of a person who was in close proximity to someone else as well as if they were wearing a mask. The cameras could even carry out regular temperature checks using thermal imaging. The issue is that employees must feel comfortable with these technologies. The right communication was critical to relay to workers the purpose of the tool (safety rather than policing). As soon as the workers were able to understand this purpose in terms of its benefits, they started responding to it well;
- Mobile phones were turned into an automatic alert system. When workers got too close to one another, their mobile phones started vibrating and ringing so as to create a warning system with no need for cameras;
- Wearable devices in the form of a wrist band buzzed when social distancing was breached.

Furthermore, if a worker became sick, data were collected from their wearables and contact tracing was quickly employed. This will, in turn, enable businesses to operate and isolate people who fall in the risk group. These technologies are also being used in some airports, hotels, and offices. This could be a way of building safer workplaces even when the pandemic is over. Such types of strategies can transform and disrupt the way businesses are carried out and could be used to formulate strategies for combating COVID-19.

## Conclusion and Future Directions

The novel coronavirus has affected more than 42 million people throughout the globe and has been termed a global emergency by the WHO. The virus is highly contagious and spreads through physical contact and socializing; therefore, in order to combat such diseases, several digital contact tracing mechanisms have been developed by companies, researchers, and governments. Contact tracing techniques have been practiced for many years now, but with time they have evolved and now utilize wireless technologies and mobile devices. Nevertheless, users are reluctant in sharing and broadcasting their personal information and proximity details, due to certainties about how their information will be used, by whom, and for how long. Therefore, a blockchain-based digital contact tracing technique that efficiently provides contact tracing without compromising users' privacy or confidentiality is required. Blockchain provides users with the total control over their data throughout the lifecycle and allows for withdrawal at any time. Moreover, the data stored are encrypted, time stamped, and immutable, making access by unauthorized persons impossible, which promotes transparency and eliminates discrepancy.

The success of any digital contact tracing app is dependent on trust and reliability and affects performance and wide-scale user adoption rate. In addition, the coverage of digital contact tracing apps should be high, as the apps can only run on smartphone

devices, which makes its adoption in some countries like India, Bangladesh, and Pakistan, etc, challenging, as the marketing and usage of smartphones in such countries is low. People who are uncomfortable with technology may not want to run these apps, leading to unawareness of their proximity and exposure. Bluetooth is the most widely used wireless technology in digital contact tracing apps and its proximity range is relatively high. However, it is not intelligent enough to detect any object in between 2 devices such as walls or doors. It might create problems in cities where homes are congested and generate incorrect results such as false positives that might lead to panic among people [15]. To handle this, Apple-Google API has included the transmission signal strength, but this area requires more research.

Another domain that needs work in these apps is the promotion of trust between citizens (users) and governments. As the proximity of users is being broadcasted at all times with these apps along with their personal data, users can become suspicious about being surveilled by their government. Therefore, the information being shared by users must be handled lawfully, with user consent, and in compliance with the GDPR to keep personal information private and identities anonymous [48,49]. Some countries like China and India have made their contact tracing app mandatory for every citizen, while others have made it voluntary whereby only interested users can download and use it. Since the disease is spreading at a faster rate, it becomes necessary to have trust among citizens to improve adoption rates. Moreover, these apps should be transparent and ask for consent to assure users in terms of security and privacy. Apart from the aforementioned concerns, other issues needing improvement that affects app performance and user acceptance rates include glitches in the apps, interoperability, and proper detection of devices [50].

## Acknowledgments

## Conflicts of Interest

None declared.

## References

1. Novel Coronavirus – China. World Health Organization. 2020 Jan 12. URL: https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/ [accessed 2020-02-01]
2. COVID-19 coronavirus pandemic. Worldometers. 2020. URL: https://www.worldometers.info/coronavirus/ [accessed 2020-10-25]
3. Chan JF, Yuan S, Kok K, To KK, Chu H, Yang J, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. The Lancet 2020 Feb;395(10223):514-523. [doi: 10.1016/s0140-6736(20)30154-9]
4. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. N Engl J Med 2020 Mar 26;382(13):1199-1207 [FREE Full text] [doi: 10.1056/NEJMoa2001316] [Medline: 31995857]
5. World Health Organization. Contact tracing in the context of COVID-19. World Health Organization. 2020 May 10. URL: https://www.who.int/publications/i/item/contact-tracing-in-the-context-of-covid-19 [accessed 2020-07-17]
6. Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, Abeler-Dörner L, et al. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. Science 2020 May 08;368(6491):eabb6936 [FREE Full text] [doi: 10.1126/science.abb6936] [Medline: 32234805]
7. Golinelli D, Boetto E, Carullo G, Nuzzolese AG, Landini MP, Fantini MP. Adoption of Digital Technologies in Health Care During the COVID-19 Pandemic: Systematic Review of Early Scientific Literature. J Med Internet Res 2020 Nov 06;22(11):e22280 [FREE Full text] [doi: 10.2196/22280] [Medline: 33079693]
8. Contact tracing. World Health Organization. 2017 May 9. URL: https://www.who.int/news-room/q-a-detail/contact-tracing [accessed 2021-01-13]
9. Privacy policy. Zero. 2020 May 19. URL: https://www.usezero.org/legal/privacy-policy [accessed 2021-01-13]
10. Bay J, Kek J, Tan A, Hau CS, Yongquan L, Tan J, et al. BlueTrace: A privacy-preserving protocol for community-driven contact tracing across borders. Government Technology Agency. 2020. URL: https://bluetrace.io/static/bluetrace_whitepaper-938063656596c104632def383eb33b3c.pdf [accessed 2021-01-13]
11. Privacy policy. COVID Trace. 2020 Mar 29. URL: https://covidtrace.com/privacy-policy/ [accessed 2021-01-13]
12. HowWeFeel. 2020. URL: https://howwefeel.org/ [accessed 2021-01-13]
13. Privacy policy. NOVID. 2020 Nov 30. URL: https://www.novid.org/privacy [accessed 2021-01-13]
14. Privacy policy. COVID Shield. 2020 May 13. URL: https://www.covidshield.app/privacy/ [accessed 2021-01-13]
15. Shorey R, Miller B. he Bluetooth technology: merits and limitations. 2000 Presented at: Proceedings of the IEEE International Conference on Personal Wireless Communications; 2000; Hyderabad, India p. 80-84. [doi: 10.1109/icpwc.2000.905777]
16. ShareTrace privacy policy. ShareTrace. URL: https://www.sharetrace.org/privacy [accessed 2021-01-13]

17.   Safe2 privacy policy. Safe2. 2020 May 17. URL: https://safe2.org/privacy-policy/ [accessed 2021-01-13]

18.   Aarogya Setu mobile app. Government of India. 2021. URL: https://www.mygov.in/aarogya-setu-app/ [accessed 2021-01-13]

19.   Singh HJL, Couch D, Yap K. Mobile Health Apps That Help With COVID-19 Management: Scoping Review. JMIR Nurs 2020;3(1):e20596 [FREE Full text] [doi: 10.2196/20596] [Medline: 32897271]

20.   Privacy-Preserving Contact Tracing. Apple Inc / Google LLC. 2020. URL: https://covid19.apple.com/contacttracing [accessed 2020-09-10]

21.   Huang Z, Guo H, Lee YM, Ho EC, Ang H, Chow A. Performance of Digital Contact Tracing Tools for COVID-19 Response in Singapore: Cross-Sectional Study. JMIR Mhealth Uhealth 2020 Oct 29;8(10):e23148 [FREE Full text] [doi: 10.2196/23148] [Medline: 33006944]

22.   Thomas R, Michaleff ZA, Greenwood H, Abukmail E, Glasziou P. Concerns and Misconceptions About the Australian Government's COVIDSafe App: Cross-Sectional Survey Study. JMIR Public Health Surveill 2020 Nov 04;6(4):e23081 [FREE Full text] [doi: 10.2196/23081] [Medline: 33048826]

23.   Pan-European Privacy-Preserving Proximity Tracing. PePP-PT eV iG. 2020. URL: https://www.science4covid19.pt/pan-european-privacy-preserving-proximity-tracing/ [accessed 2021-01-13]

24.   Troncoso C, Payer M, Hubaux JP, Salathé M, Larus J, Bugnion E, et al. Decentralized privacy-preserving proximity tracing. arXiv Preprint posted online May 25, 2020. [FREE Full text]

25.   Exposure Notification. Apple Inc / Google LLC. 2020. URL: https://www.google.com/covid19/exposurenotifications/ [accessed 2021-01-13]

26.   Ahmed J, Yildirim S, Nowostawski M, Abomhara M, Ramachandra R, Elezaj O. Towards Blockchain-Based GDPR-Compliant Online Social Networks: Challenges, Opportunities and Way Forward. In: Arai K, Kapoor S, Bhatia R, editors. Advances in Information and Communication. FICC 2020. Advances in Intelligent Systems and Computing, vol 1129. Cham: Springer; 2020:113-129.

27.   Article 4 GDPR - Definitions | General Data Protection Regulation (GDPR). Intersoft Consulting. URL: https://gdpr-info.eu/art-4-gdpr/ [accessed 2020-09-15]

28.   Li J, Guo X. Global Deployment Mappings and Challenges of Contact-tracing Apps for COVID-19. SSRN Journal Preprint posted online May 24, 2020. [FREE Full text] [doi: 10.2139/ssrn.3609516]

29.   Abeler J, Bäcker M, Buermeyer U, Zillessen H. COVID-19 Contact Tracing and Data Protection Can Go Together. JMIR Mhealth Uhealth 2020 Apr 20;8(4):e19359 [FREE Full text] [doi: 10.2196/19359] [Medline: 32294052]

30.   Ramirez N. 12 Facts about GDPR Compliance Regulations You Need to Know. Osano. 2020 Dec 9. URL: https://www.osano.com/articles/gdpr-compliance-regulations [accessed 2020-09-15]

31.   Kampf G, Todt D, Pfaender S, Steinmann E. Persistence of coronaviruses on inanimate surfaces and their inactivation with biocidal agents. J Hosp Infect 2020 Mar;104(3):246-251 [FREE Full text] [doi: 10.1016/j.jhin.2020.01.022] [Medline: 32035997]

32.   Tidy J. Coronavirus: Israel enables emergency spy powers. BBC News. 2020 Mar 17. URL: https://www.bbc.com/news/technology-51930681 [accessed 2020-04-23]

33.   Kim MJ, Denyer S. A travel log of the times in South Korea: Mapping the movements of coronavirus carriers. The Washington Post. 2020 Mar 17. URL: https://www.washingtonpost.com/world/asia_pacific/coronavirus-south-korea-tracking-apps/2020/03/13/2bed568e-5fac-11ea-ac50-18701e14e06d_story.html?utm_source=facebook&utm_medium=social&utm_campaign=wp_world [accessed 2020-04-23]

34.   Gaus A. Trolling attacks and the need for new approaches to privacy torts. USFL Rev 2012;47:353-376 [FREE Full text]

35.   Altuwaiyan T, Hadian M, Liang X. EPIC: Efficient Privacy-Preserving Contact Tracing for Infection Detection. 2018 Presented at: 2018 IEEE International Conference on Communications (ICC); May 20-24, 2018; Kansas City, MO p. 1-6 URL: https://ieeexplore.ieee.org/document/8422886 [doi: 10.1109/ICC.2018.8422886]

36.   Warner J, Johnston R. A simple demonstration that the global positioning system (GPS) is vulnerable to spoofing. J Sec Admin 2002;25(2):19-28.

37.   Dar AB, Lone AH, Zahoor S, Khan AA, Naaz R. Applicability of mobile contact tracing in fighting pandemic (COVID-19): Issues, challenges and solutions. Comput Sci Rev 2020 Nov;38:100307 [FREE Full text] [doi: 10.1016/j.cosrev.2020.100307] [Medline: 32989380]

38.   Brownfield M, Gupta Y, Davis N. Wireless sensor network denial of sleep attack. 2005 Presented at: Proceedings from the Sixth Annual IEEE SMC Information Assurance Workshop; June 15-17, 2005; West Point, NY p. 356-364 URL: https://ieeexplore.ieee.org/document/1495974 [doi: 10.1109/IAW.2005.1495974]

39.   Feng Z, Ning J, Broustis I, Pelechrinis K, Krishnamurthy SV, Faloutsos M. Coping with packet replay attacks in wireless network. 2011 Presented at: 8th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks; 2011; Salt Lake City, UT p. 368-376. [doi: 10.1109/SAHCN.2011.5984919]

40.   Warner J, Johnston R. GPS spoofing countermeasures. Homeland Security Journal 2003 [FREE Full text]

41.   Andronio N, Zanero S, Maggi F. HelDroid: Dissecting and Detecting Mobile Ransomware. In: Bos H, Monrose F, Blanc G, editors. Research in Attacks, Intrusions, and Defenses. RAID 2015. Lecture Notes in Computer Science, vol 9404. Cham: Springer; 2015:382-404.

42. Durneva P, Cousins K, Chen M. The Current State of Research, Challenges, and Future Research Directions of Blockchain Technology in Patient Care: Systematic Review. J Med Internet Res 2020 Jul 20;22(7):e18619 [FREE Full text] [doi: 10.2196/18619] [Medline: 32706668]

43. Xu H, Zhang L, Onireti O, Fang Y, Buchanan WJ, Imran MA. BeepTrace: Blockchain-enabled Privacy-preserving Contact Tracing for COVID-19 Pandemic and Beyond. IEEE Internet of Things Journal 2020:1-14 [FREE Full text] [doi: 10.1109/jiot.2020.3025953]

44. Khurshid A. Applying Blockchain Technology to Address the Crisis of Trust During the COVID-19 Pandemic. JMIR Med Inform 2020 Sep 22;8(9):e20477 [FREE Full text] [doi: 10.2196/20477] [Medline: 32903197]

45. COVID-19 contact tracing: data protection expectations on app development. Information Commissioner's Office. URL: https://ico.org.uk/media/for-organisations/documents/2617676/ico-contact-tracing-recommendations.pdf [accessed 2020-07-15]

46. Al Nsour M, Bashier H, Al Serouri A, Malik E, Khader Y, Saeed K, et al. The Role of the Global Health Development/Eastern Mediterranean Public Health Network and the Eastern Mediterranean Field Epidemiology Training Programs in Preparedness for COVID-19. JMIR Public Health Surveill 2020 Mar 27;6(1):e18503 [FREE Full text] [doi: 10.2196/18503] [Medline: 32217506]

47. Kucharski A, Klepac P, Conlan A, Kissler SM, Tang M, Fry H, et al. Effectiveness of isolation, testing, contact tracing and physical distancing on reducing transmission of SARS-CoV-2 in different setting. medRxiv Preprint posted online April 29, 2020. [FREE Full text] [doi: 10.1101/2020.04.23.20077024]

48. Chowdhury M, Ferdous M, Biswas K, Chowdhury N, Muthukkumarasamy V. COVID-19 Contact Tracing: Challenges and Future Directions. Preprints Preprint posted online June 19, 2020. [doi: 10.20944/preprints202006.0240.v2]

49. Cheng W, Hao C. Case-Initiated COVID-19 Contact Tracing Using Anonymous Notifications. JMIR Mhealth Uhealth 2020 Jun 22;8(6):e20369 [FREE Full text] [doi: 10.2196/20369] [Medline: 32501802]

50. Chan J, Foster D, Gollakota S, Horvitz E, Jaeger J, Kakade S. PACT: Privacy sensitive protocols and mechanisms for mobile contact tracing? arXiv Preprint posted online Apr 7, 2020 [FREE Full text]

## Abbreviations

**AI:** artificial intelligence
**API:** application programming interface
**AWS:** Amazon Web Services
**GDPR:** General Data Protection Regulation
**SSL:** Secure Socket Layer
**TCN:** Temporary Contact Numbers
**WHO:** World Health Organization

XSL•FO

**RenderX**

<u>Original Paper</u>

# Classification of the Disposition of Patients Hospitalized with COVID-19: Reading Discharge Summaries Using Natural Language Processing

Marta Fernandes[1,2,3*], PhD; Haoqi Sun[1,2,3*], PhD; Aayushee Jain[1,2], MSc; Haitham S Alabsi[1,3], MD; Laura N Brenner[3,4,5], MD; Elissa Ye[1,2], MSc; Wendong Ge[1,2,3], PhD; Sarah I Collens[1], BSc; Michael J Leone[1], MSc; Sudeshna Das[1,3], PhD; Gregory K Robbins[3,6*], MD, MPH; Shibani S Mukerji[1,3*], MD, PhD; M Brandon Westover[1,2,3,7*], MD, PhD

[1]Department of Neurology, Massachusetts General Hospital, Boston, MA, United States

[2]Clinical Data Animation Center, Boston, MA, United States

[3]Harvard Medical School, Boston, MA, United States

[4]Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, Boston, MA, United States

[5]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, United States

[6]Division of Infectious Diseases, Massachusetts General Hospital, Boston, MA, United States

[7]McCance Center for Brain Health, Massachusetts General Hospital, Boston, MA, United States

[*]these authors contributed equally

**Corresponding Author:**
Marta Fernandes, PhD
Department of Neurology
Massachusetts General Hospital
50 Staniford St
Boston, MA, 02114
United States
Phone: 1 6508621154
Email: <u>mbentofernandes@mgh.harvard.edu</u>

## *Abstract*

**Background:** Medical notes are a rich source of patient data; however, the nature of unstructured text has largely precluded the use of these data for large retrospective analyses. Transforming clinical text into structured data can enable large-scale research studies with electronic health records (EHR) data. Natural language processing (NLP) can be used for text information retrieval, reducing the need for labor-intensive chart review. Here we present an application of NLP to large-scale analysis of medical records at 2 large hospitals for patients hospitalized with COVID-19.

**Objective:** Our study goal was to develop an NLP pipeline to classify the discharge disposition (home, inpatient rehabilitation, skilled nursing inpatient facility [SNIF], and death) of patients hospitalized with COVID-19 based on hospital discharge summary notes.

**Methods:** Text mining and feature engineering were applied to unstructured text from hospital discharge summaries. The study included patients with COVID-19 discharged from 2 hospitals in the Boston, Massachusetts area (Massachusetts General Hospital and Brigham and Women's Hospital) between March 10, 2020, and June 30, 2020. The data were divided into a training set (70%) and hold-out test set (30%). Discharge summaries were represented as bags-of-words consisting of single words (unigrams), bigrams, and trigrams. The number of features was reduced during training by excluding n-grams that occurred in fewer than 10% of discharge summaries, and further reduced using least absolute shrinkage and selection operator (LASSO) regularization while training a multiclass logistic regression model. Model performance was evaluated using the hold-out test set.

**Results:** The study cohort included 1737 adult patients (median age 61 [SD 18] years; 55% men; 45% White and 16% Black; 14% nonsurvivors and 61% discharged home). The model selected 179 from a vocabulary of 1056 engineered features, consisting of combinations of unigrams, bigrams, and trigrams. The top features contributing most to the classification by the model (for each outcome) were the following: "appointments specialty," "home health," and "home care" (home); "intubate" and "ARDS" (inpatient rehabilitation); "service" (SNIF); "brief assessment" and "covid" (death). The model achieved a micro-average area

under the receiver operating characteristic curve value of 0.98 (95% CI 0.97-0.98) and average precision of 0.81 (95% CI 0.75-0.84) in the testing set for prediction of discharge disposition.

**Conclusions:** A supervised learning–based NLP approach is able to classify the discharge disposition of patients hospitalized with COVID-19. This approach has the potential to accelerate and increase the scale of research on patients' discharge disposition that is possible with EHR data.

## Introduction

The COVID-19 pandemic continues to present challenges for health care systems around the world [1-8], with over 32.7 million COVID-19 cases confirmed and 991,000 deaths worldwide as of September 27, 2020 [6]. The SARS-CoV-2 virus first appeared in Wuhan, China, in December 2019. The first case in the United States was confirmed January 20, 2020 [9], followed by rapid spread [2]. By the end of April, Massachusetts became the third hardest hit state, trailing New York and New Jersey [10].

To prepare for a possible second wave in Massachusetts, we set out to conduct a large-scale study of factors associated with outcomes in hospitalized patients at 2 large academic Boston hospitals. This effort required the significant task of reviewing medical records for over 1000 patients. For structured parts of the electronic health record (EHR), automated data extraction is straightforward. However, some essential information is exclusively or most reliably available only in semistructured or unstructured narrative medical notes, including patient-reported symptoms, examination findings, or social habits. Thus, developing automated approaches to EHR information extraction wherever possible is critical for more complete patient phenotyping.

Natural language processing (NLP) deals with automated analysis of unstructured text data. Recent advances in NLP machine learning have empowered computers to do several tasks such as machine translation, speech recognition, speech synthesis, semantic understanding, and text summarization [11,12]. NLP has the advantage of being much faster than human chart review of medical records [13-16].

Here we present an automated approach, using NLP, to extract a specific outcome from hospital discharge summaries: discharge destination or "disposition" (ie, anticipated location or status following discharge). Dispositions of interest included home, inpatient rehabilitation center, skilled nursing inpatient facility (SNIF), and death. Discharge disposition of patients with COVID-19 from health care facilities is important due to the high risk of transmission of the disease within nursing homes

and hospitals when patients are discharged to locations other than home, and also because it represents an important measure closely related to functional outcome and level of disability following hospitalization, as well as overall costs of care. Furthermore, this information has the potential to aid health care facilities in resource planning to better prepare for the incoming flow of patients. Although our model is tailored for discharge disposition, the approach we developed is generalizable to other outcomes available in discharge summaries.

## Methods

### Study Overview

Data were extracted from the hospital electronic medical record under a research protocol approved for a waiver of informed consent by the Partners Healthcare Institutional Review Board. Clinical data were retrospectively analyzed for all adult patients who tested positive for SARS-CoV-2 infection between March 10 and June 30, 2020. A total of 1737 patients admitted to 2 major Boston hospitals, 1232 from Massachusetts General Hospital (MGH) and 505 from Brigham and Women's Hospital (BWH), were included. Only patients with a physician discharge summary and available known ground-truth discharge disposition were included.

### Data Collection and Processing

#### Overview

Data consisted of discharge summaries, which are unstructured free-text notes written by physicians, and a ground-truth record of discharge disposition, used to assess the accuracy of the NLP results. The methodology for note preprocessing is shown in Figure 1. The upper part of the figure provides an overview of the text extraction for each field on the list of extraction fields depicted in Table 1. The lower part of the figure shows the methodology steps where the text extracted from all the fields is processed for modeling. The data were randomly stratified into train and test sets for modeling, which we address in the Model Development section.

**Figure 1.** Methodology steps for discharge summary notes preprocessing and modeling. The list of extraction field is depicted in Table 1.
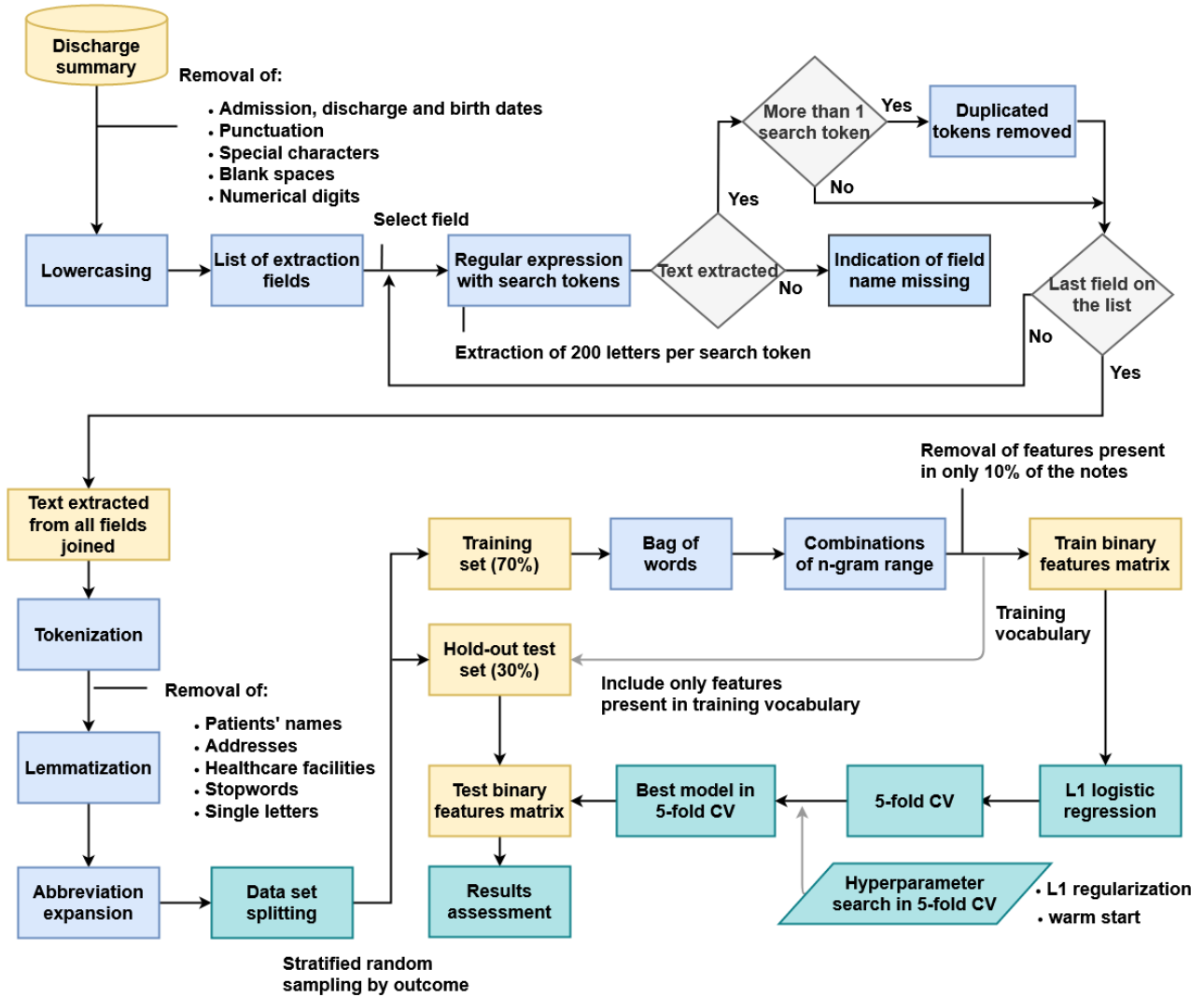
**Table 1.** Information captured from discharge summaries, grouped in fields, and respective search tokens used in the regular expression.

| Field | Search token |
| --- | --- |
| Discharge disposition | "discharge," "discharged," "dispo," "skilled nursing," "snf" |
| Diagnosis | "diagnosis," "diagnoses," "problem," "reason for admission," "chief complaint" |
| Surgeries | "surgeries this admission" |
| Treatments | "treatments" |
| Tests | "tests" |
| Allergies | "allergies," "allergic" |
| Diet | "diet," "nutrition" |
| Medical history | "history" |
| Hospital course | "hospital course" |
| Laboratory results | "labs" |
| Activity | "activity," "activities" |
| Physical exam | "discharge exam," "physical exam" |
| Physical therapy | "physical therapy" |
| Occupational therapy | "occupational therapy" |
| Discharge instructions | "instructions" |
| Follow-up care | "follow up" |
| Discharge plan | "discharge plan" |
| Additional orders | "additional orders" |
| Code status | "code status" |

### Document Preprocessing

Admission, discharge, and birth dates were removed from the discharge summaries, as well as punctuation, special characters, blank spaces, and numerical digits. Notes were then subjected to lowercasing, tokenization, and correction using lemmatization, a procedure for obtaining the root form of the word, using vocabulary (dictionary importance of words) and morphological (word structure and grammar relations) analysis. *WordNetLemmatizer* from NLTK library in Python (Version 3.7; Python Software Foundation) was used with a part-of-speech (POS) tag specified as a verb. Patients' names, addresses, health care facilities, and hospital unit names were removed, as well as single letters. Abbreviation expansion and spelling corrections were performed for a small list of frequently used clinical words (Table S1 in Multimedia Appendix 1). A list of commonly used and less informative stopwords was also removed from the notes (Table S2 in Multimedia Appendix 1).

### Processing of Specific Discharge Summary Fields

Discharge summaries at MGH and BWH are semistructured, with a series of named fields containing specific types of mostly free-text information (Table 1). We present an example of discharge summary notes with protected health information removed (Table S3 in Multimedia Appendix 1). Text fields were identified based on information extracted from the notes using regular expressions with search tokens (Table 1). The function "str.extractall" from Python was used to extract a length

of 200 letters of text onwards from all instances where the search token appeared.

Some notes contained a "discharge disposition" field used to list the discharge disposition. We deleted this field to avoid an overly "easy" solution, because this field is not universally available, and because we wished to assess how well the approach is able to perform when structured data is unavailable. In a field where more than one extraction was performed (ie, with more than one search token), the corresponding results were joined, and duplicated words were removed. To illustrate with an example, for the "Diet" field, using the regular expressions with search tokens "diet" and "nutrition," 200 letters were captured for each search token, for a total of 400 letters. Since there might be repeated information in the discharge summary regarding diet and nutrition recommendations, duplicated words were removed from the captured text. Where no data was captured with the search tokens, an indication of missingness was set with the name of the field and the suffix "_missing."

The texts extracted from all fields (depicted in Table 1) were joined to create a reduced version of the discharge summary, which was then subjected to tokenization, lemmatization, and abbreviation expansion, as described in the Document Preprocessing subsection. The vocabulary used for modeling was created based on these reduced versions of the discharge summaries contained in the training set. Documents were represented as a binary bag-of-words (BoW; ie, an ordered series of binary vectors indicating whether a given n-gram [word or

sequence of 2 or 3 words] is present in the document, disregarding grammar and word order). The function *CountVectorizer* was used with its default parameters from Python, except for the n-gram range, which was set as unigrams (1 word), bigrams (2 consecutive words), and trigrams (3 consecutive words). As a first step to reduce dimensionality, only features present in at least 10% of the reduced version of the discharge summary notes were considered. Multiclass logistic regression with the least absolute shrinkage and selection operator (LASSO) [17] was used to further sparsify the model.

## Outcome Measure

The multiclass outcome measure was discharge disposition, composed of the classes: home, inpatient rehabilitation, SNIF, and death. "Home" included "home or self-care," "home-health care services," and patients who "left against medical advice." SNIF included "Skilled Nursing Facility" and "Custodial Care Facility."

## Model Development

The training algorithm used the one-vs-rest scheme for multiclassification, where a binary problem was fitted for each class and the class weight was balanced. Logistic regression [18] with LASSO regularization was used as the classification model. The model estimator ☒ is depicted in equation 1 and the LASSO regularization objective can be written as in equation 2. ☒ corresponds to the design input matrix and ☒ corresponds to the vector of observations, where *n* is the number of observations, in this case the number of discharge summaries or number of patients, and *p* the number of features in ☒. The vector of regression coefficients is given by ☒ corresponds to the L1 norm of this coefficients vector, and λ is the regularization parameter that controls the amount of shrinkage. The regularization adds a penalty on the weights to prevent overfitting [19]. The inverse of the regularization strength *C* was varied for the values {0.005, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5}.



Stratified random sampling was used to split the data set into a training set (70%) and a hold-out test set (30%). A randomized search was used for hyperparameter tuning during training with 100 iterations of 5-fold cross-validation. The solver was set to "liblinear" and the "warm start" hyperparameter was varied between true/false, where "true" corresponded to reusing the solution of the previous call to fit as initialization, and "false" corresponded to erasing the previous solution.

## Performance Measures

The $R^2$ coefficient of determination score was used in cross-validation scoring to select the best model configuration in the training data. The one standard error rule was used to select the regularization parameter. The simplest model, whose $R^2$ mean score fell within 1 standard deviation of the maximum $R^2$, was selected.

To measure model performance on test data, the area under the receiver operating characteristic curve (AUROC) was calculated. The ROC curve is a function of recall (sensitivity) versus the false positive rate (FPR; ie, 1–specificity; Table S1 in Multimedia Appendix 1). The pair ($Recall_k$, $FPR_k$) is called an operating point for this curve, where k is a threshold that is varied to generate the ROC curve. The equations for these metrics are presented in Table S4 in Multimedia Appendix 1.

The area under the precision-recall curve (AUPRC), which is an important measure in the presence of class imbalance, was also calculated. The pair ($Recall_k$, $Precision_k$) is referred to as an operating point for this curve. Average precision (AP; Table S3 in Multimedia Appendix 1) summarizes this plot as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight.

The $F_1$-score (Table S4 in Multimedia Appendix 1) was also assessed as another performance metric commonly reported for data sets with imbalanced numbers across classes [20].

In total, 100 iterations of bootstrap random sampling with replacement were performed to calculate 95% CIs for performance metrics.

## *Results*

### Summary of Patient Population

From 1917 patients' medical records, 1752 had a physician discharge summary and a discharge disposition within the categories of home, inpatient rehabilitation, SNIF, and death. Only adults (aged ≥18 years) were included in the analysis, leaving a study cohort of 1737 patients. The cohort was split into train and test sets using stratified random sampling according to outcome. Age in the train and test sets was balanced, with a median of 62 and 60 years old, respectively (Table 2). The majority of patients were White (n=774; median 44.6%) and Black or African American (n=285; median 16.4%). Most were discharged home (n=1052; 60.6%). Among all patients with COVID-19 in this sample, there were 243 (14.0%) nonsurvivors.

**Table 2.** Baseline characteristics of the study patient population stratified by train and test sets.

| Characteristic | Train set (n=1215) | Test set (n=522) | Total (N=1737) |
|---|---|---|---|
| Age (years), median (SD) | 62.0 (18.2) | 60.0 (18.2) | 61.0 (18.2) |
| **Gender, n (%)** | | | |
| Female | 545 (44.9) | 244 (46.7) | 789 (45.4) |
| Male | 670 (55.1) | 278 (53.3) | 948 (54.6) |
| **Race, n (%)** | | | |
| White | 533 (43.9) | 241 (46.2) | 774 (44.6) |
| Hispanic or Latino | 52 (4.2) | 19 (3.6) | 71 (4.1) |
| Black or African American | 204 (16.8) | 81 (15.5) | 285 (16.4) |
| Asian | 46 (3.8) | 21 (4.0) | 67 (3.9) |
| American Indian or Alaska Native | 31 (2.5) | 13 (2.5) | 44 (2.5) |
| Native Hawaiian or other Pacific Islander | 2 (0.2) | 1 (0.2) | 3 (0.2) |
| Unknown[a] | 347 (28.6) | 146 (28.0) | 493 (28.3) |
| **Institution, n (%)** | | | |
| Massachusetts General Hospital | 881 (72.5) | 351 (67.2) | 1232 (70.9) |
| Brigham and Women's Hospital | 334 (27.5) | 171 (32.8) | 505 (29.1) |
| **Discharge disposition, n (%)** | | | |
| Home | 736 (60.6) | 316 (60.5) | 1052 (60.6) |
| Inpatient rehabilitation | 102 (8.4) | 44 (8.4) | 146 (8.4) |
| Skilled nursing inpatient facility | 207 (17.0) | 89 (17.1) | 296 (17.0) |
| Death | 170 (14.0) | 73 (14.0) | 243 (14.0) |

[a]Unknown includes "other," "declined," or "unavailable."

The preprocessed data set for modeling was created based on the notes extracted in all fields except the "discharge disposition" and "code status" fields, as described in the Processing of Specific Discharge Summary Fields subsection. Before dimensionality reduction, where features present in at least 10% of the reduced version of the discharge summary notes were considered, there were a total of 15,182 tokens (unigrams). After applying this dimensionality reduction step, we were left with 477 tokens. With this set of tokens, 3497 combinations of n-grams were generated, leading to a total of 1056 features with duplicates removed. Thus, the total number of candidate features in the training vocabulary was 1056, including 460 unigrams, 329 bigrams, and 267 trigrams.

## Modeling Results

The best model configuration parameters and performance results in the hold-out test set are presented in Table 3 with 95% CIs. The corresponding confusion matrices normalized by precision and recall are presented in Figure 2. The performance discriminated by discharge outcome is presented in Table 4.

Higher performance was obtained for the outcomes of home discharge and death compared to inpatient rehabilitation and SNIF discharge outcomes. The model presented higher recall (0.95) and precision (1.0) for the death outcome. Home disposition also presented high performance for these metrics. For this model, 2 deceased patients were classified as discharged home. In experiments, for models where we included the discharge disposition field, extracted from the discharge summary, all deceased patients were correctly classified. The inpatient rehabilitation outcome presented the lowest recall (0.61) and 12 patients with this outcome were incorrectly classified by the model as discharged to SNIF. The outcome of disposition to SNIF presented the lowest precision (0.68) overall and 20 patients discharged home were incorrectly predicted as discharged to SNIF. Compared to the initial set of features in the training vocabulary, the final model contained approximately 83% fewer features, with a total of 179 features. The relative importance of the top 30 model features is presented in Figure 3, where the importance for each feature consisted of the sum of the absolute coefficients' values across the outcomes.

**Figure 2.** Confusion matrices for the best model evaluated in the hold-out test set normalized (A) by recall and (B) by precision. SNIF: skilled nursing inpatient facility.



(a)  (b)
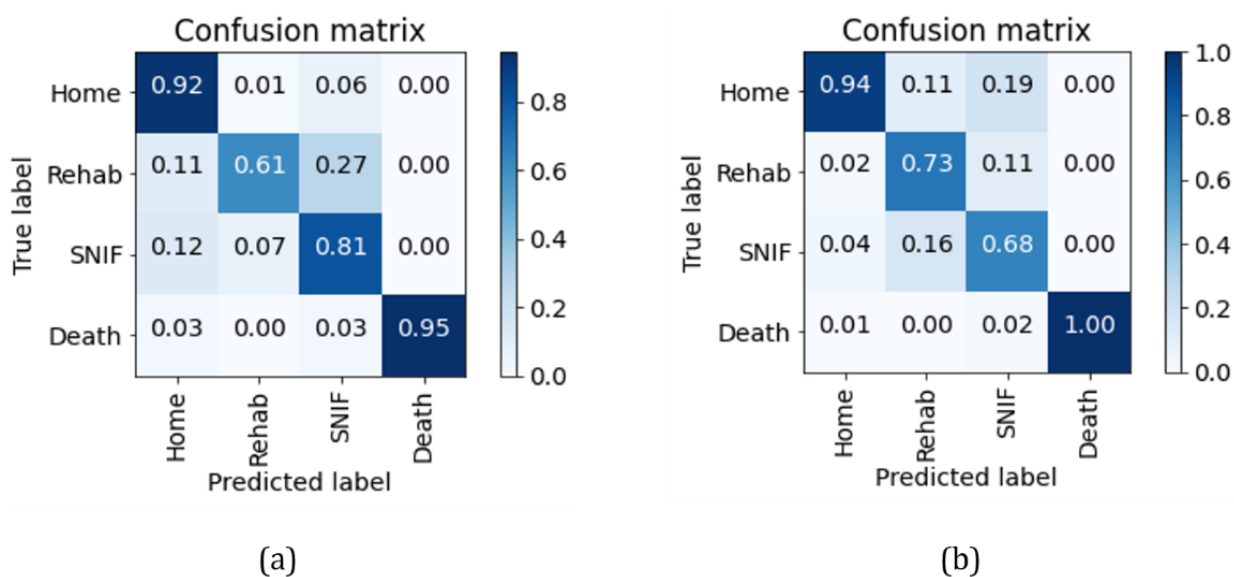
**Table 3.** Model performance in the hold-out test set and configuration parameters.

| Area under the receiver operating characteristic curve[a] | Accuracy[a] | Recall[a] | F$_1$ score[a] | Average precision[a] | Precision[a] | Parameters |
|---|---|---|---|---|---|---|
| 0.98 (0.97-0.98) | 0.88 (0.85-0.90) | 0.88 (0.85-0.90) | 0.88 (0.85-0.90) | 0.81 (0.75-0.84) | 0.88 (0.85-0.90) | Number of features (unigrams, bigrams, trigrams): 179 (95, 52, 32); C=0.09; warm start: true |

[a]The 95% CIs of bootstrapping results are in parentheses.

**Figure 3.** Relative importance of top 30 features obtained with the model coefficients estimates for (A) the sum of the absolute coefficients values and (B) the coefficients values discriminated by outcome. Coef: coefficient.
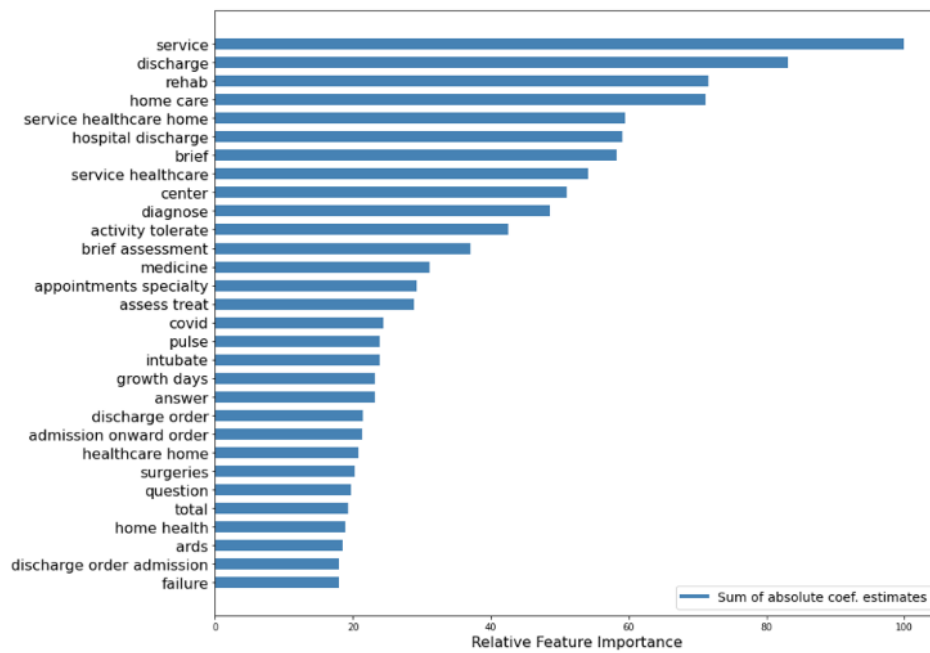


(a)
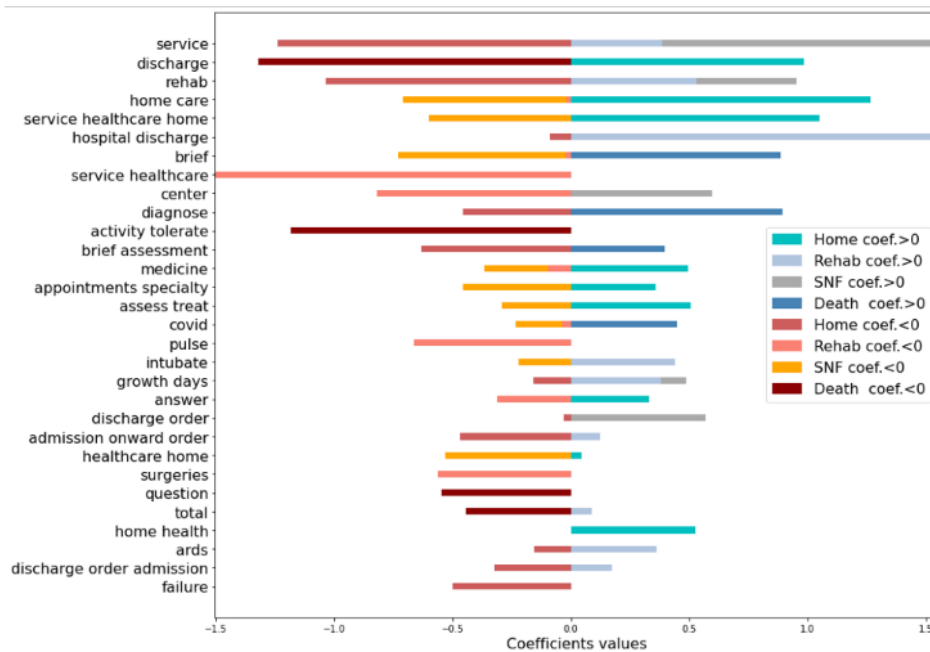


(b)

**Table 4.** Model performance in the hold-out test set by discharge outcome.

| Outcome | Area under the receiver operating characteristic curve[a] | Accuracy[a] | Recall[a] | $F_1$ score[a] | Average precision[a] | Precision[a] |
|---|---|---|---|---|---|---|
| Home | 0.97 (0.95-0.98) | 0.92 (0.89-0.94) | 0.92 (0.89-0.95) | 0.93 (0.91-0.95) | 0.92 (0.88-0.94) | 0.94 (0.91-0.97) |
| Rehab | 0.95 (0.91-0.98) | 0.95 (0.93-0.97) | 0.61 (0.53-0.76) | 0.67 (0.53-0.78) | 0.48 (0.32-0.64) | 0.73 (0.58-0.86) |
| Skilled nursing inpatient facility | 0.93 (0.88-0.96) | 0.90 (0.87-0.92) | 0.81 (0.72-0.88) | 0.74 (0.64-0.79) | 0.58 (0.46-0.66) | 0.68 (0.58-0.75) |
| Death | 1.00 (1.00-1.00) | 0.99 (0.99-1.00) | 0.95 (0.90-0.98) | 0.97 (0.95-0.99) | 0.95 (0.91-0.98) | 1.00 (1.00-1.00) |

[a]The 95% CIs of bootstrapping results are in parentheses.

"Service" was the feature assigned the highest importance for classification of the discharge outcomes. For inpatient rehabilitation and SNIF dispositions, the coefficient values for this feature are positive, which indicates that this term will most likely appear in the preprocessed notes for both outcomes. "Home care," "healthcare home," and "home health" were assigned a positive coefficient value for home disposition. "Service healthcare home" was also assigned high importance for this outcome, suggesting that this feature is related to patients discharged home with home health care services provided. "Medicine" and "appointments specialty" were also important for this outcome. "Rehab" had positive coefficients for both inpatient rehabilitation and SNIF dispositions. "Intubate" and "ARDS" (acute respiratory distress syndrome) are important features for inpatient rehabilitation disposition. For death, "discharge" and "activity tolerate" presented negative coefficient values, indicating that these features are unlikely to appear in discharge summaries of deceased patients. "Brief assessment" and "brief" are assigned high coefficient values for this outcome. "Covid" was assigned a positive coefficient value for predicting death, while the term was given negative values for inpatient rehabilitation and SNIF.

Training performance is depicted in Figure S1 in Multimedia Appendix 2, with the curve corresponding to the $R^2$ scores for the different values of the inversed regularization strength. The top 15 features and their relative importance obtained with LASSO regularization are presented for each outcome (Figure S2 in Multimedia Appendix 2). Blue bars correspond to features with positive coefficient values and red bars to features with negative coefficient values. The areas under the ROC and Precision-Recall curves for the best model are also presented (Figure S3 in Multimedia Appendix 2). We also assessed how the model performance and the features selected as the most important in the train set varied with the dimension of the train set (Figure S4 in Multimedia Appendix 2). The hold-out test set for model evaluation was fixed and the train set dimension was varied from 10% to 100% of the original train set, with 1215 patients. We observed that the best performance was achieved with a higher number of patients in the train set (ie, the original train set of 100%). However, with 50% versus 100% of the original train set, the model achieved good performance for 1018 (versus 1056) vocabulary features (AUROC 0.97 versus 0.98 and AP 0.79 versus 0.81, respectively). We assessed the common features between each train set and the original train set (Figure S5 in Multimedia Appendix 2). Among the top 30

features, there were 10 common features between the 50% and the original train sets. A higher number of common features was found for the train set with 90% of the original train set, with a total of 17 common features. Finally, we observed that more than half of the features in the top 30 from the original train set were selected as top 30 in at least two train sets (Figure S6 in Multimedia Appendix 2).

## Discussion

### Principal Findings

In this study, a machine learning–based NLP pipeline was developed to classify the discharge disposition of adult patients hospitalized with COVID-19. The model achieved near-perfect identification of patients with outcomes of home disposition or death. For the intermediate outcomes of inpatient rehabilitation or SNIF, performance was imperfect but also acceptable. Due to this classification task being relatively easy, more complex and time-consuming modeling approaches, such as recurrent neural networks or bidirectional encoder representations from transformers were not considered. We acknowledge that for harder tasks, these approaches can improve performance. The final method is automated, thus enabling large-scale rapid processing of thousands of discharge summaries, a task that is infeasible when relying on manual chart review.

### Limitations

The present analysis was limited to a cohort of patients with COVID-19, who may have specific medical symptoms related to the disease. Therefore, as future work, it is proposed to extend the model to other cohorts. Further, although results spanned 2 hospitals, they are located in the same geographic region (Boston, Massachusetts). Thus, our cohort may not be representative of other US and non-US populations. Moreover, decision making for discharge disposition may vary for different hospitals, according to the number of SNIFs or rehabilitation centers in the geographic area, which may affect the generalizability of the model. The models were developed with textual information from discharge summaries, while the addition of other clinical features (eg, physical or occupational therapy reports, social work or case manager notes) was not considered, which is a limitation of the study and can be pursued in future work.

XSL•FO

RenderX

## Comparison With Prior Work

Extraction of information from clinical narratives is a growing application of NLP in health care. NLP has been used to extract information from hospital discharge notes about medical conditions such as postsurgical sepsis [21], pneumonia [22], or other potential medical problems [23], as well as to identify critical illness [24,25], detect adverse events [26], predict risk of rehospitalization [27], extract medication information [28], and risk stratify patients [29]. To the best of our knowledge, ours is the first work on classifying hospital discharge disposition based on discharge summary notes using machine learning and NLP.

## Conclusions

This study shows that a supervised learning–based NLP approach can be used to accurately classify the discharge disposition of hospitalized patients with COVID-19 in an automated fashion. This model, and the NLP approach used to develop it, have the potential to accelerate and increase the scale of research that is possible with EHR data.

## Conflicts of Interest

None declared.

Multimedia Appendix 1
Methodology.
[DOC File , 555 KB - medinform_v9i2e25457_app1.doc ]

Multimedia Appendix 2
Results.
[DOC File , 399 KB - medinform_v9i2e25457_app2.doc ]

## References

1. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, the Northwell COVID-19 Research Consortium, et al. Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. JAMA 2020 May 26;323(20):2052-2059 [FREE Full text] [doi: 10.1001/jama.2020.6775] [Medline: 32320003]
2. Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, et al. First Case of 2019 Novel Coronavirus in the United States. N Engl J Med 2020 Mar 05;382(10):929-936. [doi: 10.1056/nejmoa2001191]
3. Fauci A, Lane H, Redfield R. Covid-19 - Navigating the Uncharted. N Engl J Med 2020 Mar 26;382(13):1268-1269 [FREE Full text] [doi: 10.1056/NEJMe2002387] [Medline: 32109011]
4. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. The Lancet 2020 Mar;395(10229):1054-1062. [doi: 10.1016/s0140-6736(20)30566-3]
5. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. JAMA 2020 Apr 07;323(13):1239-1242. [doi: 10.1001/jama.2020.2648] [Medline: 32091533]
6. CDC COVID-19 Response Team. Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) - United States, February 12-March 16, 2020. MMWR Morb Mortal Wkly Rep 2020 Mar 27;69(12):343-346 [FREE Full text] [doi: 10.15585/mmwr.mm6912e2] [Medline: 32214079]
7. He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. Nat Med 2020 May 15;26(5):672-675. [doi: 10.1038/s41591-020-0869-5] [Medline: 32296168]
8. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. Ann Intern Med 2020 May 05;172(9):577-582 [FREE Full text] [doi: 10.7326/M20-0504] [Medline: 32150748]
9. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. JAMA 2020 Mar 17;323(11):1061-1069 [FREE Full text] [doi: 10.1001/jama.2020.1585] [Medline: 32031570]
10. The Covid-19 Tracker. STAT. URL: https://www.statnews.com/feature/coronavirus/covid-19-tracker/ [accessed 2021-01-21]

11. Nallapati R, Zhou B, Nogueira dos santos C, Gulcehre C, Xiang B. Abstractive text summarization using sequence-to-sequence RNNs and beyond. ArXiv Preprint posted online on February 19, 2016. [FREE Full text] [doi: 10.18653/v1/k16-1028]

12. Hirschberg J, Manning CD. Advances in natural language processing. Science 2015 Jul 17;349(6245):261-266. [doi: 10.1126/science.aaa8685] [Medline: 26185244]

13. Wilbur WJ, Rzhetsky A, Shatkay H. New directions in biomedical text annotation: definitions, guidelines and corpus construction. BMC Bioinformatics 2006 Jul 25;7(1):356 [FREE Full text] [doi: 10.1186/1471-2105-7-356] [Medline: 16867190]

14. Buchan NS, Rajpal DK, Webster Y, Alatorre C, Gudivada RC, Zheng C, et al. The role of translational bioinformatics in drug discovery. Drug Discov Today 2011 May;16(9-10):426-434. [doi: 10.1016/j.drudis.2011.03.002] [Medline: 21402166]

15. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc 2011 Sep 01;18(5):544-551 [FREE Full text] [doi: 10.1136/amiajnl-2011-000464] [Medline: 21846786]

16. Uzuner Ö, South B, Shen S, DuVall S. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011;18(5):552-556 [FREE Full text] [doi: 10.1136/amiajnl-2011-000203] [Medline: 21685143]

17. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological) 2018 Dec 05;58(1):267-288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]

18. Cramer JS. The Origins of Logistic Regression. SSRN Journal 2003 Jan 25. [doi: 10.2139/ssrn.360300]

19. Bühlmann P, van de Geer S. Statistics for High-Dimensional Data: Methods, Theory and Applications. Heidelberg, Germany: Springer; 2011.

20. Azari A, Janeja V, Levin S. Imbalanced learning to predict long stay Emergency Department patients. 2015 Presented at: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2015; Washington, DC p. 807-814. [doi: 10.1109/BIBM.2015.7359790]

21. Arvind V, Cho B, Ukogu CO, Kim J, Cho SK. Wednesday, September 26, 2018 2:00 PM – 3:00 PM Integrating Technology into Practice: 59. Natural language processing of electronic medical records can identify sepsis following orthopedic surgery. In: The Spine Journal. 2018 Aug Presented at: Proceedings of the 33rd Annual Meeting of the North American Spine Society; September 26-29, 2018; Los Angeles, CA p. S29. [doi: 10.1016/j.spinee.2018.06.068]

22. Yetisgen-Yildiz M, Glavan B, Xia F, Vanderwende L, Wurfel M. Identifying patients with pneumonia from free-text intensive care unit reports. 2011 Presented at: ICML 2011 Workshop on Learning from Unstructured Clinical Text; 2011; Bellevue, WA URL: https://faculty.washington.edu/melihay/publications/ICML_2011.pdf

23. Meystre S, Haug P. Improving the sensitivity of the problem list in an intensive care unit by using natural language processing. AMIA Annu Symp Proc 2006:554-558 [FREE Full text] [Medline: 17238402]

24. Weissman GE, Harhay MO, Lugo RM, Fuchs BD, Halpern SD, Mikkelsen ME. Natural Language Processing to Assess Documentation of Features of Critical Illness in Discharge Documents of Acute Respiratory Distress Syndrome Survivors. Annals ATS 2016 Sep;13(9):1538-1545. [doi: 10.1513/annalsats.201602-131oc]

25. Marafino BJ, Park M, Davies JM, Thombley R, Luft HS, Sing DC, et al. Validation of Prediction Models for Critical Care Outcomes Using Natural Language Processing of Electronic Health Record Data. JAMA Netw Open 2018 Dec 07;1(8):e185097 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.5097] [Medline: 30646310]

26. Murff HJ, Forster AJ, Peterson JF, Fiskio JM, Heiman HL, Bates DW. Electronically screening discharge summaries for adverse medical events. J Am Med Inform Assoc 2003;10(4):339-350 [FREE Full text] [doi: 10.1197/jamia.M1201] [Medline: 12668691]

27. Kang Y, Hurdle J. Predictive Model for Risk of 30-Day Rehospitalization Using a Natural Language Processing/Machine Learning Approach Among Medicare Patients with Heart Failure. Journal of Cardiac Failure 2020 Oct;26(10):S5. [doi: 10.1016/j.cardfail.2020.09.023]

28. Yang H. Automatic extraction of medication information from medical discharge summaries. J Am Med Inform Assoc 2010 Sep 01;17(5):545-548 [FREE Full text] [doi: 10.1136/jamia.2010.003863] [Medline: 20819861]

29. Lehman L, Saeed M, Long W, Lee J, Mark R. Risk stratification of ICU patients using topic models inferred from unstructured progress notes. AMIA Annu Symp Proc 2012;2012:505-511 [FREE Full text] [Medline: 23304322]

## Abbreviations

**AP:** average precision
**ARDS:** acute respiratory distress syndrome
**AUPRC:** area under the precision recall curve
**AUROC:** area under the receiver operating characteristic curve
**BoW:** bag-of-words
**BWH:** Brigham and Women's Hospital
**EHR:** electronic health record
**FPR:** false positive rate
**ICU:** intensive care unit

XSL•FO
RenderX

**LASSO:** least absolute shrinkage and selection operator
**MGH:** Massachusetts General Hospital
**NLP:** natural language processing
**ROC:** receiver operating characteristic
**SNIF:** skilled nursing inpatient facility

XSL•FO
**RenderX**

Original Paper

# Development and Validation of a Machine Learning Approach for Automated Severity Assessment of COVID-19 Based on Clinical and Imaging Data: Retrospective Study

Juan Carlos Quiroz[1,2*], PhD; You-Zhen Feng[3*], MD; Zhong-Yuan Cheng[3*], MD; Dana Rezazadegan[1,4], PhD; Ping-Kang Chen[3], MD; Qi-Ting Lin[3], MD; Long Qian[5], PhD; Xiao-Fang Liu[6,7], PhD; Shlomo Berkovsky[1], PhD; Enrico Coiera[1], PhD; Lei Song[8], MD; Xiaoming Qiu[9], MD; Sidong Liu[1], PhD; Xiang-Ran Cai[3], MD

[1]Centre for Health Informatics, Australian Institute of Health Innovation, Faculty of Medicine, Health and Human Sciences, Macquarie University, Macquarie Park, Australia

[2]Centre for Big Data Research in Health, University of New South Wales, Sydney, Australia

[3]Medical Imaging Centre, The First Affiliated Hospital of Jinan University, Guangzhou, China

[4]Department of Computer Science and Software Engineering, Swinburne University of Technology, Melbourne, Australia

[5]Department of Biomedical Engineering, Peking University, Beijing, China

[6]Institute of Robotics and Automatic Information System, College of Artificial Intelligence, Nankai University, Tianjin, China

[7]School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

[8]Department of Radiology, Xiangyang Central Hospital, Affiliated Hospital of Hubei University of Arts and Science, Xiangyang, China

[9]Department of Radiology, Huangshi Central Hospital, Affiliated Hospital of Hubei Polytechnic University, Edong Healthcare Group, Huangshi, China

[*]these authors contributed equally

**Corresponding Author:**
Sidong Liu, PhD
Centre for Health Informatics, Australian Institute of Health Innovation
Faculty of Medicine, Health and Human Sciences
Macquarie University
75 Talvera Road
Macquarie Park, 2113
Australia
Phone: 61 29852729
Email: sidong.liu@mq.edu.au

## Abstract

**Background:** COVID-19 has overwhelmed health systems worldwide. It is important to identify severe cases as early as possible, such that resources can be mobilized and treatment can be escalated.

**Objective:** This study aims to develop a machine learning approach for automated severity assessment of COVID-19 based on clinical and imaging data.

**Methods:** Clinical data—including demographics, signs, symptoms, comorbidities, and blood test results—and chest computed tomography scans of 346 patients from 2 hospitals in the Hubei Province, China, were used to develop machine learning models for automated severity assessment in diagnosed COVID-19 cases. We compared the predictive power of the clinical and imaging data from multiple machine learning models and further explored the use of four oversampling methods to address the imbalanced classification issue. Features with the highest predictive power were identified using the Shapley Additive Explanations framework.

**Results:** Imaging features had the strongest impact on the model output, while a combination of clinical and imaging features yielded the best performance overall. The identified predictive features were consistent with those reported previously. Although oversampling yielded mixed results, it achieved the best model performance in our study. Logistic regression models differentiating between mild and severe cases achieved the best performance for clinical features (area under the curve [AUC] 0.848; sensitivity 0.455; specificity 0.906), imaging features (AUC 0.926; sensitivity 0.818; specificity 0.901), and a combination of clinical and imaging features (AUC 0.950; sensitivity 0.764; specificity 0.919). The synthetic minority oversampling method further improved the performance of the model using combined features (AUC 0.960; sensitivity 0.845; specificity 0.929).

XSL•FO
RenderX

**Conclusions:** Clinical and imaging features can be used for automated severity assessment of COVID-19 and can potentially help triage patients with COVID-19 and prioritize care delivery to those at a higher risk of severe disease.

## Introduction

COVID-19 has overwhelmed health systems worldwide [1,2]. Considering the various complications associated with COVID-19 [3-5], methods that help triage patients with COVID-19 can help prioritize care delivery to individuals at a high risk of severe or critical illness. COVID-19 severity can be categorized as follows: mild, ordinary, severe, and critical [6]. Severe and critical cases require intensive care and more health care resources than mild and ordinary cases. A high rate of false-positive severe or critical cases could overwhelm health care resources (ie, beds in the intensive care unit). Moreover, delays in identifying severe or critical cases would lead to delayed treatment of patients at a higher risk of mortality. Therefore, it is important to identify severe cases as early as possible, such that resources can be mobilized and treatment can be escalated.

Chest computed tomography (CT) scans provide important diagnostic and prognostic information [7,8]; consequently, they have been the focus of numerous recent studies using machine learning techniques for COVID-19–related prediction tasks [9-21]. Previous studies have focused on mortality predictions [9], diagnosis (identifying COVID-19 cases and differentiating them from other pulmonary diseases or no disease) [10-15,19,22-25], and severity assessment and disease progression [16-18,23]. Most current approaches have used deep learning methods and imaging features from CT scans [10-15,19,22-24] and X-ray imaging [18,20,21] with popular architectures including ResNet [10,12,14,23], U-Net [11,17], Inception [15,22], Darknet [20], and other convolutional neural networks (NNs) [18,21,26,27]. Recent reviews provide more details regarding these architectures [1,28-32].

Although automated assessment of chest CT scans to predict COVID-19 severity is of great clinical importance, few studies have focused on it [16-18,23]. Automated assessment of chest CT scans can substantially reduce the image reading time for radiologists, provide quantitative data that can be compared across patients and time points, and can be clinically applicable in disease detection and diagnosis, progression tracking, and prognosis [8]. While CT scans are an important diagnostic tool, previous studies reported that clinical data, such as symptoms, comorbidities, and laboratory findings, differed between patients with COVID-19 admitted to intensive care units and those who were not [33], and these data help predict the mortality risk [9]. A previous study compared the imaging data and clinical data of 81 patients with confirmed COVID-19 and suggested that the combination of imaging features with clinical and laboratory findings facilitated an early diagnosis of COVID-19 [34].

In this study, we used patient clinical data and imaging data to predict disease severity among patients with COVID-19. Considering this as a putative binary classification task, we predicted whether a patient diagnosed with COVID-19 is likely to have mild or severe disease. This study has 3 objectives. First, we compared the predictive power of clinical and imaging data for disease severity assessment by testing three machine learning models: logistic regression (LR) [35], gradient boosted trees (eg, XGBoost) [36], and NNs [37]. Second, since our cohort data are highly imbalanced, with the majority of cases being of mild/ordinary severity, we tested 4 oversampling methods to address the imbalanced classification issue [38-41]. Third, we interpreted the importance of features by using the Shapley Additive Explanations (SHAP) framework and identified features with the highest predictive power [42]. The predictive models evaluated herein yielded high accuracy and identified predictive imaging and clinical features consistent with those reported previously.

## Methods

### Participants

This retrospective study was performed using data collected by 2 hospitals in the Hubei Province, China. The study cohort consisted of patients with COVID-19 diagnosed through RT–PCR analysis of nasopharyngeal swab samples. A total of 346 patients from 2 hospitals were retrospectively enrolled, including 230 (66.5%) patients from Huang Shi Central Hospital (HSCH) and 116 (33.5%) from Xiang Yang Central Hospital (XYCH). These patients were admitted to hospital between January 1 and February 23, 2020, and underwent chest CT upon initial hospitalization. All participants provided written consent. This study was approved by the institutional review board of both hospitals (approval number LL-2020-032-02). Table 1 summarizes the demographic characteristics of the patients in the 2 cohorts.

**Table 1.** Demographic characteristics of the patients in the 2 cohorts (N=346).

| Category | Variables | | |
|---|---|---|---|
| | HSCH[a] (n=230) | XYCH[b] (n=116) | Total |
| **COVID-19 severity, n (%)** | | | |
| Mild | 7 (3.0) | 1 (0.9) | 8 (2.3) |
| Ordinary | 212 (92.2) | 104 (89.7) | 316 (91.3) |
| Severe | 7 (3.0) | 6 (5.2) | 13 (3.8) |
| Critical | 4 (1.7) | 5 (4.3) | 9 (2.6) |
| Age (years), mean (SD) | 49.0 (14.4) | 47.5 (17.2) | 48.5 (15.4) |
| Gender ratio (female to male) | 120:110 | 57:59 | 177:169 |

[a]HSCH: Huang Shi Central Hospital.

[b]XYCH: Xiang Yang Central Hospital.

## Imaging and Clinical Data

Chest CT scans of patients were collected upon initial hospitalization and preprocessed using intensity normalization, contrast limited adaptive histogram equalization, and gamma adjustment, using the same preprocessing pipeline as in our previous study [43]. We performed lung segmentation in the chest CT images by using an established model "R231CovidWeb" [44], which was pretrained using a large, diverse data set of non–COVID-19 chest CT scans and further fine-tuned with an additional COVID-19 data set [45]. CT slices with <3 mm$^2$ of lung tissue were excluded from the data sets since they provide limited or no information about the lung. Lung lesions were segmented using EfficientNetB7 U-Net [16],

which was also pretrained using a public COVID-19 data set [45]. The model indicated four types of lesions: ground-glass opacities, consolidations, pleural effusions, and other abnormalities. The volume of each lesion type and the total lesion volume were calculated from the segmentation maps as the imaging features and were further normalized by the lung volume. Figure 1 shows representative results of lung and lesion segmentation of a mild case and a severe case, wherein the upper row presents 3D models of the lung and lesions reconstructed using 3D Slicer (v4.6.2) [46], and the lower row presents axial chest CT slices with the lung and lesion (green: ground-glass opacities, yellow: consolidation, and brown: pleural effusion) boundaries overlaid on the CT slices.

**Figure 1.** Representative chest computed tomography scans and the lung and lesion models of (A) a mild COVID-19 case and (B) a severe COVID-19 case.



(a) Mild                                          (b) Severe
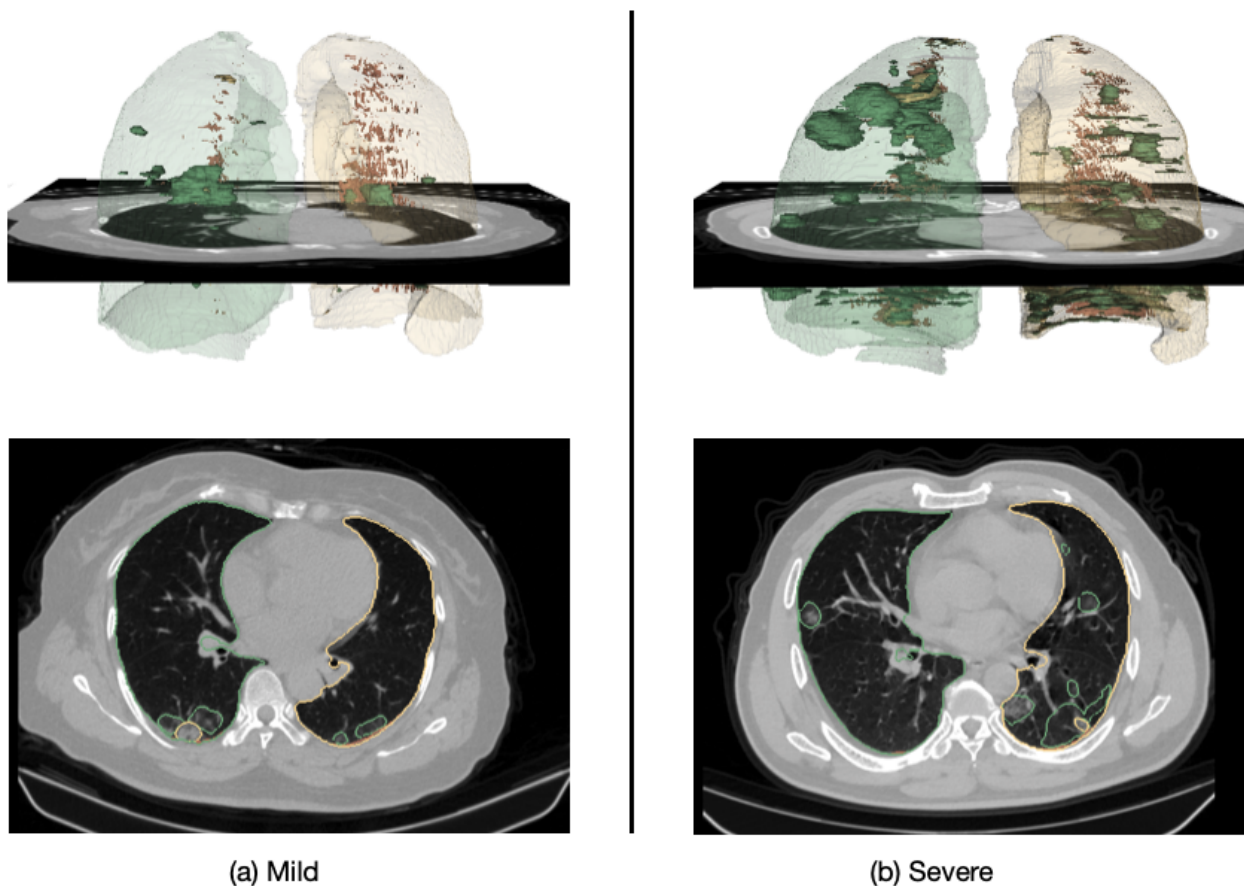
Clinical data collected from the patients included demographic characteristics, signs, symptoms, comorbidities, and the following 18 laboratory findings: white blood cell count ($\times 10^9$/L), neutrophil count ($\times 10^9$/L), lymphocyte count ($\times 10^9$/L), hemoglobin (g/L), platelets ($\times 10^9$/L), prothrombin time (s), activated partial thromboplastin time (s), D-dimer (nmol/L), C-reactive protein (mg/L), albumin (g/L), alanine aminotransferase (μkat/L), aspartate aminotransferase (μkat/L), total bilirubin (μmol/L), potassium (mmol/L), sodium (mmol/L), creatinine (μmol/L), creatine kinase (μkat/L), and lactate dehydrogenase (μkat/L).

All features were either continuous or binary—all binary features include signs, symptoms, and comorbidities. Continuous features were standardized to be centered around 0 (SD 1). Figure 2 shows the structure and dimensions of the features used in this study. These features were grouped into four feature sets: demographic characteristics and symptoms (a subset of the available clinical features), clinical features (demographic characteristics, signs and symptoms, and laboratory findings), imaging features extracted from the chest CT scans through deep learning methods, and a combination of clinical and imaging features.

**Figure 2.** Structure and dimensions of the feature sets. COPD: chronic obstructive pulmonary disorder.
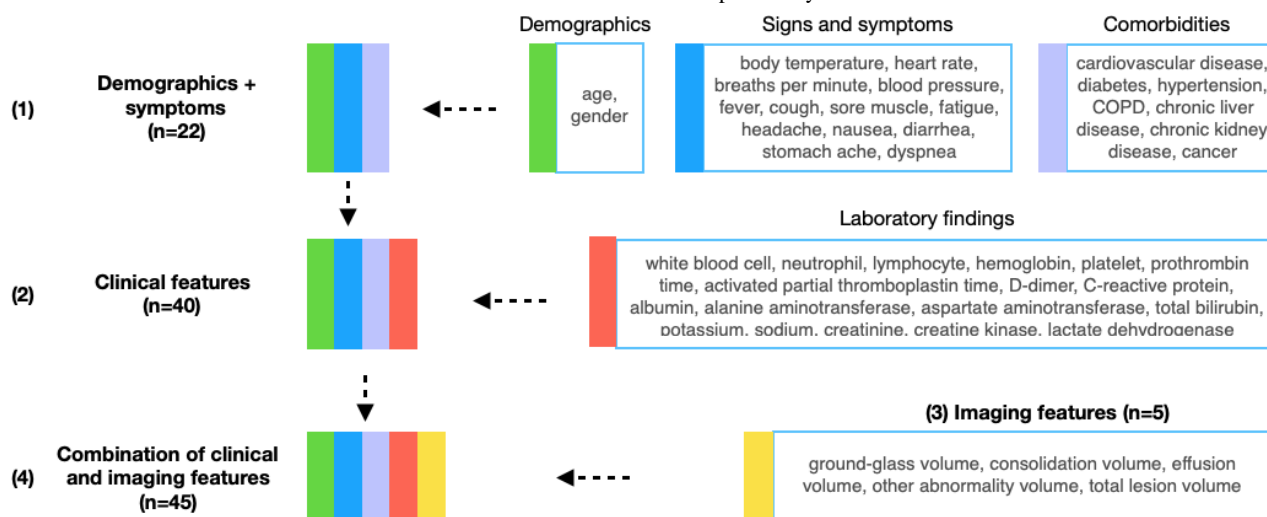


## Severity Assessment Models

We trained and compared three models to predict case severity: LR (with scikit-learn) [47], gradient boosted trees (XGBoost) [36], and an NN (fast.ai) [48]. We used the HSCH data (230 samples) for training and validation using 5-fold repeated stratified cross-validation. The XYCH data (116 samples) were withheld for testing. We reported the results for the test set with the area under the curve (AUC) and F1 scores averaged through independent runs.

Hyperparameter exploration and tuning were performed using the training/validation set. A random search was performed to tune the hyperparameters of LR and XGBoost. For NN, we used a 4-layer, fully connected architecture, with the first hidden layer having 200 nodes and a second hidden layer of 100 nodes. We determined the learning rate (0.01) using Learning Rate Finder [49]. All other NN parameters were set to default values. We explored a different number of nodes in the first and second hidden layers, with 200×100 images yielding the best results in the validation set. Of 346 patients, 167 (48%) had at least one missing feature (5.7 on average, mostly for the laboratory findings). Missing feature values were imputed with the mean for each feature.

## Oversampling

The majority of cases in our data set were of mild/ordinary severity, with only a few cases of severe/critical severity. The imbalance ratio for the entire data set was 0.07; training/validation set, 0.05; and testing set, 0.10. We tested four oversampling methods to increase the ratio of the minority class: synthetic minority oversampling (SMOTE) [38], Adaptive

Synthetic sampling [39], geometric SMOTE [40], and a conditional generative adversarial network (CTGAN) model for tabular data [41]. For these methods, we oversampled the training set, trained a model using the oversampled data, and reported results on the same test set. We adjusted the resampling ratio of all methods to 0.3 (thus setting the imbalance ratio to 0.3). Using CTGAN for oversampling, we fitted the CTGAN model with the training set, performed sampling to generate synthetic data, using only synthetic data for the minority class (severe/critical), and this was repeated until the minority-to-majority class ratio approached 0.3.

## *Results*

### Patient Characteristics

Table 2 summarizes the patients' characteristics. The differences between the mild/ordinary and severe/critical groups were assessed with the Mann-Whitney *U* test and Fisher exact test. The median age of the entire cohort was 49 (IQR 38-59) years. The median age of patients with mild/ordinary COVID-19 was 48.5 (IQR 37.0-57.3) years and that of patients with severe/critical COVID-19 was 63.0 (IQR 52.5-69.5) years. We observed significant differences between patients with severe/critical COVID-19 and those with mild/normal COVID-19 with respect to age ($P<.001$) and comorbidities including cardiovascular disease ($P=.002$), hypertension ($P=.002$), diabetes ($P=.01$), and cancer ($P=.01$). From among all signs and symptoms, an increased respiration rate ($P=.002$) and dyspnea ($P<.001$) were more common among patients with severe/critical COVID-19 than among those with mild/ordinary COVID-19.

**Table 2.** Demographics and baseline characteristics of patients with confirmed COVID-19 (N=346). Symptoms including cardiovascular disease and shortness of breath were more likely in cases of severe/critical COVID-19.

| Characteristics | Patients | | | |
| --- | --- | --- | --- | --- |
| | All patients | Mild/ordinary | Severe/critical | P value[a] |
| Sample size, n | 346 | 324 | 22 | N/A[b] |
| **Demographic characteristics** | | | | |
| Age (years), median (IQR) | 49.0 (38.0-59.0) | 48.5 (37.0-57.3) | 63.0 (52.5-69.5) | <.001 |
| **Gender, n (%)** | | | | .38 |
| Female | 177 (51.2) | 168 (51.9) | 9 (41.0) | |
| Male | 169 (48.8) | 156 (48.1) | 13 (59.0) | |
| **Comorbidities, n (%)** | | | | |
| Cardiovascular disease | 40 (11.6) | 32 (9.9) | 8 (36.0) | .002 |
| Diabetes | 34 (9.8) | 28 (8.6) | 6 (27.0) | .01 |
| Hypertension | 51 (14.7) | 42 (13.0) | 9 (41.0) | .002 |
| Chronic obstructive pulmonary disease | 11 (3.2) | 9 (2.8) | 2 (9.0) | .15 |
| Chronic liver disease | 7 (2.0) | 7 (2.2) | 0 (0) | N/A |
| Chronic kidney disease | 4 (1.2) | 3 (0.9) | 1 (5.0) | .20 |
| Cancer | 8 (2.3) | 5 (1.5) | 3 (14.0) | .01 |
| **Signs, median (IQR)** | | | | |
| Body temperature | 37.8 (37-38.3) | 37.8 (37-38.3) | 38.1 (37.1-39) | .11 |
| Heart rate | 85 (80-90) | 85 (80-90) | 90 (80-101.8) | .11 |
| Breaths per minute | 20 (20-21) | 20 (20-21) | 21 (20-28) | .002 |
| Blood pressure high | 120 (119.5-130.0) | 120 (118.5-130.0) | 127 (120-146.5) | .07 |
| Blood pressure low | 74 (69-80) | 74 (69-80) | 79.5 (71-89) | .08 |
| **Symptoms, n (%)** | | | | |
| Fever | 275 (79.5) | 256 (79.0) | 19 (86.0) | .59 |
| Cough | 238 (68.8) | 220 (67.9) | 18 (82.0) | .24 |
| Fatigue | 118 (34.1) | 108 (33.3) | 10 (45.0) | .25 |
| Dyspnea | 32 (9.2) | 23 (7.1) | 9 (41.0) | <.001 |
| Sore muscle | 38 (11.0) | 35 (10.8) | 3 (14.0) | .72 |
| Headache | 34 (9.9) | 31 (9.6) | 3 (14.0) | .47 |
| Diarrhea | 23 (6.6) | 20 (6.2) | 3 (14.0) | .17 |
| Nausea | 9 (2.6) | 7 (2.2) | 2 (9.0) | .11 |

[a]P values were compared using mild/ordinary and severe/critical cases were obtained with Mann-Whitney U test and Fisher exact test. As no patient in our cohort had a stomach ache, this feature was not factored into our model.

[b]N/A: not applicable.

## Prediction of COVID-19 Severity at Baseline

Data from the HSCH (230 patients, 66.5%) were used for training and validation, and data from the XYCH (116 patients, 33.5%) were used as the independent test set. We compared model performance using four feature sets: demographic characteristics and symptoms, clinical features, imaging features, and a combination of clinical and imaging features (Figure 2). The optimal classification threshold for the sensitivity, specificity, and F1 score was identified using the Youden index

[50]. Table 3 shows the severity assessment performance of an LR model, an XGBoost model, and a 4-layer fully connected NN model. Overall, LR models outperformed the other evaluated models, achieving the highest AUC, F1 score, and sensitivity for all four feature sets. While imaging features yielded substantially better results than clinical features, the combination of clinical and imaging features benefited only the LR model. Hence, the LR model displayed the best performance (AUC 0.950; F1 score 0.604; sensitivity 0.764; specificity 0.919) upon using the combination of clinical and imaging features.

XSL•FO
RenderX

**Table 3.** Results of using different feature sets (values in italics indicate the best results).

| Feature sets and model | Area under the curve | F1 score | Sensitivity | Specificity |
|---|---|---|---|---|
| **Demographics + symptoms** | | | | |
| LR[a] | *0.819* | *0.382* | *0.627* | 0.825 |
| XGB[b] | 0.763 | 0.363 | 0.318 | *0.956* |
| NN[c] | 0.730 | 0.332 | 0.427 | 0.880 |
| **Clinical** | | | | |
| LR | *0.848* | *0.387* | *0.455* | 0.906 |
| XGB | 0.787 | 0.286 | 0.227 | *0.962* |
| NN | 0.647 | 0.237 | 0.309 | 0.881 |
| **Imaging** | | | | |
| LR | *0.926* | *0.593* | *0.818* | 0.901 |
| XGB | 0.904 | 0.486 | 0.636 | 0.896 |
| NN | 0.845 | 0.555 | 0.600 | *0.936* |
| **Clinical + imaging** | | | | |
| LR | *0.950* | *0.604* | *0.764* | 0.919 |
| XGB | 0.904 | 0.520 | 0.473 | *0.965* |
| NN | 0.782 | 0.413 | 0.486 | 0.907 |

[a]LR: logistic regression.

[b]XGB: XGBoost.

[c]NN: neural network.

## Prediction at Baseline Severity With Oversampling

Since the cohort was highly imbalanced, with the majority of cases being of mild/ordinary severity (imbalance ratio 0.07), we applied four oversampling methods to increase the ratio of severe/critical cases: SMOTE [38], Adaptive Synthetic sampling [39], geometric SMOTE [40], and CTGAN [41]. Figure 3 shows the differences in AUC values and F1 scores obtained through oversampling, with negative values indicating a reduction in AUC or F1 scores and positive values indicating the opposite trend. Oversampling resulted in greater improvements in the F1 score than in the AUC. The greatest improvement in the F1 score (0.09) was observed for the clinical features (clinical) with XGBoost and SMOTE (XGB-smo); however, the AUC decreased by 0.08 with the same method. Considering both AUC and F1 scores simultaneously, the combination of clinical and imaging features (clinical + imaging) benefited most from oversampling. In particular, the AUC and F1 score for clinical + imaging features were increased by 0.01 and 0.06, respectively, using LR with SMOTE (LR-smo).

**Figure 3.** Differences in the (A) area under the curve values and (B) F1 scores with oversampling and without oversampling. Positive values (blue) indicate oversampling resulting in higher values, negative values (red) indicating oversampling resulting in lower values. smo: synthetic minority oversampling; ada: Adaptive Synthetic sampling; geo = geometric synthetic minority oversampling; gan: conditional generative adversarial network; LR: logistic regression; NN: neural network; XGB: XGBoost.



(a) differences in AUC values
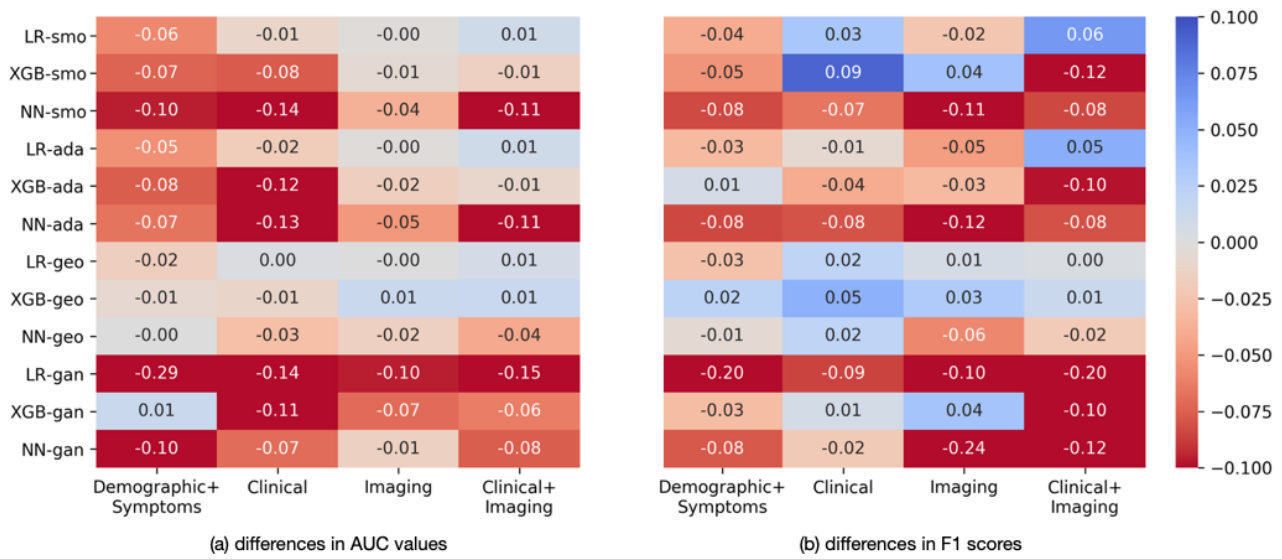
(b) differences in F1 scores

Table 4 presents the best results of the evaluated models using various feature sets after oversampling. Oversampling did not improve the performance of the LR model for the demographic characteristics + symptoms features, but SMOTE and geometric SMOTE increased the F1 scores for clinical features and imaging features, respectively. Notably, the performance of the LR model (Table 3) was optimal for the combination of clinical and imaging features, with improvements in the AUC (0.960 vs 0.950), F1 score (0.668 vs 0.604), sensitivity (0.845 vs 0.764), and specificity (0.929 vs 0.919), after oversampling with SMOTE.

**Table 4.** The best results obtained using different feature sets after oversampling (arrow indicates improved performance after oversampling).

| Feature sets | Results | | | | |
|---|---|---|---|---|---|
| | Model | Area under the curve | F1 | Sensitivity | Specificity |
| Demographics + symptoms | LR[a,b] | 0.819 | 0.382 | 0.627 | 0.825 |
| Clinical | LR – smo[c] | 0.837 | 0.421 ↑ | 0.518 ↑ | 0.902 |
| Imaging | LR – geo[d] | 0.926 | 0.599 ↑ | 0.818 | 0.904 ↑ |
| Clinical + imaging | LR – smo | 0.960 ↑ | 0.668 ↑ | 0.845 ↑ | 0.929 ↑ |

[a]LR: logistic regression.

[b]No improvement after oversampling.

[c]smo: synthetic minority oversampling.

[d]geo: geometric synthetic minority oversampling.

## Model Interpretation

We used the SHAP framework [42] to interpret the output of the best-performing LR model through SMOTE oversampling. This framework helps determine the importance of a feature by comparing model predictions with or without the feature.

Figure 4 shows a SHAP plot summarizing how the values of each feature impact the model output of the LR model using all features (clinical and imaging features), with features sorted in descending order of importance. Figure 4A shows the feature importance scores sorted by the average impact on the model output, and Figure 4B shows the SHA*P* values of individual features. Furthermore, 4 imaging features, including consolidation volume (consolidation_val), total lesion volume (lesion_vol), ground-glass volume (groundglass_vol), and volume of other abnormalities (other_vol), are among the top 6 features, their high values increasing the likelihood of the model to predict a severe/critical COVID-19 case. Low albumin levels, high C-reactive protein levels, a high leukocyte count, and low lactate dehydrogenase levels make the model more likely to predict a critical/severe COVID-19 case. Moreover, older age and male gender increased the likelihood of the model to predict severe/critical COVID-19 cases.

**Figure 4.** (A) Feature importance, evaluated using the mean SHAP (Shapely Addictive Explanations) values, in the logistic regression (LR) model using all features. (B) SHAP plot for the LR model using all features. Each point represents a feature instance, and the color indicates the feature value (red: high, blue: low). Negative SHAP values indicate feature instances contributing to a model output of a mild/ordinary COVID-19 case, whereas positive SHAP values indicate features contributing to a model output of a severe/critical COVID-19 case.
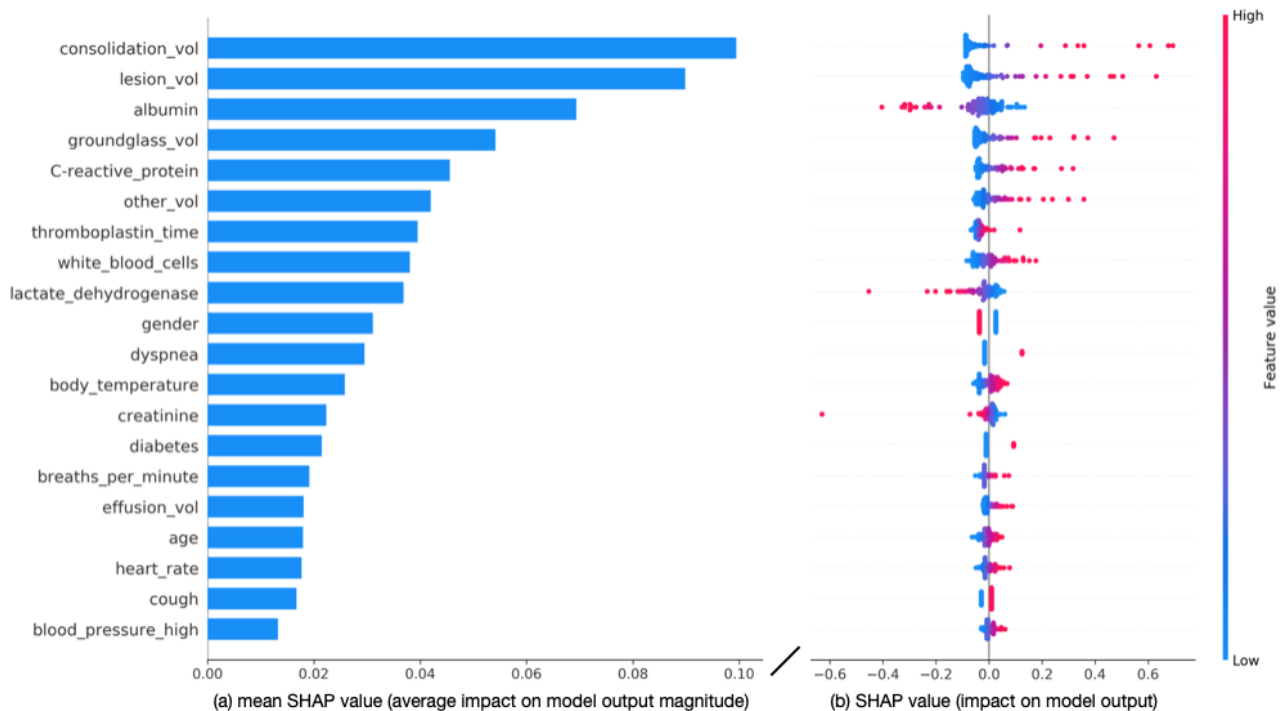


## Discussion

### Principal Findings

In our cohort of patients with COVID-19, fever, cough, and fatigue were the most common symptoms, consistent with previous studies on COVID-19 [34]. The incidence of dyspnea and an increased respiratory rate was significantly higher in severe cases. Some symptoms such as sore muscle, headache, diarrhea, and nausea were present in 9–38 (2.6%-11.0%) of patients and did not differ significantly between mild and severe cases. Patients with severe COVID-19 tended to be of older age and had comorbidities (including cardiovascular disease, diabetes, hypertension, and cancer), concurrent with previous studies [1,3,5,34]. We observed no difference between males and females in our cohort, although the model did rely on gender for increasing the likelihood of predicting a severe/critical case.

A combination of clinical and imaging features yielded the best performance. Imaging features had the strongest impact on model output, with high values of consolidation volume, lesion volume, ground-glass volume, and other volume increasing the likelihood of the model to predict a severe case of COVID-19. Ground-glass opacity is an important feature of COVID-19 [14]. The inclusion of clinical features further improved the accuracy of severity assessment, with findings such as albumin levels, C-reactive protein levels, thromboplastin time, white blood cell counts, and lactate dehydrogenase levels being amongst the most informative features, concurrent with a previous study that also used laboratory findings to predict COVID-19–related mortality [9]. Furthermore, C-reactive protein was associated with a significant risk of critical illness in a study of 5279 patients with laboratory-confirmed COVID-19 [5]. Our model

also relied on symptoms and patient characteristics such as gender, dyspnea, body temperature, diabetes, and respiratory rate for differentiating between mild and severe cases. Clinical features alone (demographics, signs, symptoms, and laboratory results), resulted in low sensitivity. Therefore, dependence on only clinical features poses the risk of predicting mild/ordinary COVID-19 among patients at the risk of critical/severe illness.

Oversampling yielded mixed results, although it revealed the best model performance in our study. The best model without oversampling (ie, the LR model) also yielded remarkable findings (AUC 0.950; F1 0.604; sensitivity 0.764; specificity 0.919), and SMOTE oversampling further improved the model performance (AUC 0.960; F1 0.668; sensitivity 0.845; specificity 0.929). Considering the propensity of health care data to be imbalanced [51-54], our results suggest the need for further analysis of oversampling methods for medical data sets. Self-supervision [55,56] may also help improve the performance of models using imbalanced medical data sets; in particular, future studies should evaluate the impact of self-supervision on tabular medical data.

### Clinical Implications

The rapid spread of COVID-19 has overwhelmed health care systems, necessitating methods for efficient disease severity assessment. Our results indicate that clinical and imaging features can facilitate automated severity assessment of COVID-19. While our study would benefit from a larger data set, our results are encouraging because we trained the models with data from one hospital only and tested them using an independent data set from another hospital, albeit with high predictive accuracy.

The proposed methods and models would be useful in several clinical scenarios. First, the proposed models are fully automated and can expedite the assessment process, saving time in reading CT scans or evaluating patients through a scoring system. These models can be useful in hospitals that are overwhelmed by a high volume of patients during the outbreak by identifying severe cases as early as possible, such that treatment can be escalated. Our models, with low sensitivity and high specificity, are best used in combination with a model with high sensitivity and low specificity. A high-sensitivity model can identify patients with severe COVID-19, and our model (with high specificity) could identify false-positives; that is, patients with mild COVID-19 who were wrongly identified as having severe COVID-19.

Our models were developed and validated using 4 different feature sets, providing the flexibility to accommodate patients with different available data. For example, if a patient has neither a chest CT scan nor a blood test, the model based on demographics and symptoms can still achieve reasonably good prediction performance (AUC 0.819; sensitivity 0.627; specificity 0.825). Availability of the patients' clinical and imaging features can improve the model's sensitivity and specificity, with the potential to triage patients with COVID-19 (eg, prioritizing care for patients at a higher risk of mortality).

### Limitations and Future Prospects

Our data set consisted of 346 patients with confirmed COVID-19, with data on 230 (66.5%) patients from HSCH used for training/validation and data on 116 (33.5%) patients from XYCH used for testing. Our data set was highly imbalanced, which could have made models overfit to the majority class. In addition, only the baseline data for patients were used in this study; therefore, we could not assess how early can COVID-19 progression be detected. We intend to further investigate the longitudinal data and design computational models to predict disease progression in our future studies.

While we explored various NN configurations, the results were not comparable to those of LR, presumably owing to the limited data set and the low dimensionality of the feature vectors. In this study, we used a complex NN model (EfficientNetB7 U-Net) to extract the imaging features and tested various models for classification using the combination of imaging features and tabular clinical data. Such 2-stage processing may simplify the classification task for these models, thereby reducing the need for another NN model for classification owing to low dimensionality of the features. Further exploration of NN architectures for tabular data would likely improve the performance of the NN model, especially if more data are available.

During training and validation, the performance of the models across cross-validation folds showed high variance owing to the small number of positive cases in the validation fold. A larger dataset would improve the reliability and robustness of the models. The data also consisted of COVID-19 cases which were confirmed through RT–PCR analysis of nasopharyngeal swabs. As such, our model is limited to differentiating severe/critical cases from mild/ordinary cases of COVID-19 and not for diagnosing COVID-19 or differentiating COVID-19 cases from those of other respiratory tract infections. Further studies are required to determine the efficacy of the severity assessments, including data from asymptomatic patients.

Using the Prediction Model Study Risk of Bias Assessment Tool [57], our models are at a high risk of bias owing to a potential bias in the participants domain (the cohort including participants [mean age 48.5 years, SD 15.4 years] who were admitted to hospitals) and the analysis domain (small sample size and class imbalance). Our models are at a low risk of bias in the predictor and outcome domains.

### Conclusions

This study presents a novel method for severity assessment of patients diagnosed with COVID-19. Our results indicate that clinical and imaging features can be used for automated severity assessment of COVID-19. While imaging features had the strongest impact on the model's performance, inclusion of clinical features and oversampling yielded the best performance in our study. The proposed method may potentially help triage patients with COVID-19 and prioritize care for patients at a higher risk of severe disease.

### Authors' Contributions

SL, XQ, and X-RC share corresponding author responsibilities; SL (sidong.liu@mq.edu.au) will respond to technical inquiries, and XQ (qxm2020cov@163.com), and X-RC (caixran@jnu.edu.cn) will respond to inquiries regarding clinical data and applications.

### Conflicts of Interest

None declared.

XSL•FO

RenderX

## References

1. Siordia J. Epidemiology and clinical features of COVID-19: A review of current literature. J Clin Virol 2020 Jun;127:104357 [FREE Full text] [doi: 10.1016/j.jcv.2020.104357] [Medline: 32305884]

2. Angulo FJ, Finelli L, Swerdlow DL. Reopening Society and the Need for Real-Time Assessment of COVID-19 at the Community Level. JAMA 2020 Jun 09;323(22):2247-2248. [doi: 10.1001/jama.2020.7872] [Medline: 32412582]

3. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, the Northwell COVID-19 Research Consortium, et al. Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. JAMA 2020 May 26;323(20):2052-2059 [FREE Full text] [doi: 10.1001/jama.2020.6775] [Medline: 32320003]

4. Madjid M, Safavi-Naeini P, Solomon SD, Vardeny O. Potential Effects of Coronaviruses on the Cardiovascular System: A Review. JAMA Cardiol 2020 Jul 01;5(7):831-840. [doi: 10.1001/jamacardio.2020.1286] [Medline: 32219363]

5. Petrilli CM, Jones SA, Yang J, Rajagopalan H, O'Donnell L, Chernyak Y, et al. Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: prospective cohort study. BMJ 2020 May 22;369:m1966 [FREE Full text] [doi: 10.1136/bmj.m1966] [Medline: 32444366]

6. Jin X, Pang B, Zhang J, Liu Q, Yang Z, Feng J, et al. Core Outcome Set for Clinical Trials on Coronavirus Disease 2019 (COS-COVID). Engineering (Beijing) 2020 Oct;6(10):1147-1152 [FREE Full text] [doi: 10.1016/j.eng.2020.03.002] [Medline: 32292626]

7. Inui S, Fujikawa A, Jitsu M, Kunishima N, Watanabe S, Suzuki Y, et al. Chest CT Findings in Cases from the Cruise Ship with Coronavirus Disease (COVID-19). Radiology: Cardiothoracic Imaging 2020 Apr 01;2(2):e200110-e201180 [FREE Full text] [doi: 10.1148/ryct.2020200110] [Medline: 32565152]

8. Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, et al. Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases. Radiology 2020 Feb 26:200642. [doi: 10.1148/radiol.2020200642] [Medline: 32101510]

9. Yan L, Zhang H, Xiao Y, Wang M, Sun C, Liang J, et al. A Machine Learning-Based Model for Survival Prediction in Patients With Severe COVID-19 Infection. medRxiv Preprint posted online March 17, 2020. [FREE Full text] [doi: 10.1101/2020.02.27.20028027]

10. Butt C, Gill J, Chun D, Babu BA. Deep learning system to screen coronavirus disease 2019 pneumonia [retracted in: Appl Intell. 2020]. Appl Intell 2020 Apr 22. [doi: 10.1007/s10489-020-01714-3]

11. Chen J, Wu L, Zhang J, Zhang L, Gong D, Zhao Y, et al. Deep Learning-Based Model for Detecting 2019 Novel Coronavirus Pneumonia on High-Resolution Computed Tomography: A Prospective Study. medRxiv Preprint posted online March 01, 2020. [FREE Full text] [doi: 10.1101/2020.02.25.20021568]

12. Gozes O, Frid-Adar M, Greenspan H, Browning P, Zhang H, Ji W, et al. Rapid AI Development Cycle for the Coronavirus (COVID-19) Pandemic: Initial Results for Automated Detection and Patient Monitoring using Deep Learning CT Image Analysis. arXiv Preprint posted online March 24, 2020 [FREE Full text]

13. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, et al. Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy. Radiology 2020 Aug;296(2):E65-E71 [FREE Full text] [doi: 10.1148/radiol.2020200905] [Medline: 32191588]

14. Song Y, Zheng S, Li L, Zhang X, Zhang X, Huang Z, et al. Deep Learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) With CT Images. medRxiv Preprint posted online February 25, 2020. [FREE Full text] [doi: 10.1101/2020.02.23.20026930]

15. Wang S, Kang B, Ma J, Zeng X, Xiao M, Guo J, et al. A Deep Learning Algorithm Using CT Images to Screen for Coronavirus Disease (COVID-19). medRxiv Preprint posted online April 24, 2020. [doi: 10.1101/2020.02.14.20023028]

16. Feng Y, Liu S, Cheng Z, Quiroz J, Rezazadegan D, Chen P, et al. Severity Assessment Progression Prediction of COVID-19 Patients Based on the LesionEncoder FrameworkChest CT. medRxiv Preprint posted online August 06, 2020. [doi: 10.1101/2020.08.03.20167007]

17. Chaganti S, Balachandran A, Chabin G, Cohen S, Flohr T, Georgescu B, et al. Quantification of Tomographic Patterns associated with COVID-19 from Chest CT. arXiv Preprint posted online April 2, 2020. [Medline: 32550252]

18. Wong A, Lin Z, Wang L, Chung A, Shen B, Abbasi A, et al. Towards computer-aided severity assessment: training and validation of deep neural networks for geographic extent and opacity extent scoring of chest X-rays for SARS-CoV-2 lung disease severity. arXiv Preprint posted online August 22, 2020. [FREE Full text]

19. Bai HX, Wang R, Xiong Z, Hsieh B, Chang K, Halsey K, et al. Artificial Intelligence Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Origin at Chest CT. Radiology 2020 Sep;296(3):E156-E165 [FREE Full text] [doi: 10.1148/radiol.2020201491] [Medline: 32339081]

20. Ozturk T, Talo M, Yildirim E, Baloglu U, Yildirim O, Rajendra Acharya U. Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput Biol Med 2020 Jun;121:103792 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.103792] [Medline: 32568675]

21. Wang L, Lin ZQ, Wong A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci Rep 2020 Nov 11;10(1):19549 [FREE Full text] [doi: 10.1038/s41598-020-76550-z] [Medline: 33177550]

XSL•FO

RenderX

22. Mei X, Lee H, Diao K, Huang M, Lin B, Liu C, et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. Nat Med 2020 Aug;26(8):1224-1228 [FREE Full text] [doi: 10.1038/s41591-020-0931-3] [Medline: 32427924]

23. Xiao L, Li P, Sun F, Zhang Y, Xu C, Zhu H, et al. Development and Validation of a Deep Learning-Based Model Using Computed Tomography Imaging for Predicting Disease Severity of Coronavirus Disease 2019. Front Bioeng Biotechnol 2020;8:898 [FREE Full text] [doi: 10.3389/fbioe.2020.00898] [Medline: 32850746]

24. Qian X, Fu H, Shi W, Chen T, Fu Y, Shan F, et al. M Lung-Sys: A Deep Learning System for Multi-Class Lung Pneumonia Screening From CT Imaging. IEEE J Biomed Health Inform 2020 Dec;24(12):3539-3550. [doi: 10.1109/JBHI.2020.3030853] [Medline: 33048773]

25. Hu S, Gao Y, Niu Z, Jiang Y, Li L, Xiao X, et al. Weakly Supervised Deep Learning for COVID-19 Infection Detection and Classification From CT Images. IEEE Access 2020 Aug;8(8):118869-118883. [doi: 10.1109/ACCESS.2020.3005510] [Medline: 33156775]

26. Wang Z, Liu Q, Dou Q. Contrastive Cross-Site Learning With Redesigned Net for COVID-19 CT Classification. IEEE J Biomed Health Inform 2020 Oct;24(10):2806-2813. [doi: 10.1109/JBHI.2020.3023246] [Medline: 32915751]

27. Babukarthik RG, Adiga VAK, Sambasivam G, Chandramohan D, Amudhavel J. Prediction of COVID-19 Using Genetic Deep Learning Convolutional Neural Network (GDCNN). IEEE Access 2020;8:177647-177666. [doi: 10.1109/access.2020.3025164]

28. Ng M, Lee EYP, Yang J, Yang F, Li X, Wang H, et al. Imaging Profile of the COVID-19 Infection: Radiologic Findings and Literature Review. Radiology: Cardiothoracic Imaging 2020 Feb 01;2(1):e200034. [doi: 10.1148/ryct.2020200034]

29. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ 2020 Apr 07;369:m1328 [FREE Full text] [doi: 10.1136/bmj.m1328] [Medline: 32265220]

30. Shi F, Wang J, Shi J, Wu Z, Wang Q, Tang Z, et al. Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19. IEEE Rev Biomed Eng 2021;14:4-15. [doi: 10.1109/RBME.2020.2987975] [Medline: 32305937]

31. Jamshidi M, Lalbakhsh A, Talla J, Peroutka Z, Hadjilooei F, Lalbakhsh P, et al. Artificial Intelligence and COVID-19: Deep Learning Approaches for Diagnosis and Treatment. IEEE Access 2020;8:109581-109595. [doi: 10.1109/ACCESS.2020.3001973]

32. Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos Solitons Fractals 2020 Oct;139:110059 [FREE Full text] [doi: 10.1016/j.chaos.2020.110059] [Medline: 32834612]

33. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020 Feb 15;395(10223):497-506 [FREE Full text] [doi: 10.1016/S0140-6736(20)30183-5] [Medline: 31986264]

34. Shi H, Han X, Jiang N, Cao Y, Alwalid O, Gu J, et al. Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. Lancet Infect Dis 2020 Apr;20(4):425-434 [FREE Full text] [doi: 10.1016/S1473-3099(20)30086-4] [Medline: 32105637]

35. Bishop C. In: Jordan M, Kleinberg J, Schölkopf B, editors. Pattern Recognition and Machine Learning (Information Science and Statistics). New York, NY: Springer-Verlag; 2006.

36. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. New York, NY: Association for Computing Machinery; 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 2016; New York, NY. [doi: 10.1145/2939672.2939785]

37. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. 1943. Bull Math Biol 1990;52(1-2):99-115; discussion 73. [Medline: 2185863]

38. Chawla NV, Bowyer K, Hall L, Kegelmeyer W. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res 2002 Jun 01;16:321-357. [doi: 10.1613/jair.953]

39. He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 Presented at: IEEE International Joint Conference on Neural Networks; 1-8 June 2008; Hong Kong URL: https://ieeexplore.ieee.org/document/4633969 [doi: 10.1109/IJCNN.2008.4633969]

40. Douzas G, Bacao F. Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE. Information Sciences 2019 Oct;501:118-135 [FREE Full text] [doi: 10.1016/j.ins.2019.06.007]

41. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling Tabular data using Conditional GAN. arXiv Preprint posted online October 28, 2019. [FREE Full text]

42. Lundberg S, Lee S. A Unified Approach to Interpreting Model Predictions. arXiv Preprint posted online November 25, 2017. [FREE Full text]

43. Liu S, Shah Z, Sav A, Russo C, Berkovsky S, Qian Y, et al. Isocitrate dehydrogenase (IDH) status prediction in histopathology images of gliomas using deep learning. Sci Rep 2020 May 07;10(1):7733 [FREE Full text] [doi: 10.1038/s41598-020-64588-y] [Medline: 32382048]

44. Hofmanninger J, Prayer F, Pan J, Röhrich S, Prosch H, Langs G. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. Eur Radiol Exp 2020 Aug 20;4(1):50 [FREE Full text] [doi: 10.1186/s41747-020-00173-2] [Medline: 32814998]

45. COVID-19 CT segmentation dataset. MedSeg. 2020. URL: http://medicalsegmentation.com/covid19/ [accessed 2021-02-05]

46. Kikinis R, Pieper S, Vosburgh K. 3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support. In: Intraoperative Imaging and Image-Guided Therapy. New York, NY: Springer; 2013:277-289.

47. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12:2825-2830.

48. Howard J, Gugger S. Fastai: A Layered API for Deep Learning. Information 2020 Feb 16;11(2):108 [FREE Full text] [doi: 10.3390/info11020108]

49. Smith L. Cyclical Learning Rates for Training Neural Networks. : IEEE; 2017 Presented at: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV); 24-31 March 2017; Santa Rosa, CA. [doi: 10.1109/wacv.2017.58]

50. Youden WJ. Index for rating diagnostic tests. Cancer 1950 Jan;3(1):32-35. [doi: 10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3] [Medline: 15405679]

51. Krawczyk B, Schaefer G, Woźniak M. A hybrid cost-sensitive ensemble for imbalanced breast thermogram classification. Artif Intell Med 2015 Nov;65(3):219-227. [doi: 10.1016/j.artmed.2015.07.005] [Medline: 26319694]

52. Jiang J, Liu X, Zhang K, Long E, Wang L, Li W, et al. Automatic diagnosis of imbalanced ophthalmic images using a cost-sensitive deep convolutional neural network. Biomed Eng Online 2017 Nov 21;16(1):132 [FREE Full text] [doi: 10.1186/s12938-017-0420-1] [Medline: 29157240]

53. Gan D, Shen J, An B, Xu M, Liu N. Integrating TANBN with cost sensitive classification algorithm for imbalanced data in medical diagnosis. Computers & Industrial Engineering 2020 Feb;140:106266 [FREE Full text] [doi: 10.1016/j.cie.2019.106266]

54. Zhang L, Yang H, Jiang Z. Imbalanced biomedical data classification using self-adaptive multilayer ELM combined with dynamic GAN. Biomed Eng Online 2018 Dec 04;17(1):181 [FREE Full text] [doi: 10.1186/s12938-018-0604-3] [Medline: 30514298]

55. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 2013 Aug;35(8):1798-1828. [doi: 10.1109/TPAMI.2013.50] [Medline: 23787338]

56. Raina R, Battle A, Lee H, Packer B, Ng A. Self-Taught Learning: Transfer Learning From Unlabeled Data. New York, NY: Association for Computing Machinery; 2007 Presented at: ICML '07: Proceedings of the 24th International Conference on Machine learning; June 2007; Corvalis, OR. [doi: 10.1145/1273496.1273592]

57. Wolff R, Moons K, Riley R, Whiting P, Westwood M, Collins G, PROBAST Group†. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med 2019 Jan 01;170(1):51-58 [FREE Full text] [doi: 10.7326/M18-1376] [Medline: 30596875]

## Abbreviations

**AUC:** area under the curve
**CT:** computed tomography
**CTGAN:** conditional generative adversarial network
**HSCH:** Huang Shi Central Hospital
**LR:** logistic regression
**NHMRC:** National Health and Medical Research Council
**NN:** neural network
**SHAP:** Shapley Additive Explanations
**SMOTE:** synthetic minority oversampling
**XYCH:** Xiang Yang Central Hospital

XSL•FO
RenderX

XSL•FO
**RenderX**

Original Paper

# Lexicon Development for COVID-19-related Concepts Using Open-source Word Embedding Sources: An Intrinsic and Extrinsic Evaluation

Soham Parikh[1*], MS; Anahita Davoudi[2*], MS, PhD; Shun Yu[3], MD; Carolina Giraldo[4], BSc; Emily Schriver[5], MSc; Danielle Mowery[6], MS, PhD

[1]School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA, United States

[2]Department of Biostatistics, Epidemiology, & Informatics, University of Pennsylvania, Philadelphia, PA, United States

[3]Division of Hematology/Oncology, Department of Medicine, Hospital of the University of Pennsylvania, Philadelphia, PA, United States

[4]Philadelphia College of Osteopathic Medicine, Philadelphia, PA, United States

[5]Data Analytics Center, Penn Medicine, Philadelphia, PA, United States

[6]Department of Biostatistics, Epidemiology, & Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, United States

[*]these authors contributed equally

**Corresponding Author:**
Danielle Mowery, MS, PhD
Department of Biostatistics, Epidemiology, & Informatics
Institute for Biomedical Informatics
University of Pennsylvania
A206 Richards Hall,
3700 Hamilton Walk
Philadelphia, PA, 19104-6021
United States
Phone: 1 215 746 6677
Email: dlmowery@pennmedicine.upenn.edu

## Abstract

**Background:**   Scientists are developing new computational methods and prediction models to better clinically understand COVID-19 prevalence, treatment efficacy, and patient outcomes. These efforts could be improved by leveraging documented COVID-19–related symptoms, findings, and disorders from clinical text sources in an electronic health record. Word embeddings can identify terms related to these clinical concepts from both the biomedical and nonbiomedical domains, and are being shared with the open-source community at large. However, it's unclear how useful openly available word embeddings are for developing lexicons for COVID-19–related concepts.

**Objective:**   Given an initial lexicon of COVID-19–related terms, this study aims to characterize the returned terms by similarity across various open-source word embeddings and determine common semantic and syntactic patterns between the COVID-19 queried terms and returned terms specific to the word embedding source.

**Methods:**   We compared seven openly available word embedding sources. Using a series of COVID-19–related terms for associated symptoms, findings, and disorders, we conducted an interannotator agreement study to determine how accurately the most similar returned terms could be classified according to semantic types by three annotators. We conducted a qualitative study of COVID-19 queried terms and their returned terms to detect informative patterns for constructing lexicons. We demonstrated the utility of applying such learned synonyms to discharge summaries by reporting the proportion of patients identified by concept among three patient cohorts: pneumonia (n=6410), acute respiratory distress syndrome (n=8647), and COVID-19 (n=2397).

**Results:**   We observed high pairwise interannotator agreement (Cohen kappa) for symptoms (0.86-0.99), findings (0.93-0.99), and disorders (0.93-0.99). Word embedding sources generated based on characters tend to return more synonyms (mean count of 7.2 synonyms) compared to token-based embedding sources (mean counts range from 2.0 to 3.4). Word embedding sources queried using a qualifier term (eg, dry cough or muscle pain) more often returned qualifiers of the similar semantic type (eg, "dry" returns consistency qualifiers like "wet" and "runny") compared to a single term (eg, cough or pain) queries. A higher proportion of patients had documented fever (0.61-0.84), cough (0.41-0.55), shortness of breath (0.40-0.59), and hypoxia (0.51-0.56)

retrieved than other clinical features. Terms for dry cough returned a higher proportion of patients with COVID-19 (0.07) than the pneumonia (0.05) and acute respiratory distress syndrome (0.03) populations.

**Conclusions:** Word embeddings are valuable technology for learning related terms, including synonyms. When leveraging openly available word embedding sources, choices made for the construction of the word embeddings can significantly influence the words learned.

## *Introduction*

### Background

COVID-19 has become a pandemic that is felt throughout the world. Scientists are developing new methods for determining infection rates, disease burden, treatment efficacy, and patient outcomes [1]. Our ability to detect and phenotype patients with COVID-19 and controls for clinical and translational studies requires clinical symptomatology, radiological imaging, laboratory tests, and associated disorders derived from electronic health record data [2]. Much of this information can be locked within the electronic health record clinical notes [3]. To accurately characterize each patient's COVID-19 profile for study, we must develop natural language processing systems to reliably extract COVID-19–related information. One of the first steps to extracting this information is developing lexicons with adequate coverage for all synonyms describing each COVID-19 concept. In the clinical domain, lexicons have been developed using several techniques: standardized vocabularies [4], lexico-syntactic patterns [5], term expansion [6], and distributional semantics [7]. Moreover, word embedding technologies have become increasingly popular for identifying semantically and syntactically-related terms within vector spaces by assessing the *distributional hypothesis* that "words that share a common, relative vector space will often also share a common, semantic relatedness" [7].

### Word Embeddings

Word embeddings represent a word in a vector space while preserving its contextualized usage. Word embeddings have been leveraged to learn synonyms to develop lexicons [8]. These vectors are commonly learned by training algorithms like Word2Vec [9], FastText [10], and global vectors for word representation (GloVe) [11] on large corpora, including domain-independent texts (eg, internet web pages like Wikipedia and CommonCrawl, and social media like Twitter and Reddit) and domain-dependent texts (eg, clinical notes like the Medical Information Mart for Intensive Care III [MIMIC III] database notes [12] and biomedical research articles like PubMed). These domain-dependent embeddings may capture richer biomedical information than domain-independent embeddings (eg, *standard GloVe embeddings*) and are becoming increasingly available to the community at large. For example, *BioASQ* released their embeddings trained using the Word2Vec algorithm on 11 million biomedical abstracts from PubMed [13]. Pyysalo et al [14] trained embeddings using Word2Vec on a combination of

PubMed and PubMed Central articles along with Wikipedia to combine open domain and biomedical knowledge (*biomedical natural language processing* [*BioNLP*] corpus). Zhang et al [15] (*BioWordVec* corpus) and Flamholz et al [16] (*ClinicalEmbeddings* corpus) also leveraged PubMed and PubMed Central articles in addition to clinical notes from the MIMIC III to train embeddings using the FastText, GloVe, and Word2Vec algorithms [12].

### Word Embedding Evaluations

Systematic evaluations of word embeddings can be broadly classified into two categories, *intrinsic* and *extrinsic* evaluations. Intrinsic evaluations typically evaluate these word embeddings against human annotations by measuring the similarity or relationship between the queried and returned word pairs. Pakhomov et al [17,18] and Pedersen et al [19] have developed data sets containing pairs of biomedical terms along with their degree of relatedness as rated by human annotators. Furthermore, Pakhomov et al [17] and Hliaoutakis [20] have annotated pairs of medical terms for their semantic similarity. One intrinsic evaluation for validating these human annotations entails computing the Spearman coefficient between word pairs. Others have intrinsically evaluated word embeddings by clustering biomedical terms from the Unified Medical Language System and Ranker [21], and assessing the cluster quality using metrics like the Davies-Bouldin index and the Dunn index. Word embeddings have advanced the state of the art for many intrinsic natural language processing subtasks (ie, reading comprehension [22], natural language inference [23], text summarization [24], vocabulary development [8], and document classification [25]). An extrinsic or summative evaluation of clinical word embeddings can involve evaluating the performance of machine learning models by using word embeddings to complete a biomedical research task or clinical operation such as patient phenotyping [26,27], patient fall prediction [25], and patient hospital readmission prediction [28].

### COVID-19 and Word Embeddings

In recent years, there has been extensive work to leverage biomedical and clinical texts for developing clinical word embeddings to create concept lexicons [29]. For example, clinical word embeddings have been trained to identify drugs [30], substance abuse terms [8], and anatomical locations [16]. More recently, word embeddings have been used to understand the COVID-19 pandemic. For example, Schild et al [31] trained word2vec models for learning terms related to "virus" ("corona," "covid," "wuflu," "coronovirus," "coronavirus") for

understanding the emergence of sinophobic behavior on web communities like Twitter and 4chan's /pol/ facing COVID-19 outbreaks. Klein et al [32] applied pretrained Bidirectional Encoder Representations from Transformers to identify Twitter users with probable or possible COVID-19 infection using their self-reported Twitter messages and temporal-spatial information. However, to our knowledge, there has been no intrinsic evaluation of openly available word embeddings to identify COVID-19 terms related to symptoms, findings, and disorder concepts for encoding clinical notes.

Our long-term goal is to develop a COVID-19 information extraction system to support a variety of purposes, including clinical and translational research, observational studies, clinical trials, public health monitoring, and hospital capacity monitoring. Our short-term goal is to conduct an *intrinsic evaluation* to qualitatively analyze and compare various openly available word embedding sources by categorizing the most similar words returned for symptoms, findings, and disorders related to COVID-19, and to identify common patterns between

returned terms and their associated COVID-19 query terms to better understand which of these word embedding sources and their configurations could support synonym discovery. An additional short term goal is to conduct an *extrinsic evaluation* to apply these terms and their learned synonyms to the discharge summaries of patients with pneumonia, acute respiratory distress syndrome (ARDS), and COVID-19, and report the proportion of patients identified, with terms representing each concept for each disorder cohort.

## Methods

In this University of Pennsylvania Institute Review Board–approved study (#831895, #843620), we conducted a literature review of open-source word embeddings. We identified 7 publicly available sources and characterized each resource according to the training source, unit of processing, context window embedding technology, preprocessing, embedding technology used, returned units, embedding size, and vocabulary size (Table 1).

**Table 1.** Description of word embedding sources used.

| Name | Author and source | Training source | Unit | Context window | Preprocess (reduce case, remove stop words, term types) | Embedding technology (gensim, Fast-Text, GloVe[a], BERT[b], EL-MO, etc) | Returned unit (1-3 ngrams) | Embedding size | Vocab size |
|---|---|---|---|---|---|---|---|---|---|
| BioNLP[c] Lab PubMed + PMC[d] W2V | Pyysalo et al 2013 [14,33] | PubMed/ PMC articles | Token | 5 | Mixed case, no stop words, skip-grams | word2Vec | 1 ngram | 200 | ~4 billion tokens |
| BioNLP Lab-Wiki + PubMed + PMC W2V | Pyysalo et al 2013 [14,33] | Wikipedia, PubMed/ PMC articles | Token | 5 | Mixed case, no stop words, skip-grams | word2Vec | 1 ngram | 200 | ~5.4 billion tokens |
| BioASQ | Tsatsaronis et al 2015 [13,34] | PubMed abstracts | Token | 5 | Lowercase, no stop words, continuous bag of words | word2Vec | 1 ngram | 200 | ~1.7 billion tokens |
| Clinical Embeddings W2V300 | Flamholz et al 2019 [16,35] | PubMed/ PMC/ MIMIC III[e] | Token | 7 | Lowercase, include stop words, skip-grams | word2Vec | 1-3 ngrams | 300 | ~300k tokens |
| BioWordVec Extrinsic | Zhang et al 2019 [15,36] | PubMed + MeSH[f] | Character | 5 | lowercase, include stop words | FastText | 1-3 ngrams | 200 | ~2.3 billion tokens |
| BioWordVec Intrinsic | Zhang et al 2019 [15,36] | PubMed + MeSH | Character | 20 | Lowercase, include stop words | FastText | 1-3 ngrams | 200 | ~2.3 million tokens |
| Standard GloVe Embeddings | Pennington et al 2014 [11] | Common Crawl | Token | 10 | Mixed case | GloVe | 1 ngram | 300 | ~2.1 billion tokens |

[a]GloVe: global vectors for word representation.

[b]BERT: Bidirectional Encoder Representations from Transformers.

[c]BioNLP: biomedical natural language processing.

[d]PMC: PubMed Central.

[e]MIMIC III: Medical Information Mart for Intensive Care III.

[f]MeSH: Medical Subject Headings.

## Constructing the Reference Standard

We generated a list of terms for COVID-19–related semantic categories of *symptoms* ("fever," "high fever," "cough," "wet cough," "dry cough," "congestion," "nasal congestion," "pain," "chest pain," "muscle pain," "shortness of breath," "dyspnea," "tachypnea," "malaise," "headache," "sore throat"), *findings* ("hypoxia," "opacities," "bilateral opacities," "infiltrates," "lung infiltrates"), and *disorders* ("ARDS," "respiratory distress," "acute respiratory distress syndrome," "pneumonia") described in Cascella et al [1]. We queried each word embedding source detailed in Table 1 using these COVID-19–related phrases and retrieved the top 20 phrases based on ranked cosine similarity (terms closest to 1.0 signifying high similarity). Three annotators (a biomedical informatician, a clinical general internist and informatician, and a second-year medical student) encoded each returned phrase with the following semantic class types:

- *Negation (black)*: a negation of the query term (eg, "afebrile" is a negation of "fever")
- *Synonyms (green)*: a lexical variant of the query term with highly similar or synonymous meaning, including misspellings and short forms (eg, "ARDS" is a synonym for "Acute Respiratory Distress Syndrome")
- *Symptom/signs (yellow)*: any symptom, observation, finding, or syndrome that is not a synonym of the query term (eg, "fever" is a symptom returned by "cough")
- *Disease/disorders (blue)*: any disease, disorder, or diagnosis that is not a synonym for the query term (eg, "pneumonia" is a disorder returned by "dyspnea")
- *Hyponym (light red)*: a more specific semantic type of the query term (eg, "ground-glass opacities" is a hyponym of "opacities")
- *Hypernym (dark red)*: a broader semantic type of the query term (eg, "cough" is a hypernym of "productive cough")
- *Qualifiers (teal)*: any nonclinical temporal, spatial, quality, extent, or size descriptor (eg, "dry" is a qualifier for "cough")
- *Anatomical location (orange)*: any clinical anatomical or positional descriptor (eg, "lower lobe" is an anatomical location)
- *Therapeutic (purple)*: any medication, therapy, or procedure (eg, "mechanical ventilation" is a therapeutic device)
- *Other (grey)*: any semantic type that was not among the aforementioned or a nonclinical type (eg, "traffic" returned for "congestion")

## Assessing Interannotator Agreement

For each annotator pair, we computed the interannotator agreement for the semantic class types for each queried term using Cohen kappa [37] using sklearn [38]. Specifically, for each queried phrase (eg, "fever"), each annotator encoded the *semantic type* of the returned candidate term compared to the queried term (eg, returned term "pyrexia" encoded as a *synonym* for queried term "fever"). We report the overall interannotator agreement by category (symptom, finding, and disorder) and by queried term ("fever," "dry cough"). We also depict semantic disagreements between each pair of annotators using heat maps generated using matplotlib [39].

## Analyzing the Similarity Between COVID-19 Queried and Returned Terms

We depict the broad range of terms returned across openly available word embedding sources. For each queried term, the returned term will maintain the same semantic type across word embedding sources but might return a different cosine similarity or occur in only select sources. Therefore, for all unique returned terms within the top 20 ranked by cosine similarity, we visualized the returned term based on its frequency among the word embedding sources at any rank using word clouds generated with matplotlib. The size of the word is a weighted representation of how frequently the returned term occurred across the seven-word embedding source; the score is bounded between 0.14 (observed within only one of seven word embedding sources) and 1.0 (observed within all seven word embedding sources). Additionally, of the terms that occurred *at least once* among the top 20 ranked terms across the seven embeddings, we plotted the range of cosine similarities. Observed top-ranked terms may have cosine similarity values ranging from 0 to 1.0. If a top-ranked term was not found within another embedding source, the term received a value of –1.

## Assessing the Semantic Distribution Patterns for Returned Candidate Terms by Source

We determined the distribution of semantic classes among returned candidates for each queried term according to word embedding source. Our goal is to identify common semantic themes among the queried-returned term pairs that might be driven by the word embedding source construction. We performed a content analysis with simple mean comparisons for each semantic category as well as terms with and without modifiers across embedding sources to identify additional association patterns (Table 2).

**Table 2.** Queried terms (symptoms, findings, and disorders) with and without modifiers.

| Category and term without modifier | With modifier |
| --- | --- |
| **Symptoms** | |
| "fever" | "high fever" |
| "cough" | "wet cough," "dry cough" |
| "congestion" | "nasal congestion" |
| "pain" | "chest pain," "muscle pain" |
| **Findings** | |
| "opacities" | "bilateral opacities" |
| "infiltrates" | "lung infiltrates" |
| **Disorders** | |
| "ARDS"[a] | "respiratory distress," "acute respiratory distress syndrome" |

[a]ARDS: acute respiratory distress syndrome.

## Generating Symptom Severity Profiles for Patients With Pneumonia, ARDS, and COVID-19

As a proof of concept, we compared the proportion of patients that can be classified according to COVID-19 illness severity groups using terms indicative of their clinical features for three cohorts: patients with *pneumonia*, *ARDS*, and *COVID-19*. For the patients with pneumonia and ARDS cohorts, we queried all inpatient encounters and their resulting discharge summaries with COVID-19–related disorders: ARDS (International Classification of Diseases [ICD] codes: 518.5, 518.81, 518.82) and pneumonia (ICD codes: 480-488) from the MIMIC III database [12]. For the patients with COVID-19 cohort, we queried all COVID-19 inpatient encounters from our EPIC PennChart COVID-19 registry from March 2020 to August 2020 and the resulting discharge summaries. In Table 3, we denote the clinical findings associated with COVID-19 respiratory illness severity categories [1]. We applied the expanded lexicon for COVID-19 respiratory illness severity clinical features using synonyms detected from all embedding approaches (*keywords + embedding expansion*). For each cohort, we report the proportion of patients with the clinical feature documented within one or more discharge summaries.

**Table 3.** Clinical findings according to the COVID-19 respiratory illness severity groups.

| COVID-19 respiratory illness severity | Clinical features |
| --- | --- |
| Mild illness | Mild fever, cough (dry), sore throat, malaise, headache, muscle pain, nasal congestion |
| Moderate pneumonia | Cough and shortness of breath |
| Severe pneumonia/acute respiratory distress syndrome | Fever is associated with severe dyspnea, respiratory distress, tachypnea, and hypoxia |

## *Results*

We queried seven embedding sources with 15 symptom terms, five finding terms, and four disorder terms, resulting in 10,080 annotations (top 20 returned candidate terms × 25 queried terms × seven word embedding sources × three annotators).

### Assessing Interannotator Agreement

We observed high overall pairwise interannotator agreement between annotators (ie, A#=Annotator#) for each semantic category: symptoms (0.86-0.99), findings (0.93-0.99), and disorders (0.93-0.99). For A1/A2 and A2/A3, we observed low to moderate interannotator agreement for "malaise" (0.40-0.41), "muscle pain" (0.6), "headache" (0.65-0.68), and "dry cough" (0.68). For A3/A1, interannotator agreement was consistently high (≥0.93). In Figure 1, we report the distribution of each queried term's overall agreement between paired annotators. The color bar represents the third annotator pair. Overall agreement by COVID-19 category and by queried term for each annotator pair can be found in Multimedia Appendix 1.

**Figure 1.** Pairwise interannotator agreement according to semantic category for each queried term.
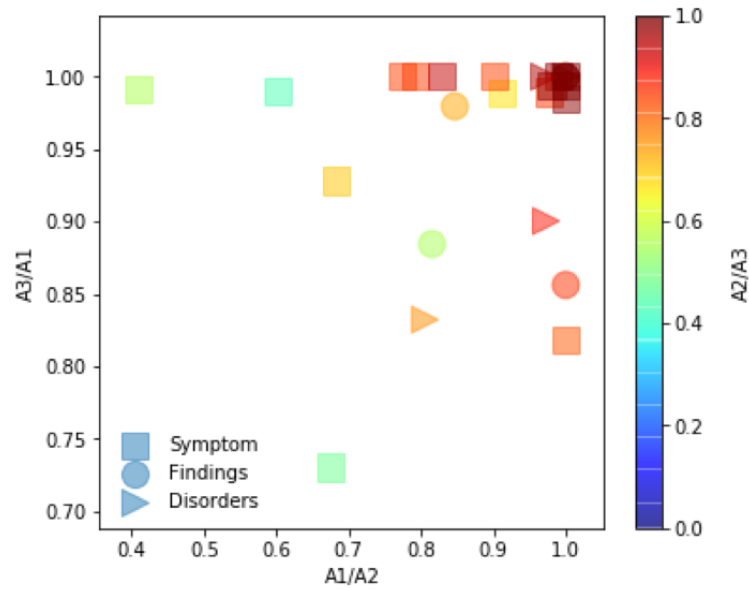


In Figures 2-4, for each returned term, we also computed interannotator agreement across semantic types. Across annotator pairs, we observed high interannotator agreement for all semantic types. Each heat map depicts systematic differences between annotators. In Figure 2, A1/A2 more often disagreed about whether a returned term was a hypernym, hyponym, or negation. In Figure 3, A2/A3 more often disagreed about whether a returned term was a synonym, disease or disorder, hypernym, hyponym, other, or negation. In Figure 4, A3/A1 most often disagreed about whether a returned term was a negation or other term.

**Figure 2.** A1/A2 interannotator agreement of returned terms according to semantic type.
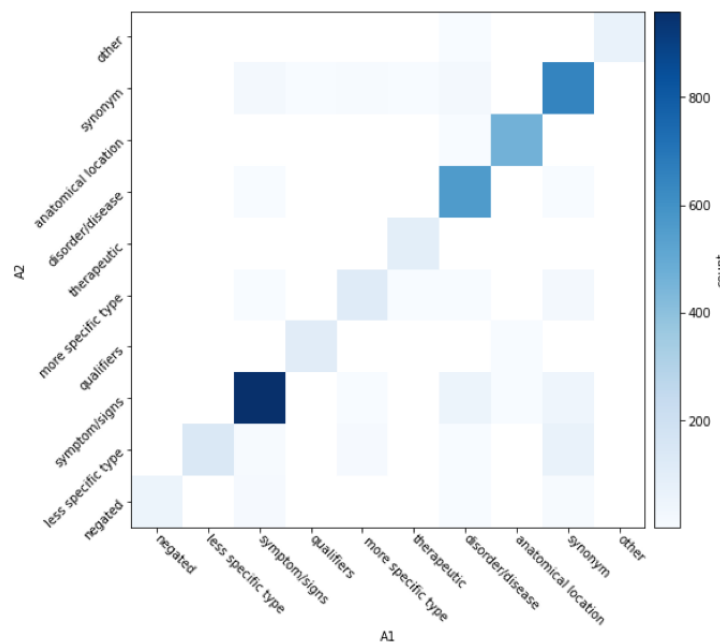
**Figure 3.** A2/A3 interannotator agreement of returned terms according to semantic type.
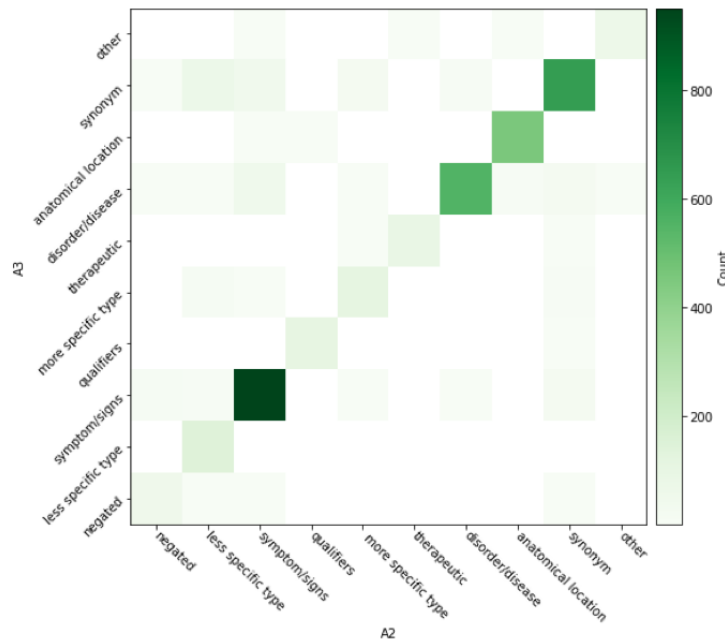


**Figure 4.** A1/A3 interannotator agreement of returned terms by semantic type.
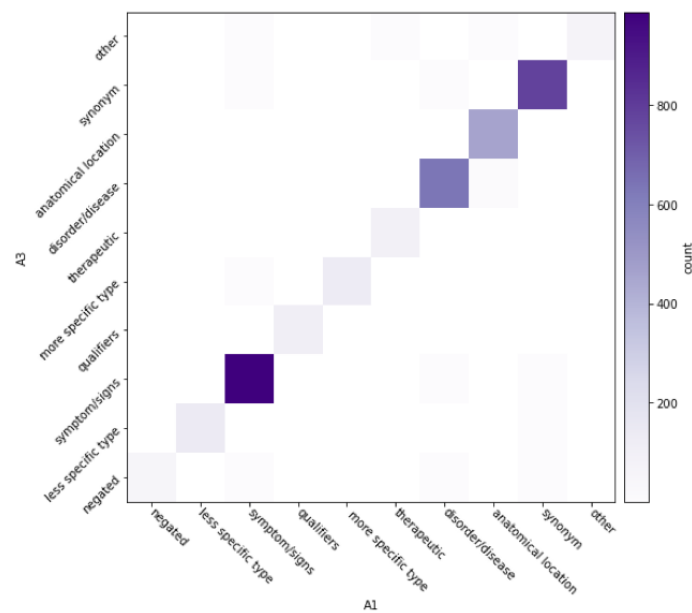


## Analyzing the Similarity for Returned Candidate Terms

We report the broad range of queried terms returned across word embedding sources. For brevity, we depict three COVID-19–related concepts, one of each semantic category: *symptom* ("fever"; Figure 5), *finding* ("lung infiltrates"; Figure 6), and *disorder* ("acute respiratory distress syndrome"; Figure 7). For "fever," synonyms (eg, "pyrexia," "fevers," and "febrile") and signs or symptoms (eg, "chills" and "diarrhea")

were common among the returned terms. For "lung infiltrates," the most frequent semantic types included anatomical locations (eg, "lungs," and "peribronchial") and hypernyms (eg, "infiltrate" and "infiltration") were among the returned terms. For "ARDS," disease or disorders (eg, "SARS" [severe acute respiratory syndrome] and "aSARS-CoV"), synonyms (eg, "ards" and "respiratory-distress-syndrome"), and hypernyms (eg, "syndromee" and "syndrome-critical" were observed commonly among the returned terms.

**Figure 5.** Word cloud depicting each returned term for "fever." Colors correspond to semantic class types.



**Figure 6.** Word cloud depicting each returned term for "lung infiltrates." Colors correspond to semantic class types.



**Figure 7.** Word cloud depicting each returned term for "acute respiratory distress syndrome." Colors correspond to semantic class types. ALI: acute lung injury; ARDS: acute respiratory distress syndrome; ARI: acute respiratory infection; SARS: severe acute respiratory syndrome; SARS-CoV: severe acute respiratory syndrome–related coronavirus.



In Figure 8, we observe that, given a queried term (eg, "fever," "lung infiltrates," and "acute respiratory distress syndrome"), returned terms differ by cosine similarity and variance. For example, some returned terms have high cosine similarity and low variability (left most in red and orange only), while others demonstrate variable cosine similarity and high variability (right most in all colors). Examples of returned terms with *high cosine similarity and low variability* include "fever": "fevers," "fevering," and "pyrexia"; "lung infiltrates": "infiltration," "infiltrates," and "peribronchial"; and "acute respiratory distress syndrome": "syndrome(ARDS)," "aSARS," and "syndromeards." Examples of returned terms with *variable cosine similarity and high variability* include "fever": "fevered," "fever-based," and "fever-like"; "lung infiltrates": "infiltrational," "consolidations," and "bronchioepithelial"; and "acute respiratory distress syndrome": "syndrome-is" and "syndrome-level."

**Figure 8.** Cosine similarity measures for each unique returned term among the top 20 terms across all word embedding sources returned for the queried terms "fever," "lung infiltrates," and "acute respiratory distress syndrome." Color range indicative of cosine similarity level (0.0-1.0), not semantic type.
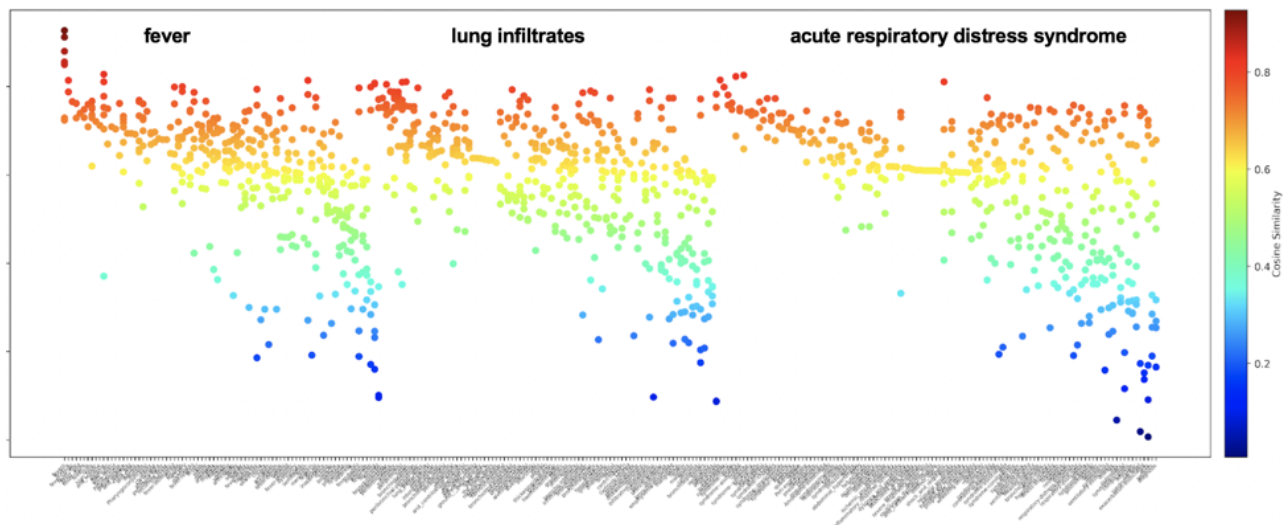


## Assessing the Distribution of Semantic Types for Returned Candidate Terms by Source

We determined the distribution of semantic classes among returned candidates for each queried term according to word embedding source. Our goal is to identify common semantic themes among the queried and returned candidate term pairs that might be driven by the word embedding source construction. We observed that the *BioWordVec Extrinsic* and *BioWordVec Intrinsic* embeddings (Figure 9 e-f) were more likely to generate synonyms (green), which is notably depicted for "fever," "headache," "hypoxia," "dyspnea," and "infiltrates." Word embedding sources generated based on characters tend to return more synonyms (mean count of 7.2 synonyms) compared to token-based embedding sources (mean count ranged from 2.04 to 3.4 synonyms). We also observed more negation terms for "hypoxia" (mean count of 2.29 negations); "congestion" (mean count of 1.57 negations); and "dry cough," "wet cough," and "tachypnea" (all had mean counts of 1.0 negations) compared to other terms (mean counts ranged from 0.00 to 0.71 negations). We observed a high mean count of hypernyms for "dry cough" (mean count of 6.43 hypernyms), "high fever" (mean count of 5.57 hypernyms), and "acute respiratory distress syndrome" (mean count of 4.43 hypernyms) over other terms (mean counts ranged from 0 to 3.29 hypernyms). Across the other word embeddings (Figure 9 a-d and g), if a symptom or sign queried term was provided, we more often observed a symptom or sign returned term (mean average of 6.62 symptoms or signs) compared to nonsymptom or sign queried terms (mean average

of 3.035 symptoms or signs). This also held true for disorders (mean average of 6.24 disorders) compared to nondisorders (mean average of 1.18 disorders). Across word embedding sources (Figure 9), we observed that qualifiers were more often returned when the queried term contained a qualifier for some terms (eg, "dry cough" and "wet cough" return time and consistency qualifiers like "wet" and "runny"; both mean counts of 4.14 qualified terms) over the nonqualified queried term "cough" (mean count of 1.71 qualified terms). Similar patterns were observed for "high fever" (mean count of 3.71 qualified terms), "fever" (mean count of 0.0 qualified terms), "bilateral opacities" (mean count of 6.14 qualified terms), and "opacities" (mean count of 2.71 qualified terms). Furthermore, if a queried term contained an anatomical location as an adjective in the term phrase (eg, "nasal congestion"), the returned terms were often anatomical locations compared to queried terms without adjectives. We observed notable differences in mean counts of returned terms with anatomical qualifiers for "nasal congestion" (mean count of 6.71 anatomical terms) and "congestion" (mean count of 0.42 anatomical terms), "chest pain" (mean count of 8.43 anatomical terms) and "pain" (mean count of 3.57 anatomical terms), and "lung infiltrates" (mean count of 10.57 anatomical terms) and "infiltrates" (mean count of 6.71 anatomical terms). In few cases, the *standard GloVe embeddings*, *BioWordVec Extrinsic*, and *BioWordVec Intrinsic* embeddings returned some terms with common term usage (eg, "congestion" returns "traffic," "bypass," or stop words such as "and," "a," and "of").

**Figure 9.** For each symptom, finding, and disorder queried term, the distribution of semantic types for returned term colored by semantic type for each embedding source: (a) BioNLP Lab PubMed + PMC W2V, (b) BioNLP LabWiki + PubMed + PMC W2V, (c) BioASQ, (d) Clinical Embeddings W2V300, (e) BioWordVec Extrinsic, (f) BioWordVec Intrinsic, and (g) Standard GloVe Embeddings. ARDS: acute respiratory distress syndrome.
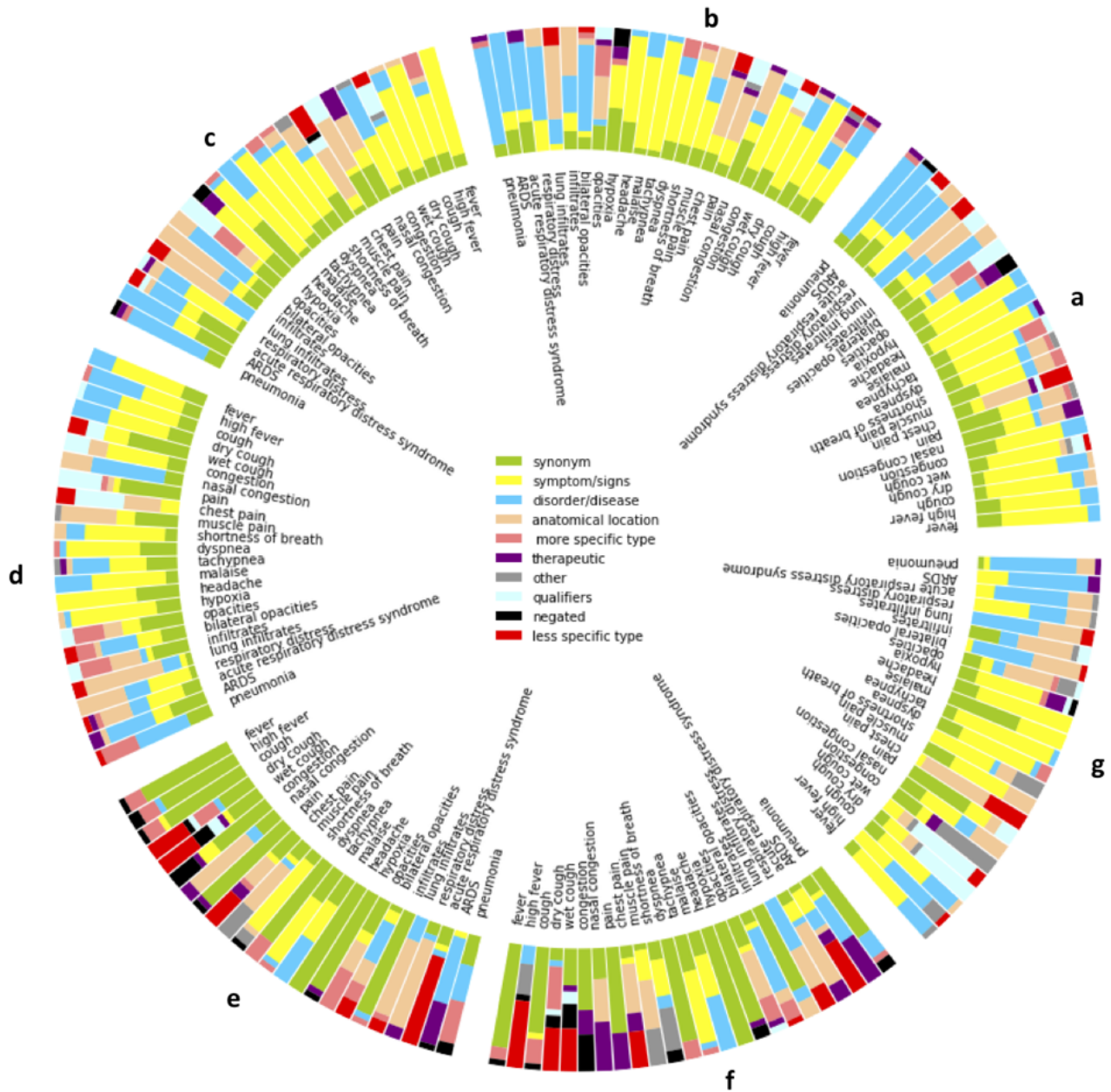


## Generating Symptom Severity Profiles for Patients With Pneumonia, ARDS, and COVID-19

Figure 10 shows the proportion of patients from each disorder cohort (pneumonia, ARDS, and COVID-19) that have one or more terms documented within their discharge summary representing clinical features from Table 3. The total number of patients in each cohort varied: pneumonia (n=6410), ARDS (n=8647), and COVID-19 (n=2397). A higher proportion of patients had documented fever (0.61-0.84), cough (0.41-0.55), shortness of breath (0.40-0.59), and hypoxia (0.51-0.56) retrieved than other clinical features. Terms for dry cough returned a higher proportion of patients with COVID-19 (0.07) than pneumonia (0.05) and ARDS (0.03) populations.

**Figure 10.** The proportion of patients with each COVID-19 clinical feature documented within their discharge summary according to disorders (pneumonia, ARDS, and COVID-19). ARDS: acute respiratory distress syndrome.
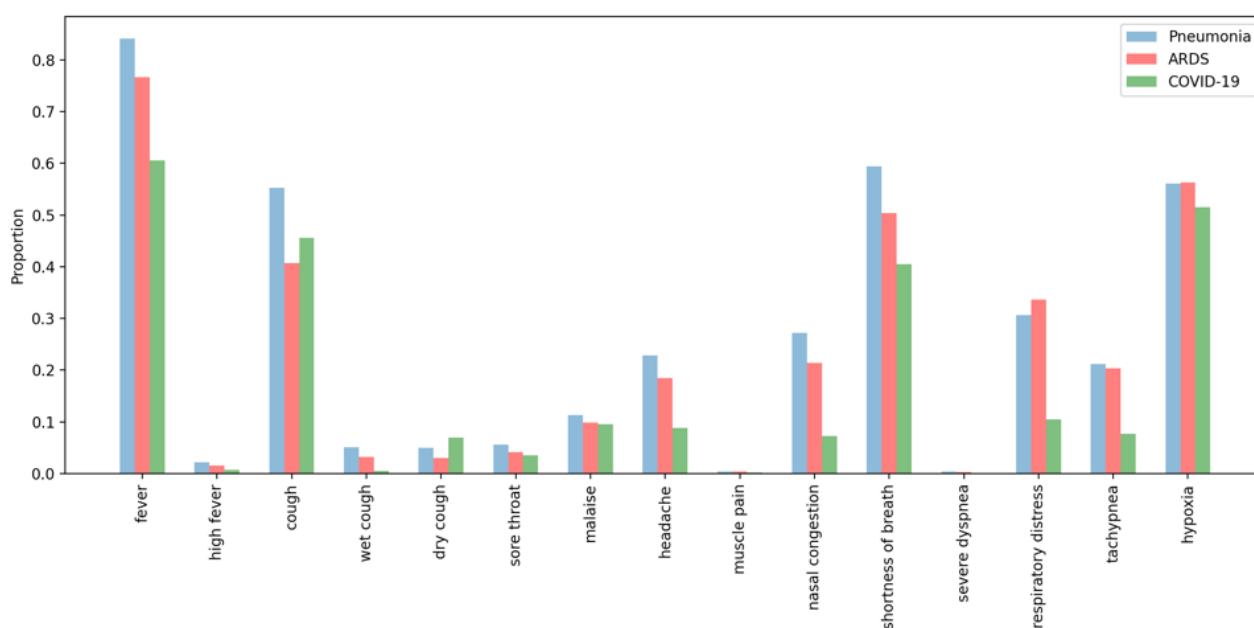


## Discussion

### Assessing Interannotator Agreement

We observed high overall pairwise interannotator agreement for the symptoms, findings, and disorder categories. Annotators A1 and A3 were more often in agreement. For the A1/A2 and A2/A3 pairs, we observed low to moderate interannotator agreement for queried terms such as "malaise," "muscle pain," "headache," and "dry cough." Annotators A1 and A3 systematically classified notably fewer returned terms as hypernyms and hyponyms than A2. For example, "migraine" is a hypernym for "headache." Additionally, A2 more easily identified negated terms through medical terminology. Many cases required more clinical domain knowledge to make these distinctions, which were easier for the general internist.

### Analyzing the Similarity Between COVID-19 Queried and Returned Terms

When analyzing the cosine similarities between queried terms and returned terms, we observed that returned terms range from *high cosine similarity and low variability* to *variable cosine similarity and high variability*. We hypothesize that terms with high cosine similarity and low variability are more likely to be synonyms and useful for training an information extraction. In practice, the presence and cosine similarities of a term varied across word embedding sources. Our ability to identify and rank likely synonyms for lexicon development may be improved with additional processing steps and comparisons between the queried and returned terms for lexical similarity [40], morphological derivation [8], and short form construction and expansion [41].

### Assessing the Semantic Distribution Patterns for Returned Candidate Terms by Source

We determined the distribution of semantic classes among returned candidates for each queried term according to the word embedding source. Our study intentions were to assess the distributional hypothesis that words with similar meanings are often used in similar contexts. Generally, if a symptom or sign queried term was provided, we often observed a symptom or sign returned term. This also held true for disorders. Furthermore, our goal was to identify common semantic themes among the queried and returned candidate term pairs that might be driven by the word embedding source construction. We observed that the *BioWordVec Extrinsic* and *BioWordVec Intrinsic* embeddings were more likely to generate synonyms. We hypothesize that this is likely due to training based on the characters rather than the token; thus, the returned terms often share a common set of characters (queried term: "*fever*"; returned term: "*fever*ish") or high lexical similarity. Character-based embeddings will often return lexical variations of the queried term. Although *BioNLP*, *BioASQ*, and *Clinical Embeddings* generated fewer synonyms, these were often medical terms for the lay queried term (eg, "lethargy" for "malaise," "cephalea" for "headache," and "rhinorrhea" for "nasal congestion"). To maximize the diversity of learned synonyms, multiple embeddings could be most beneficial. Returned negated terms were expressed with prefixes (eg, "non-pneumonia-related"), suffixes (eg, "fever-free"), or medical terminology (eg, "normoxia"). Hypernyms were commonly observed among queried terms with an adjectival phrase (eg, "high fever," "muscle pain," "dry cough," and "lung infiltrates"). Moreover, we observed that qualifiers were often returned when the queried term contained a qualifier (eg, time, consistency, and anatomical location qualifiers). For developing a clinical information extraction system, these returned terms can be useful for brainstorming synonyms as inclusionary terms as well as antonyms as exclusionary terms. We suspect that a mix of hypernyms and qualifiers were often returned, given the semantics of the individual parts of the queried phrase. It was not surprising that *standard GloVe* embeddings returned some terms with a nonclinical word sense (eg, "congestion" returns

"traffic" or "bypass") because they were trained using the CommonCrawl domain-independent corpora. Similarly, *BioWordVec Extrinsic* and *BioWordVec Intrinsic* occasionally return stop words, as these were not removed prior to training and perhaps should be for detecting meaningful synonyms.

## Generating Symptom Severity Profiles for Patients With Pneumonia, ARDS, and COVID-19

We created an expanded lexicon of COVID-19 respiratory illness clinical features (Table 3) using synonyms detected from all embedding approaches. We assessed the proportion of patients from three disorder cohorts (pneumonia, ARDS, and COVID-19) with each clinical feature documented within their discharge summary. We observed that terms indicative of clinical features for fever, cough, shortness of breath, and hypoxia retrieved a higher proportion of patients than clinical features. For fever and cough, our lexicons for capturing contextualized mentions of these clinical features (eg, high fever or wet or dry cough) retrieved modest proportions of patient cases. This is likely due to the variability of qualitative and quantifications of these symptoms (eg, productive cough and fever of 102° F) in discharge summaries. Terms indicative of dry cough returned a higher proportion of patients with COVID-19 than pneumonia and ARDS populations. This is not surprising given that this is a prominent symptom reported among patients with COVID-19.

## Limitations and Future Work

Our study has a few notable limitations. We began this study during the early stages of the COVID-19 pandemic when the symptomatology was less understood. COVID-19 is a heterogeneous disease with emerging symptomatology identified through ongoing clinical observational studies. Emerging

COVID-19–related symptomatology (ie, loss of smell, loss of taste, and COVID toes) were not included in our analysis, as their association with COVID-19 were not well understood at the time of our study. We leveraged existing word embedding sources to better understand the utility of embeddings for synonym generation. We recognize that further experimentation is needed to support broader claims of their utility. As a proof of concept of patient information retrieval, we applied an expanded lexicon of terms representing clinical features of COVID-19 to three disorder cohorts (pneumonia, ARDS, and COVID-19). Although these terms retrieved a high proportion of patients, we acknowledge that additional terms might be necessary to accurately identify these features and that contextualization (ie, negation, severity, experiencer, and temporality [42-44]) is critical to generating accurate patient profiles. We look forward to addressing these issues as next steps within our clinical information extraction pipeline powered by Linguamatics [45]. These text-derived and contextualized variables will be available through our clinical research databases—COVID-19 Informatics for Integrating Biology and the Bedside database [46] and Penn Genotype and Phenotype database supported by the Observational Medical Outcomes Partnership common data model [47]—at the end of Spring 2021.

## Conclusion

Word embeddings are a valuable technology for learning semantically and syntactically related terms including synonyms and useful for text classification and information extraction tasks. When leveraging openly available word embedding sources, choices made in the development of the embeddings can significantly influence the types of phrases and information learned.

## Conflicts of Interest

None declared.

Multimedia Appendix 1
Overall agreement by COVID-19 category and by queried term for each annotator pair.
[DOCX File , 15 KB - medinform_v9i2e21679_app1.docx ]

## References

1.  Cascella M, Rajnik M, Cuomo A, Dulebohn SC, Di Napoli E. Features, Evaluation, and Treatment of Coronavirus. Treasure Island, FL: StatPearls Publishing; 2020.
2.  Moore JH, Barnett I, Boland MR, Chen Y, Demiris G, Gonzalez-Hernandez G, et al. Ideas for how informaticians can get involved with COVID-19 research. BioData Min 2020;13:3 [FREE Full text] [doi: 10.1186/s13040-020-00213-y] [Medline: 32419848]

3.    Chapman A, Peterson K, Turano A, Box T, Wallace K, Jones M. A natural language processing system for national COVID-19 surveillance in the US Department of Veterans Affairs. 2020 Presented at: 1st Workshop on NLP for COVID-19 at ACL 2020; July 2020; Online URL: https://www.aclweb.org/anthology/2020.nlpcovid19-acl.10

4.    Castro VM, Dligach D, Finan S, Yu S, Can A, Abd-El-Barr M, et al. Large-scale identification of patients with cerebral aneurysms using natural language processing. Neurology 2017 Jan 10;88(2):164-168 [FREE Full text] [doi: 10.1212/WNL.0000000000003490] [Medline: 27927935]

5.    Liu K, Chapman WW, Savova G, Chute CG, Sioutos N, Crowley RS. Effectiveness of lexico-syntactic pattern matching for ontology enrichment with clinical documents. Methods Inf Med 2011;50(5):397-407 [FREE Full text] [doi: 10.3414/ME10-01-0020] [Medline: 21057720]

6.    Zeng QT, Redd D, Rindflesch T, Nebeker J. Synonym, topic model and predicate-based query expansion for retrieving clinical documents. AMIA Annu Symp Proc 2012;2012:1050-1059 [FREE Full text] [Medline: 23304381]

7.    Henriksson A, Conway M, Duneld M, Chapman WW. Identifying synonymy between SNOMED clinical terms of varying length using distributional analysis of electronic health records. AMIA Annu Symp Proc 2013;2013:600-609 [FREE Full text] [Medline: 24551362]

8.    Velupillai S, Mowery DL, Conway M, Hurdle J, Kious B. Vocabulary development to support information extraction of substance abuse from psychiatry notes. 2016 Presented at: The 15th Workshop on Biomedical Natural Language Processing; August 2016; Berlin, Germany. [doi: 10.18653/v1/w16-2912]

9.    Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. Red Hook, NY: Curran Associates Inc; 2013:3111-31119.

10.   Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Computational Linguistics 2017 Dec;5:135-146. [doi: 10.1162/tacl_a_00051]

11.   Pennington J, Socher R, Manning C. GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014 Presented at: EMNLP '14; October 2014; Doha, Qatar p. 1532-1543. [doi: 10.3115/v1/d14-1162]

12.   Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035. [doi: 10.1038/sdata.2016.35] [Medline: 27219127]

13.   BioASQ releases continuous space word vectors obtained by applying Word2Vec to PubMed abstracts. BioASQ. URL: http://bioasq.org/news/bioasq-releases-continuous-space-word-vectors-obtained-applying-word2vec-pubmed-abstracts [accessed 2020-04-18]

14.   Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou T. Distributional semantics resources for biomedical text processing. In: Proceedings of the 5th International Symposium on Languages in Biology and Medicine. 2013 Presented at: LBM '13; December 12-13, 2013; Tokyo, Japan.

15.   Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci Data 2019 May 10;6(1):52. [doi: 10.1038/s41597-019-0055-0] [Medline: 31076572]

16.   Flamholz ZN, Ungar LH, Weissman GE. Word embeddings trained on published case reports are lightweight, effective for clinical tasks, and free of protected health information. medRxiv. Preprint posted online December 4, 2019. [doi: 10.1101/19013268]

17.   Pakhomov S, McInnes B, Adam T, Liu Y, Pedersen T, Melton G. Semantic similarity and relatedness between clinical terms: an experimental study. AMIA Annu Symp Proc 2010 Nov 13;2010:572-576 [FREE Full text] [Medline: 21347043]

18.   Pakhomov SVS, Pedersen T, McInnes B, Melton GB, Ruggieri A, Chute CG. Towards a framework for developing semantic relatedness reference standards. J Biomed Inform 2011 Apr;44(2):251-265 [FREE Full text] [doi: 10.1016/j.jbi.2010.10.004] [Medline: 21044697]

19.   Pedersen T, Pakhomov SVS, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 2007 Jun;40(3):288-299 [FREE Full text] [doi: 10.1016/j.jbi.2006.06.004] [Medline: 16875881]

20.   Hliaoutakis A. Semantic similarity measures in MeSH ontology and their application to information retrieval on Medline. InteLLigence. 2005. URL: http://www.intelligence.tuc.gr/publications/Hliautakis.pdf [accessed 2021-02-15]

21.   Ranker. URL: https://www.ranker.com [accessed 2020-04-26]

22.   Dhingra B, Liu H, Salakhutdinov R, Cohen WW. A comparative study of word embeddings for reading comprehension. arXiv. Posted on March 2, 2017 2021 [FREE Full text]

23.   Nishida K, Nishida K, Asano H, Tomita J. Natural language inference with definition embedding considering context on the fly. 2018 Presented at: The Third Workshop on Representation Learning for NLP; July 2018; Melbourne, Australia. [doi: 10.18653/v1/w18-3007]

24.   Mohd M, Jan R, Shah M. Text document summarization using word embedding. Expert Syst Applications 2020 Apr;143:112958. [doi: 10.1016/j.eswa.2019.112958]

25.   Topaz M, Murga L, Gaddis KM, McDonald MV, Bar-Bachar O, Goldberg Y, et al. Mining fall-related information in clinical notes: comparison of rule-based and novel word embedding-based machine learning approaches. J Biomed Inform 2019 Feb;90:103103 [FREE Full text] [doi: 10.1016/j.jbi.2019.103103] [Medline: 30639392]

26. Bai T, Chanda AK, Egleston BL, Vucetic S. EHR phenotyping via jointly embedding medical concepts and words into a unified vector space. BMC Med Inform Decis Mak 2018 Dec 12;18(Suppl 4):123 [FREE Full text] [doi: 10.1186/s12911-018-0672-0] [Medline: 30537974]

27. Yang Z, Dehmer M, Yli-Harja O, Emmert-Streib F. Combining deep learning with token selection for patient phenotyping from electronic health records. Sci Rep 2020 Jan 29;10(1):1432. [doi: 10.1038/s41598-020-58178-1] [Medline: 31996705]

28. Liu W, Singh K, Ryan AM, Sukul D, Mahmoudi E, Waljee AK, et al. Predicting 30-day hospital readmissions using artificial neural networks with medical code embedding. bioRxiv. Preprint posted on August 20, 2019. [doi: 10.1101/741504]

29. Khattak FK, Jeblee S, Pou-Prom C, Abdalla M, Meaney C, Rudzicz F. A survey of word embeddings for clinical text. J Biomed Inform 2019 Dec;4:100057. [doi: 10.1016/j.yjbinx.2019.100057]

30. Zhao M, Masino AJ, Yang CC. A framework for developing and evaluating word embeddings of drug-named entity. 2018 Presented at: The BioNLP 2018 Workshop; July 2018; Melbourne, Australia. [doi: 10.18653/v1/w18-2319]

31. Schild L, Ling C, Blackburn J, Stringhini G, Zhang Y, Zannettou S. "Go eat a bat, Chang!": an early look on the emergence of sinophobic behavior on web communities in the face of COVID-19. arXiv. Preprint posted on April 8, 2020 [FREE Full text]

32. Klein A, Magge A, O'Connor K, Cai H, Weissenbacher D, Gonzalez-Hernandez G. A chronological and geographical analysis of personal reports of COVID-19 on Twitter. medRxiv. Preprint posted on April 22, 2020. [doi: 10.1101/2020.04.19.20069948] [Medline: 32511608]

33. Biomedical Natural Language Processing. URL: https://bio.nlplab.org/ [accessed 2021-01-26]

34. Tsatsaronis G, Balikas G, Malakasiotis P, Partalas I, Zschunke M, Alvers MR, et al. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics 2015 Apr 30;16:138 [FREE Full text] [doi: 10.1186/s12859-015-0564-6] [Medline: 25925131]

35. Weissman G. gweissman / clinical_embeddings. GitHub. URL: https://github.com/gweissman/clinical_embeddings [accessed 2021-01-25]

36. ncbi-nlp / BioWordVec. GitHub. URL: https://github.com/ncbi-nlp/BioWordVec [accessed 2021-01-25]

37. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]

38. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Machine Learning Res 2011;12:2825-2830.

39. Hunter JD. Matplotlib: A 2D Graphics Environment. Computing Sci Eng 2007 May;9(3):90-95. [doi: 10.1109/mcse.2007.55]

40. Sarker A, Gonzalez-Hernandez G. An unsupervised and customizable misspelling generator for mining noisy health-related text sources. J Biomed Inform 2018 Dec;88:98-107 [FREE Full text] [doi: 10.1016/j.jbi.2018.11.007] [Medline: 30445220]

41. Henriksson A, Moen H, Skeppstedt M, Daudaravičius V, Duneld M. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. J Biomed Semantics 2014 Feb 05;5(1):6 [FREE Full text] [doi: 10.1186/2041-1480-5-6] [Medline: 24499679]

42. Mowery DL, Jordan P, Wiebe J, Harkema H, Dowling J, Chapman WW. Semantic annotation of clinical events for generating a problem list. AMIA Annu Symp Proc 2013;2013:1032-1041 [FREE Full text] [Medline: 24551392]

43. Hong J, Davoudi A, Yu S, Mowery DL. Annotation and extraction of age and temporally-related events from clinical histories. BMC Med Inform Decis Mak 2020 Dec 30;20(Suppl 11):338 [FREE Full text] [doi: 10.1186/s12911-020-01333-5] [Medline: 33380319]

44. Mowery DL, Kawamoto K, Bradshaw R, Kohlmann W, Schiffman JD, Weir C, et al. Determining onset for familial breast and colorectal cancer from family history comments in the electronic health record. AMIA Jt Summits Transl Sci Proc 2019;2019:173-181 [FREE Full text] [Medline: 31258969]

45. Linguamatics. URL: https://www.linguamatics.com/ [accessed 2021-01-26]

46. i2b2. URL: https://www.i2b2.org/software/index.html [accessed 2021-01-26]

47. OMOP common data model. Observational Health Data Sciences and Informatics. URL: https://www.ohdsi.org/data-standardization/the-common-data-model/ [accessed 2021-01-26]

48. semantica-NLP / COVID-19_embeddings. GitHub. URL: https://github.com/semantica-NLP/COVID-19_embeddings [accessed 2021-02-15]

## Abbreviations

**ARDS:** acute respiratory distress syndrome
**BioNLP:** biomedical natural language processing
**GloVe:** global vectors for word representation
**ICD:** International Classification of Diseases
**MIMIC III:** Medical Information Mart for Intensive Care III
**SARS:** severe acute respiratory syndrome

XSL·FO
**RenderX**

Original Paper

# Rapid Response to Drive COVID-19 Research in a Learning Health Care System: Rationale and Design of the Houston Methodist COVID-19 Surveillance and Outcomes Registry (CURATOR)

Farhaan Vahidy[1], PhD, MBBS, MPH; Stephen L Jones[1], MD, MSHI; Mauricio E Tano[1], PhD; Juan Carlos Nicolas[1], BS; Osman A Khan[1], BS; Jennifer R Meeks[1], MS; Alan P Pan[1], MPH; Terri Menser[1], PhD; Farzan Sasangohar[1], PhD; George Naufal[1], PhD; Dirk Sostman[1], MD; Khurram Nasir[1], MD, MPH; Bita A Kash[1], PhD, MBA

Houston Methodist, Houston, TX, United States

**Corresponding Author:**
Farhaan Vahidy, PhD, MBBS, MPH
Houston Methodist
7550 Greenbriar Drive
Houston, TX, 77030
United States
Phone: 1 3463561479
Email: fvahidy@houstonmethodist.org

## *Abstract*

**Background:** The COVID-19 pandemic has exacerbated the challenges of meaningful health care digitization. The need for rapid yet validated decision-making requires robust data infrastructure. Organizations with a focus on learning health care (LHC) systems tend to adapt better to rapidly evolving data needs. Few studies have demonstrated a successful implementation of data digitization principles in an LHC context across health care systems during the COVID-19 pandemic.

**Objective:** We share our experience and provide a framework for assembling and organizing multidisciplinary resources, structuring and regulating research needs, and developing a single source of truth (SSoT) for COVID-19 research by applying fundamental principles of health care digitization, in the context of LHC systems across a complex health care organization.

**Methods:** Houston Methodist (HM) comprises eight tertiary care hospitals and an expansive primary care network across Greater Houston, Texas. During the early phase of the pandemic, institutional leadership envisioned the need to streamline COVID-19 research and established the retrospective research task force (RRTF). We describe an account of the structure, functioning, and productivity of the RRTF. We further elucidate the technical and structural details of a comprehensive data repository—the HM COVID-19 Surveillance and Outcomes Registry (CURATOR). We particularly highlight how CURATOR conforms to standard health care digitization principles in the LHC context.

**Results:** The HM COVID-19 RRTF comprises expertise in epidemiology, health systems, clinical domains, data sciences, information technology, and research regulation. The RRTF initially convened in March 2020 to prioritize and streamline COVID-19 observational research; to date, it has reviewed over 60 protocols and made recommendations to the institutional review board (IRB). The RRTF also established the charter for CURATOR, which in itself was IRB-approved in April 2020. CURATOR is a relational structured query language database that is directly populated with data from electronic health records, via largely automated extract, transform, and load procedures. The CURATOR design enables longitudinal tracking of COVID-19 cases and controls before and after COVID-19 testing. CURATOR has been set up following the SSoT principle and is harmonized across other COVID-19 data sources. CURATOR eliminates data silos by leveraging unique and disparate big data sources for COVID-19 research and provides a platform to capitalize on institutional investment in cloud computing. It currently hosts deeply phenotyped sociodemographic, clinical, and outcomes data of approximately 200,000 individuals tested for COVID-19. It supports more than 30 IRB-approved protocols across several clinical domains and has generated numerous publications from its core and associated data sources.

**Conclusions:** A data-driven decision-making strategy is paramount to the success of health care organizations. Investment in cross-disciplinary expertise, health care technology, and leadership commitment are key ingredients to foster an LHC system.

XSL•FO
**RenderX**

Such systems can mitigate the effects of ongoing and future health care catastrophes by providing timely and validated decision support.

## Introduction

As of December 31, 2020, over 90 million COVID-19 cases had been confirmed worldwide [1]. The COVID-19 pandemic has tested the limits of human resilience, leading to innovation in several facets of clinical and academic medicine [2,3]. Prior to the pandemic, the health care industry had already been on the precipice of a digital revolution driven by big data, machine learning, and artificial intelligence for a long time. The pandemic brought to bear a dire need for investment in robust health data infrastructures and pipelines (DIPs) such that barriers and latency to gather, assimilate, validate, and share data widely and swiftly can be minimized or eliminated [4]. Establishing and maintaining robust clinical DIPs are resource intensive and require a cross-disciplinary approach. Effective utilization of health care data to drive clinical and operational decision-making, in the context of a true learning health care (LHC) system, warrants organizational commitment—both at the technical level and as a behavioral paradigm shift.

For several health care organizations, the urgency to synthesize epidemiological and clinical evidence for understanding the rapidly evolving COVID-19 pandemic has underscored the need for innovation in two sperate yet overlapping processes: (1) the review process for approval of COVID-19–related minimal risk research while maintaining stringent federal and institutional standards of human-subject research and (2) the critical and fundamental need to establish a reliable and valid DIP to serve as the backbone for swift and accurate reporting. Organizations with an LHC focus and infrastructural investment are highly likely to be agile and adaptive to such rapidly developing needs and thus be on the forefront of combating health care catastrophes.

This paper provides an overarching account of how the needs for data accessibility, rapid research, and reliable reporting evolved in the face of the COVID-19 pandemic across a large health care system and its associated research enterprise. Both the health care system and research enterprise are located in a very populous and diverse US metropolis (Houston, Texas) that became a hub of the second wave of the COVID-19 pandemic during the summer of 2020. We share our experiences of the methodology implemented for addressing the aforementioned needs, which included (1) assembling and leveraging expertise from interdisciplinary and multispecialty teams; (2) listing considerations that include regulation and ethics of COVID-19 research; (3) leveraging organizational aspects of coordinating and harmonizing cross-institutional data and research needs; and finally, (4) the development, technical design, and implementation of the Houston Methodist COVID-19 Surveillance and Outcomes Registry (CURATOR). All these items are in line with the health care system's institutional goal of fostering a true LHC.

## Methods

### Implementation Setting

#### Greater Houston Metropolitan Area and the Houston Methodist System

Like other large metropolitan areas across the United States, the Greater Houston area experienced a rise in COVID-19 cases in early March 2020. The Greater Houston Metropolitan Statistical Area—officially designated by the Office of Budget and Management as "Houston–The Woodlands–Sugar Land"—is the fifth most populous area in the United States, with an approximate population of 7 million [5]. The Greater Houston area is also considered to be one of the nation's most ethnically diverse regions [6]. Harris County, whose county seat is Houston, is the third largest county in the United States in terms of population, whereas the city of Houston is the fourth most populated US city [5]. On March 1, 2020, there was only 1 known and officially reported case of COVID-19 in the Harris County/Houston area, which increased to almost 6000 over an 8-week period. The first surge, which peaked in mid-April, saw later a 3- to 4-fold increase in cases by early July 2020 [7]. The total number of COVID-19 cases in the 9-county Houston Metropolitan Statistical Area is estimated to be over 315,000, as of December 31, 2020 [8].

Houston Methodist, along with its centers of excellence in cancer, heart and vascular, digestive disorders, neurology, orthopedics and sports medicine, and transplant, and an academic affiliation with Weill Cornell Medicine and New York Presbyterial Hospital (New York, USA), comprises one flagship tertiary care hospital (Houston Methodist Hospital) and six large community hospitals, with an additional long-term care hospital, spanning across the Greater Houston area. Additionally, the system has an expansive emergency medicine and ambulatory health care network including an Accountable Care Organization [9]. Houston Methodist Research Institute and Houston Methodist Academic Institute lead the basic science, translational, clinical and epidemiological outcomes research, and training portfolio for the system [9,10].

#### COVID-19 Clinical, Administrative, and Research Data Needs at Houston Methodist

Houston Methodist became the clinical hub for COVID-19 in the Greater Houston area and the first in the United States to perform plasma transfusion as part of COVID-19 treatment [11]. As soon as Houston Methodist started testing for COVID-19 and providing care to infected patients, the urgent need for validated, ongoing data on COVID-19 treatment and

outcomes mandated institutional prioritization. Data requirements came from three broad categories of stakeholders. First, the frontline care teams needed data to support the clinical decision-making process; second, hospital administration and leadership needed data to efficiently manage hospitals' resources and outwardly communicate to the public; and third, clinical researchers needed data to explore innumerable important research questions. Anecdotal information on potentially beneficial therapies and effective management algorithms started flowing in, and there was a dire need to "validate" treatment efficacies and management modalities in the local context. Administrators needed quick and reliable metrics on not only the number of COVID-19 cases but also precise projections on mortality rates, length of stay, days in intensive care units (ICUs), and utilization of critical hospital resources such as ventilators and personal protective equipment. In addition, several centers of excellence and clinical departments immediately needed access to data of patients with COVID-19 to analyze important disease patterns and consequences on their respective patient populations. Consequently, there was an overwhelming outpouring of proposals and research ideas that started flowing to the Institutional Review Board (IRB). To provide rapid responses while preserving research integrity, two system-wide subcommittees were established: the Clinical Trials Task Force, which was tasked to evaluate proposals for therapeutic clinical trials, and the Retrospective Research Task Force (RRTF), which was set in place to facilitate the review and coordination of all observational (retrospective and prospective) research across the system. Many authors of this manuscript (FSV, HDS, BAK, SLJ, KN, and JRM, along with representation from the IRB and corporate and research information technology departments) constituted the membership of the RRTF.

## Current Implementation of Electronic Health Record System

The past decade has seen a dramatically increased propagation of electronic health records (EHRs) in the United States. This phenomenon was largely promoted through large US government-initiated programs to encourage the adoption of EHRs in routine practice (eg, Meaningful Use, Certification Commission for Health Information Technology; inducements in the Affordable Care Act; Health Information Technology for Economic and Clinical Health Act in 2009; and the mandatory submission of quality measures electronically). Although certain benefits of EHRs are undeniable, they are most often designed and implemented with the administrative end-user in mind. In most cases, a system with a focus on administration, with streamlined billing and coding features, is not adapted for the assimilation of research data. Coincidentally, this same system conta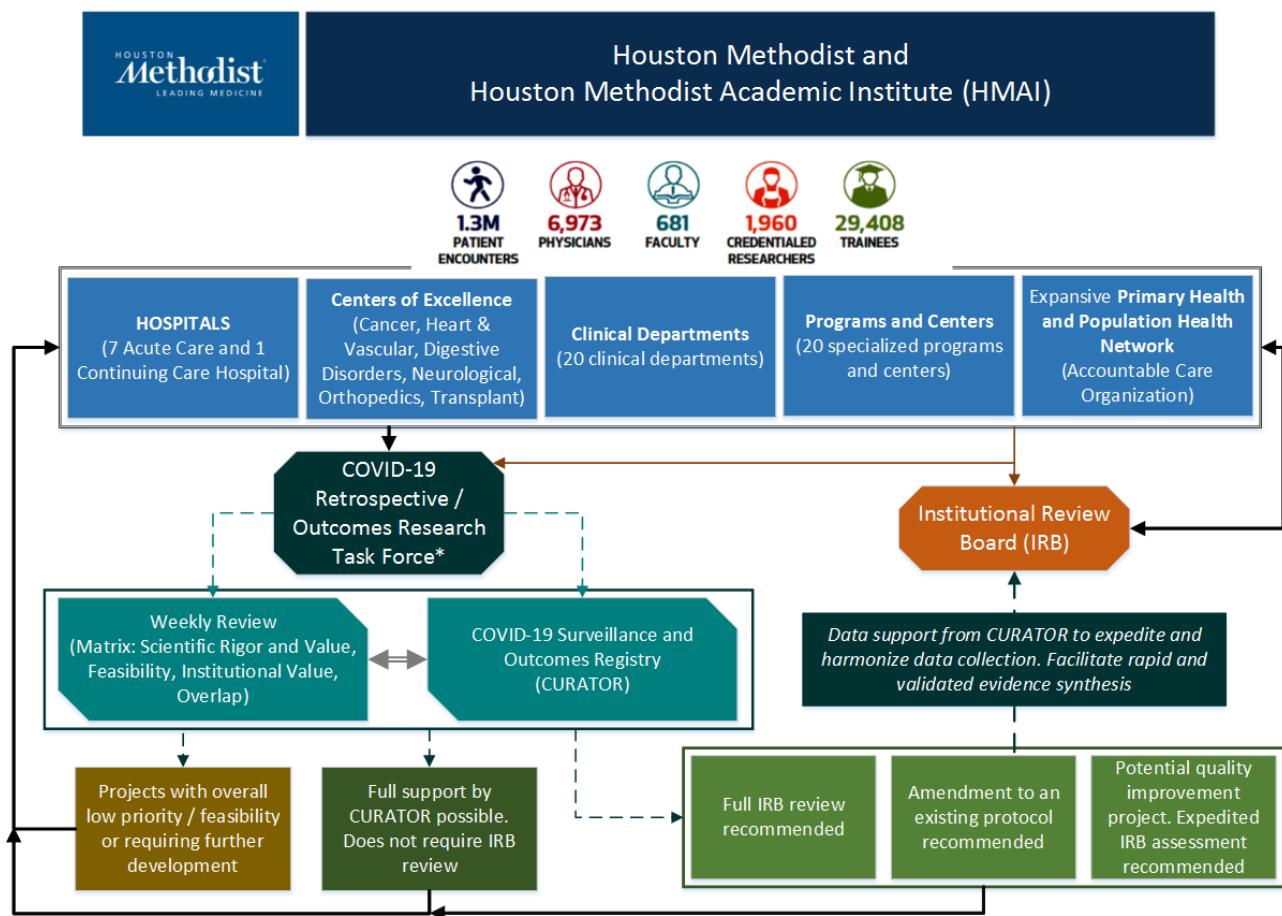ins a plethora of social, demographic, and medical information on thousands of patients in one location and is quite possibly one of the largest underutilized resources in modern medical research. However, at the time of the cusp of the COVID-19 pandemic, many health care facilities, including our own, lacked EHR add-ons that would allow for a rapid assimilation of research datasets. As the COVID-19 pandemic ensued, our research infrastructure faced an unprecedented need for validated datasets to support clinical trials and observational studies. Hence, to support research activities based on EHR, the RRTF decided to set up the Houston Methodist CURATOR. The goal of CURATOR is to serve as a unified, longitudinal, cross-institutional registry for COVID-19 data, to fulfill ongoing and long-term observational research data needs and enable availability of data for planning of prospective clinical trials.

## Results

### Structure, Workflow, and Output of the Houston Methodist COVID-19 RRTF

The RRTF was established on March 20, 2020 as a pre-IRB step after the institutional leadership effected a decision to accelerate the internal review, triage, and operationalization of a growing number of observational research protocols that were received by the IRB. The overarching clinical and academic structure of the Houston Methodist system, and the RRTF process framework in relation to the IRB, is schematically represented in Figure 1. The top two panels represent the organizational distribution of several physicians, physician scientists, translational and epidemiological scientists, and trainees spread across various hospitals, centers of excellence, clinical departments, programs and specialized centers, and an expansive primary care network. The solid black arrows represent communication pathways between investigators across this clinical and research enterprise and various elements of the COVID-19 RRTF and the Houston Methodist IRB. The RRTF initially reviewed all protocols related to COVID-19 and communicated back to the investigators directly in situations where the projects had opportunities for further development, were not technically sound, or did not require a full IRB review (see bottom-left brown and dark green text boxes in Figure 1). All other protocols, with specific comments and recommendations were forwarded to the IRB for a full evaluation (see bottom-right green text boxes in Figure 1). The Center for Outcomes Research (COR) at Houston Methodist Research Institute was tasked to set up the charter and workflow for the RRTF. The COR leadership team assembled the preliminary process documentation for the RRTF and a team comprising experts in epidemiology, health systems research, health policy, clinical domains, data sciences, information technology, and research regulation.

**Figure 1.** Schematic representation of the COVID-19 task force and the Houston Methodist COVID-19 Surveillance and Outcomes Registry (CURATOR) placement in the organizational context.



To accelerate and prioritize review of the influx of COVID-19–related protocols, all protocols received either directly or indirectly via the IRB are evaluated independently by RRTF members with a prioritization matrix and discussed on a weekly basis. The outcome of the RRTF process is communicated back to the investigators. During the extensive review activities developed by the multidisciplinary RRTF team, it was observed that many promising observational studies required similar data resources, leading to the development of a central COVID-19 data infrastructure to expedite research output for all scientists engaged in COVID-19 research [12]. For this purpose, the RRTF decided to develop and actively maintain a registry for COVID-19 surveillance and intrahospital outcomes as a key tangible output of its research-acceleration function. The design, data aspects, and front-end of this registry are addressed in the next section.

### The Houston Methodist CURATOR Protocol

#### *CURATOR Design and Cohorts*

The Houston Methodist CURATOR protocol was developed by the COR leadership and was approved by the Houston Methodist IRB on April 11, 2020. The CURATOR database comprises two cohorts. The first cohort includes all individuals who were tested (regardless of the test result) for COVID-19 at any Houston Methodist location (hospital or a free-standing clinical establishment) with any of the SARS-CoV-2 diagnostic tests, including antigen tests and the polymerase chain reaction test, or for SARS-CoV-2 serology. All clinical encounters dating back to June 2016 are included without a prospective end date. This means that the database includes, for each patient, the records of all prior (pretesting) and subsequent (posttesting) clinical encounters (ie, hospitalizations, emergency department or primary care visits, laboratory tests, imaging reports, medications, and specialists care) that happen either as standard of care, or as a part of systematic long-term follow-up (such as follow-up in specialized COVID-19 recovery clinics). The second cohort comprises COVID-19 patients who were managed at Houston Methodist facilities but originally tested elsewhere. Like the first cohort, any instances of clinical encounters prior to hospitalization or postacute care are included. More recently, CURATOR's protocol has been amended to include all individuals who have received or will be receiving a COVID-19 vaccine, regardless of their COVID-19 status.

The design elements of CURATOR allow for two salient aspects that strengthen methodological approaches in hypothesis generation and testing. First, by tracking records per patient, CURATOR creates a longitudinal array of individuals' health status. Availability of data from clinical encounters prior to testing and/or hospitalization permits granular, time-dependent, and accurate risk stratification for comorbidities based on longitudinally obtained medications, imaging, and laboratory test results, rather than cross-sectional documentation of

comorbid and pre-existing conditions at the time of COVID-19–related hospitalization or clinical encounter. Similarly, the information obtained from subsequent (post–COVID-19) encounters will provide information on recovery and outcomes. The second unique design element of CURATOR is the readily available data from a control population. By including data on all tested individuals, capturing retrospective and prospective clinical encounters of individuals who tested positive as well as those who tested negative for COVID-19, a large number of potential controls are available in CURATOR for hypothesis testing. For instance, matched case-control studies or prospective cohorts for incidence-based analyses can be used when developing COVID-19–related hypotheses. Figure 2 provides an updated schematic of the total number and proportion of individuals who underwent COVID-19 testing, those who tested positive or negative, and those who were hospitalized with at least one prior clinical encounter in the CURATOR.

**Figure 2.** Schematic representation of the total number and proportion of individuals who underwent COVID-19 testing, those who tested negative and positive, and those who were hospitalized with prior clinical encounters, based on data from COVID-19 Surveillance and Outcomes Registry (CURATOR).
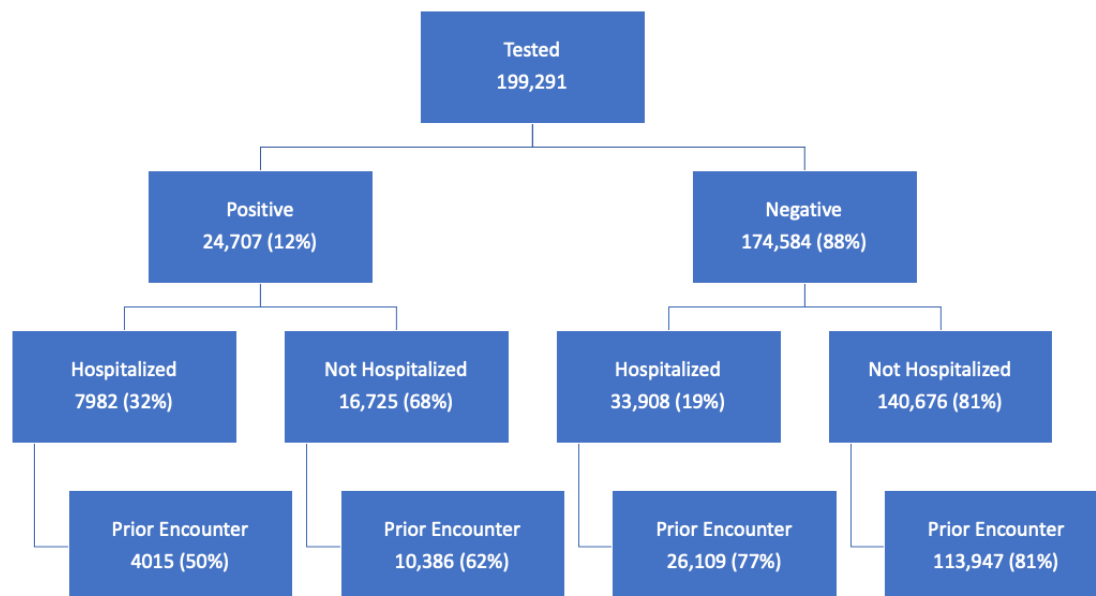


### CURATOR Data Elements and Data Structure

CURATOR is a relational structured query language database that is directly populated with the back-end data originating in one of the market-leading EHR vendors in the United States. The COR was uniquely equipped to undertake the creation and implementation of CURATOR with data scientists experienced in working with EHRs. The initial effort was to assemble back-end data by grouping disparate but related information from different locations within the EHR. Then, the database was iteratively refined to create meaningful tables and views of the data in an analytic data set that can be useful for researchers. The EHR back-end data refreshes every 24 hours based on live instances of the institutional EHRs.

The extract, transform, and load (ETL) procedures from EHR to the CURATOR database have been designed, developed, and implemented by a multispecialty team of COR scientists (big data and data science leads, epidemiologists, and physicians), data scientists, data engineers, and application analysts. In parallel, other ETL processes across the institution have also been simultaneously implemented by business intelligence teams to support clinical, operational, and administrative decision-making. The CURATOR team has undertaken an ongoing validation process with the business intelligence counterparts to streamline the ETL process and assess internal validity of COVID-19–related data across the system. Targeted smaller COR–business intelligence teams have been working together to this end. Updates and issues resulting from this cooperation are reviewed in weekly, biweekly, and monthly meetings. Currently, CURATOR is being populated with 1004 data items aggregated and organized across 87 tables and views. CURATOR is updated weekly by using an automated ETL process designed and implemented by the COR Big Data & Artificial Intelligence team. This process has been optimized to achieve efficiency and version control.

### Current Data

Currently, CURATOR contains extracted information for approximately 200,000 individuals, of whom approximately 25,000 tested positive for COVID-19, with approximately 14 million hospital encounters. For each patient, basic demographic (eg, age, sex, zip code, geocoding, marital status, and education level), ethnic identity (ethnicity and race), and baseline health (eg, BMI, IDC-10 code Charlson comorbidities, clinical morbidities, and immunizations) data are included. In addition, the CURATOR database includes the time-detailed ordering and results of laboratory tests, imaging, and procedures for each patient. For instance, the laboratory tests include cultures, real-time reverse transcription polymerase chain reaction tests, and SARS-CoV-2 antigen and antibody tests, among 3709 other COVID-19–related and unrelated tests for a total of approximately 76 million laboratory tests and results. The imaging results include multi-region computed tomography (CT)–coupled angiograms, abdomen, chest and heart CT scans,

echocardiograms, and multi-vessel interventional radiology, among 1977 other distinct imaging results, with approximately 1.8 million results in the database. The procedures include isolation; intraosseous infusions; red blood cell transfusion through peripheral veins; introduction of sera, toxoids, and vaccines into muscles; insertion of a tunnel vascular devices into the patient's chest; prone status; and 10,063 other procedures, with a total of approximately 400,000 procedures in the database. Furthermore, CURATOR contains time-resolved registries of the medications ordered (for inpatients and outpatients) and administered (inpatients), results on clinical trials, and the outcomes of each patient, including details on discharge status, discharge location, length of stay, oxygen therapy, ICU stays, the usage of mechanical ventilators, extracorporeal membrane oxygenation and endotracheal intubation, among many other variables informed by numerous research proposals that have traversed the RRTF process. The CURATOR database is continually growing to address wider research needs.

### CURATOR Integration with Other Internal and External Data Sources

CURATOR's design and implementation allows for seamless integration with other unique and siloed sources of big data across Houston Methodist. The virtual ICU (vICU) provides continuous, digitalized intensivist coverage for over 300 ICU beds at Houston Methodist. This remote ICU monitoring environment with embedded advanced telehealth capabilities captures real-time continuous physiological data on all ICU patients (including those with COVID-19) and provides an opportunity to develop predictive analytical tools to proactively identify critical risk factors and anticipate patient decompensation. The vICU platform at Houston Methodist was rapidly expanded following the COVID-19 pandemic [13]. Thus, vICU information is being integrated into CURATOR to broaden research perspectives and enrich the case histories with streaming physiologic data captured in real-time.

Additionally, Houston Methodist hosts one of the very few advanced translational imaging centers in the United States. This image center includes one of the most powerful 7-Tesla magnetic resonance imaging machines available. These advanced imaging modalities are rapidly being leveraged for the assessment, prognostication, and prediction of the effect of COVID-19 on pulmonary, cardiac, and neurological tissues. The outcomes of advanced COVID-19 imaging analyses will also be integrated into CURATOR.

Furthermore, Houston Methodist utilizes an innovative digital care navigation and data collection system for patient communication, education and awareness, and capturing patient-reported outcomes measures in postacute and long-term care setting (CareSense, MedTrak, Inc [14]). By using automated yet customized phone calls, text messages, emails, and app notifications, patients on various digital pathways are followed up with overarching goals to provide effective transition of care, promote safe recovery, and prevent complications. Success of these pathways has been previously reported and similar pathways are actively used for patients with COVID-19 [15]. The structure of CURATOR allows for seamless integration of the data sources obtained from the digital care navigation system.

Finally, CURATOR data is linkable via direct or probabilistic matching with external sources such as state-wide or national claims and administrative data sources. Certain derived data elements, such as area deprivation index [16], are now integrated into the routine workflow of CURATOR data updates.

### Regulation and Governance (Annual Audit, Review, and Stewardship)

The CURATOR protocol is governed and regulated by the Houston Methodist IRB. The protocol, training, and delegation logs, data governance policies, and data release and sharing procedures have been approved by the IRB, are maintained and updated by COR project management, and are subject to annual IRB audits. The governance committee comprises the COR leadership. All projects proposing to utilize CURATOR are subject to an independent review by the IRB. Projects led by Houston Methodist investigators that do not warrant sharing of protected health information may be exempt from an IRB review. However, these assessments are undertaken by the RRTF that has IRB representation. To date, CURATOR actively supports more than 30 IRB-approved COVID-19 research protocols across Houston Methodist (Textbox 1).

**Textbox 1.** Current list of projects at Houston Methodist approved by the Institutional Review Board and supported by COVID-19 Surveillance and Outcomes Registry (CURATOR).

---

**COVID-19 projects categorized by clinical discipline**

- **Cardiology**
  - Echocardiographic Findings in COVID-19 Patients
  - Cardiovascular Magnetic Resonance Imaging of Myocardial Damage in COVID-19 Patients
  - Vascular Disease and Complications of COVID-19
  - Troponin Elevation and Myocardial Infarction in COVID-19 Patients
  - Statin Therapy, Lipid Control, and Severe Illness in COVID-19 Among Patients With Cardiovascular Disease
  - Area Deprivation Index and Indicators of Severe COVID-19 Among Patients With Cardiovascular Disease

- **Neurology**
  - Stroke Outcomes Among COVID-19 Patients
  - Cognitive Outcomes Among COVID-19 Patients

- **Infectious disease**
  - Epidemiology of COVID-19
  - Biospecimens Related to COVID-19

- **Public health or disparities**
  - Race and Ethnic Disparities in SARS-CoV-2 Susceptibility
  - Race and Ethnic Disparities in COVID-19 Hospitalization and Mortality
  - Sex Differences in COVID-19 Outcomes
  - Characteristics and Outcomes of COVID-19 Across Various Pandemic Phases
  - Medication Outcomes Surveillance for COVID-19
  - ICU Ethics for COVID-19

- **Surgery**
  - Emergency Surgical Volumes during COVID-19 Pandemic
  - Outcomes among Transplant and Non-transplant Recipients with COVID-19
  - Surgery during COVID-19 Pandemic

- **Critical care**
  - Corticosteroid Use in COVID
  - Hydroxy Chloroquine Use and Outcomes in COVID-19
  - Proning Associated Outcomes in COVID-19
  - Tocilizumab Use and Outcomes in COVID-19
  - NISQIP and COVID-19
  - COVID-19 Treatment Algorithms and Outcomes

- **Rehab or physical therapy**
  - Physical Therapy in COVID-19 ICU

---

### Front-end User Interface

Data availability and democratization is a key component of the acceleration function of CURATOR. End-users can rapidly test hypotheses and identify feasible research lines based on preliminary studies using the database. Nonetheless, the access of end-users to data must be IRB-regulated, and processes and procedures to protect the data set from mishandling must be implemented. For this purpose, the CURATOR registry contains a web-accessible front-end that allows the end-user access to IRB-approved parts of the database via customizable, interactive charts. The charts are developed on static copies of CURATOR that are updated weekly; hence, accidental information disarrangement or system resources overconsumption is practically avoided. Our end-goal is to make the front-end of CURATOR available as a research tool across the health sciences communities. The front-end will also provide a seamless web-based communication platform between investigator teams, CURATOR management, and the IRB.

## Discussion

### Principal Findings

In this paper we describe how demanding data requirements were addressed by an administratively situated, EHR-integrated data structure for rapidly updated surveillance and outcomes data in the context of the COVID-19 pandemic.

The initial protocol was approved in less than 4 weeks from submission to the IRB review due to the critical need, via intensely responsive investigator–IRB communication. The scope involved several components that trigger particular IRB deliberation, including data pulled from various sources retrospectively and prospectively, with identifiers intact, under waiver of consent and Health Insurance Portability and Accountability Act (HIPAA) authorization, with future data banking planned. To simplify that deliberation and expedite the launch of a functional registry, components such as data sharing and data linking with potential external partners, sub-study personnel and scope of sub-studies, and secondary use of research data (including follow-up contact with patients in the registry) were relegated to future amendments or addenda to be reviewed by the IRB at a later date. This afforded the researchers' envisioned data governance committee time to convene and establish thoughtful policies on these matters; most importantly, the transparent promise of future amendments presents the IRB with information in an amount and at the time when necessary for implementation.

In our experience, it takes a unique multidisciplinary team, empowered by close contact with executive leadership, and a balance between ethics/rigor and speed during a pandemic to be able to drive impactful and meaningful observational research. At Houston Methodist, the integration of this team as a pre-IRB approval task force allowed us to design tools for fast-tracking research proposal triage, review, and operationalization. Furthermore, close contact with the hospital's leadership was key for the rapid dissemination of the RRTF duties, competences, and activities. In addition, the creation of a centralized task force unit allowed the RRTF to identify a set of common data elements across research proposals by

performing transversal analysis, thus allowing the data teams to begin data extraction in parallel with the IRB process to expedite the availability of the data to various research teams once they have secured IRB approval for their study. The insight into what was "in the research pipeline" allowed our data team to deliver data into the hands of researchers much more quickly than if IRB approval and data gatekeeping had been a serial process, as it ordinarily is.

Having one unified database as a single source of truth (SSoT) allowed us to focus resources on developing a database maintained with the highest standards. Additionally, the multidisciplinary nature of the team allows us to continuously enlarge the CURATOR database by looking at the SARS-CoV-2 pandemic from different angles and for different applications. Finally, the transition of CURATOR from a registry to a live source for hypothesis testing and research-line identification is being carried out by developing a front-end for this database. The availability of this front-end will not only reduce the querying loads to the back-end maintainers of the CURATOR database, but it will also help practitioners and researchers shorten the hypothesis test-validation cycle, leading to improved practice and research performance, respectively.

In addition to creating an SSoT, we aligned our approach and efforts with several other established principles of driving an effective digitization of health care industry [17]. First, CURATOR aims to break down data silos and create true functional interoperability between heterogenous data sources such as the traditional EHR, vICU, CareSense, and imaging data warehouses across the system. Second, we continue to evaluate and develop the analytical maturity of our informatics pipelines. As an example, the CURATOR infrastructure provides a concrete context and platform to utilize leading cloud-based technologies for analysis of continuous waveform data, develop machine learning and artificial intelligence models for image synthesis, and harness Natural Language Processing for some of the applications described below as current limitations. Third, by generating a validated cross-linkage between CURATOR and other business intelligence–driven data process across Houston Methodist, we aligned CURATOR's goals with that of the organization at large. Even though CURATOR has been set up as a research-oriented data resource, harmonization across the institution adds value and helps in engaging a wider array of stakeholders and resource allocation for continued support. Fourth, collaborating with the IRB and institutional leadership, we have set up robust governance structures that are clearly communicated and disseminated. Finally, our front-end interface provides data insight, data exploration, and communication tools that essentially facilitate regulated yet efficient data democratization and is a platform for developing further stake-holder driven applications.

The CURATOR model has significant implications for future research. In addition to providing a COVID-19–specific research platform, the CURATOR model also establishes a replicable DIP framework across several other clinical disciplines, particularly in the context of an LHC system. We believe that the focus of our institutional leadership on fostering a true LHC system enabled us to successfully resurrect and implement this infrastructure during a global pandemic. A systematic effort to

set up a similar framework across cardiology and neurology service domains is underway and significant investments have been made across other clinical domains. The CURATOR model, although catalyzed by the COVID-19 pandemic, is not a "one-and-done" project; instead, it is an ardent representation of a data centric health care organization that has poised itself to lead medicine and health care delivery and overcome health care digitization challenges of the future.

## Comparisons to Prior Work

In the wake of the COVID-19 pandemic, the need for validated data sources has been appreciated widely across the health care industry. Broadly, two approaches have been adopted and reported in literature. First, existing registries and data warehouses have been modified to include data elements pertaining to COVID-19. In most cases, such resources are clinical domain–driven, such as the American Heart Association's Get With The Guidelines Registries [18], the American Academy of Orthopedic Surgeon's Registry [19], the American College of Surgeons COVID-19 Registry [20], the American College of Radiology COVID-19 Imaging Research Registry [21], and the American Academy of Dermatology Association COVID-19 Registry [22]. This approach capitalizes on an existing network of participating organizations and has an advantage of a fairly well-established data pipeline and governance structure. However, this approach is specific to individual clinical domains and is therefore of limited utility to a wider array of stakeholders.

The second approach entails establishing dedicated data repositories for COVID-19 research, agnostic to other clinical domains. Data and information regarding such registries are, however, limited. Theoretically, these could be single or multi-institution endeavors. The Innovative Support for Patients With SARS-CoV-2 Infections Registry (INSPIRE) has been registered with ClinicalTrials.gov with a primary outcome of ascertaining incidence of myalgic encephalomyelitis or chronic fatigue syndrome across 8 institutions [23]. The INSPIRE investigators propose to enroll 3600 patients with COVID-19 and 1200 controls over a 2-year period. Other emerging examples of single-center COVID-19 registries include the Stanford University COVID-19 registry [24], Rice University COVID-19 Registry [25], and Johns Hopkins COVID-19 Precision Medicine Analytics Platform Registry (JH-CROWN) [26]. Dedicated COVID-19 registries have the strength of providing a platform for comprehensive analyses of COVID-19–related hypotheses, regardless of patients with pre-existing comorbidities or other clinical conditions. However, considerable de novo ETL efforts may be required to set up such resources. Furthermore, as the evidence indicates, establishing cross-institutional partnerships will take considerable additional effort in creating common data definitions models, harmonizing data processes, and setting up regulatory and governance structures.

CURATOR exemplifies a single, large health care institutional COVID-19 registry. However, given the pre-existing institutional commitment and investment in an LHC system, it was established at a rapid pace and, as we have discussed, conforms to several of the fundamental principles of health care

digitization. Without much information published on other institutional COVID-19 registries, a direct head-to-head comparison is not feasible. However, CURATOR capitalizes on several unique data sources and currently supports COVID-19 projects across several domains along with work already published from CURATOR data and its associated resources across Houston Methodist [7,27-32].

## Limitations

Although the architecture of CURATOR was designed and automated to retrieve new and updated data in a near-real-time implementation, it is limited to a single-center, longitudinal medical history record. In its current iteration, CURATOR cannot capture clinical encounter information from systems outside of Houston Methodist. In the context of a global pandemic, this is a significant limitation. However, the CURATOR leadership currently partners with local, regional, national, and international consortia, which provides an ongoing opportunity to establish common data element models for harmonization with external data sources. There are also significant challenges with other incomplete, yet highly relevant data (eg, presenting symptoms data is largely unstructured and incomplete with regards to salient elements such as timing, progression and severity of symptoms, as well as palliative measures). Future implementation of natural language processing pipelines is envisioned as a solution. CURATOR, like all retrospective registries, relies on "samples of convenience," and hence suffers from a certain degree of selection and information bias. Systematic selection of COVID-19 patients and planned follow-up in COVID-19 recovery clinics across Houston Methodist will minimize the influence of such potential bias. Finally, quantifiable assessment of true impact of CURATOR on reducing research timelines across our organization is not currently possible due to the limitation of resources that would be needed to perform a comparison across historical data or collect specific metrics on investigators' perspectives. However, CURATOR metrics are being actively monitored and documented and such assessments would be possible in future. Despite these limitations, CURATOR and similar efforts are powerful tools in finding the signal in the noise when confronted from every angle with the unknown, as we are, during the outbreak of any novel pathogen.

## Conclusions

In the face of rapidly evolving COVID-19 pandemic, the health care industry's challenge of meaningful digitization has been exacerbated. Developing a data-driven, clinical, operational, and research decision-making strategy is paramount to the success of health care organizations. We share our experience of how a large, tertiary care health care organization and its research enterprise rapidly adapted to this challenge and created COVID-19–centric mechanisms of efficient and validated decision-making across a complex health care enterprise. The cross-disciplinary expertise, investment in health care technology, and leadership commitment are key ingredients to establish and foster an LHC system. Such systems, if optimally developed, can mitigate the effects of ongoing and future health

care catastrophes by providing timely and validated decision    support mechanisms.

## Authors' Contributions

All authors contributed to drafting and critically reviewing this manuscript and have read and approved the final manuscript. FSV, SLJ, MET, FS, GN, and BAK wrote the manuscript and reviewed it for scientific content. FSV created Figure 1. JCN, OAK, JM, and APP provided data for Figure 2 and Textbox 1, and they reviewed the manuscript for technical content. TM, HDS, and KN reviewed the manuscript for scientific content.

## Conflicts of Interest

None declared.

## References

1.  COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). Johns Hopkins University & Medicine - Coronavirus Resource Center. URL: https://coronavirus.jhu.edu/map.html [accessed 2021-02-12]
2.  Woolliscroft JO. Innovation in response to the COVID-19 pandemic crisis. Acad Med 2020 Aug;95(8):1140-1142 [FREE Full text] [doi: 10.1097/ACM.0000000000003402] [Medline: 32282372]
3.  Farrugia G, Plutowski RW. Innovation lessons From the COVID-19 pandemic. Mayo Clin Proc 2020 Aug;95(8):1574-1577 [FREE Full text] [doi: 10.1016/j.mayocp.2020.05.024] [Medline: 32753130]
4.  Vest JR, Kash BA. Differing strategies to meet information-sharing needs: publicly supported community health information exchanges versus health systems' enterprise health information exchanges. Milbank Q 2016 Mar;94(1):77-108 [FREE Full text] [doi: 10.1111/1468-0009.12180] [Medline: 26994710]
5.  Houston Facts. Greater Houston Partnership. Houston, TX; 2019. URL: https://www.houston.org/sites/default/files/2019-08/Houston%20Facts%202019%20Final_3.pdf [accessed 2021-02-12]
6.  Emerson M, Bratter J, Howell J, Jeanty P, Cline M. Houston Region Grows More Racially/Ethnically Diverse, With Small Declines in Segregation A Joint Report Analyzing Census Data from 1990, 2000, and 2010. Kinder Institute for Urban Research & the Hobby Center for the Study of Texas. URL: https://kinder.rice.edu/sites/default/files/documents/Houston%20Region%20Grows%20More%20Ethnically%20Diverse%204-9.pdf [accessed 2021-02-12]
7.  Vahidy FS, Drews AL, Masud FN, Schwartz RL, Askary BB, Boom ML, et al. Characteristics and outcomes of COVID-19 patients during initial peak and resurgence in the Houston Metropolitan Area. JAMA 2020 Sep 08;324(10):998-1000. [doi: 10.1001/jama.2020.15301] [Medline: 32789492]
8.  COVID-19 in Texas (Dashboard). Texas Health and Human Services. URL: https://txdshs.maps.arcgis.com/apps/opsdashboard/index.html#/ed483ecd702b4298ab01e8b9cafc8b83 [accessed 2021-02-12]
9.  Facts and Statistics (2019). Houston Methodist. URL: https://www.houstonmethodist.org/newsroom/facts-statistics/ [accessed 2021-02-12]
10. About Us: Facts and Stats (2020). Houston Methodist. URL: https://www.houstonmethodist.org/research/about-us/facts-stats/ [accessed 2021-02-12]
11. Salazar E, Perez KK, Ashraf M, Chen J, Castillo B, Christensen PA, et al. Treatment of COVID-19 patients with convalescent plasma in Houston, Texas. medRxiv Preprint posted online on May 13, 2020 [FREE Full text] [doi: 10.1101/2020.05.08.20095471] [Medline: 32511574]
12. Arbab Kash B, Spaulding A, Gamm LD, Johnson CE. Healthcare strategic management and the resource based view. Journal of Strategy and Mgt 2014 Aug 12;7(3):251-264. [doi: 10.1108/jsma-06-2013-0040]
13. Dhala A, Sasangohar F, Kash B, Ahmadi N, Masud F. Rapid implementation and innovative applications of a virtual intensive care unit during the COVID-19 pandemic: case study. J Med Internet Res 2020 Sep 03;22(9):e20143 [FREE Full text] [doi: 10.2196/20143] [Medline: 32795997]
14. CareSense. URL: https://www.caresense.com/ [accessed 2021-02-12]
15. Bruce CR, Harrison P, Nisar T, Giammattei C, Tan NM, Bliven C, et al. Assessing the impact of patient-facing mobile health technology on patient outcomes: retrospective observational cohort study. JMIR Mhealth Uhealth 2020 Jun 26;8(6):e19333 [FREE Full text] [doi: 10.2196/19333] [Medline: 32589161]

16.    Knighton AJ, Savitz L, Belnap T, Stephenson B, VanDerslice J. Introduction of an Area Deprivation Index measuring
       patient socioeconomic status in an integrated health system: implications for population health. EGEMS (Wash DC)
       2016;4(3):1238 [FREE Full text] [doi: 10.13063/2327-9214.1238] [Medline: 27683670]
17.    Waller D. 10 steps to creating a data-driven culture. Harvard Business Review. 2020 Feb 06. URL: https://hbr.org/2020/
       02/10-steps-to-creating-a-data-driven-culture [accessed 2021-02-12]
18.    Alger HM, Williams JH, Walchok JG, Bolles M, Fonarow GC, Rutan C. Role of data registries in the time of COVID-19.
       Circ Cardiovasc Qual Outcomes 2020 May;13(5):e006766. [doi: 10.1161/CIRCOUTCOMES.120.006766] [Medline:
       32298145]
19.    AAOS Registry Program to Collect COVID-19 Data. American Academy of Orthopaedic Surgeons. URL: https://www.
       aaos.org/registries/covid-19-data/ [accessed 2021-02-12]
20.    ACS COVID-19 Registry. American College of Surgeons. URL: https://www.facs.org/Quality-Programs/covid19-registry
       [accessed 2021-02-12]
21.    COVID-19 Imaging Research Registry. American College of Radiology. URL: https://www.acr.org/Research/
       Clinical-Research/COVID-19-Registry [accessed 2021-02-12]
22.    COVID-19 Dermatology Registry. American Academy of Dermatology Association. URL: https://www.aad.org/member/
       practice/coronavirus/registry [accessed 2021-02-12]
23.    Hota B. Innovative Support for Patients With SARS-COV2 Infections (COVID-19) Registry (INSPIRE). ClinicalTrials.
       2020. URL: https://clinicaltrials.gov/ct2/show/NCT04610515 [accessed 2021-02-12]
24.    Stanford Medicine Research Registry for COVID-19 Studies. Stanford Medicine. URL: https://med.stanford.edu/covid19/
       research-registry.html [accessed 2021-02-12]
25.    COVID-19 Registry. Rice University. URL: https://registry.rice.edu/covid19 [accessed 2021-02-12]
26.    COVID-19 Precision Medicine Analytics Platform Registry (JH-CROWN). Johns Hopkins Institute for Clinical &
       Translational Research. URL: https://ictr.johnshopkins.edu/coronavirus/jh-crown/ [accessed 2021-02-12]
27.    Vahidy FS, Bernard DW, Boom ML, Drews AL, Christensen P, Finkelstein J, et al. Prevalence of SARS-CoV-2 infection
       among asymptomatic health care workers in the Greater Houston, Texas, Area. JAMA Netw Open 2020 Jul 01;3(7):e2016451
       [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.16451] [Medline: 32716512]
28.    Pan A, Khan O, Meeks J, Boom M, Masud F, Andrieni J, et al. Disparities in COVID-19 hospitalizations and mortality
       among Black and Hispanic patients: cross-sectional analysis from the Greater Houston Metropolitan Area. medRxiv Preprint
       posted online on August 22, 2020. [doi: 10.1101/2020.08.19.20177956]
29.    Vahidy FS, Nicolas JC, Meeks JR, Khan O, Pan A, Jones SL, et al. Racial and ethnic disparities in SARS-CoV-2 pandemic:
       analysis of a COVID-19 observational registry for a diverse US metropolitan population. BMJ Open 2020 Aug
       11;10(8):e039849 [FREE Full text] [doi: 10.1136/bmjopen-2020-039849] [Medline: 32784264]
30.    Sasangohar F, Jones SL, Masud FN, Vahidy FS, Kash BA. Provider burnout and fatigue during the COVID-19 pandemic:
       lessons learned from a high-volume intensive care unit. Anesth Analg 2020 Jul;131(1):106-111 [FREE Full text] [doi:
       10.1213/ANE.0000000000004866] [Medline: 32282389]
31.    Sasangohar F, Dhala A, Zheng F, Ahmadi N, Kash B, Masud F. Use of telecritical care for family visitation to ICU during
       the COVID-19 pandemic: an interview study and sentiment analysis. BMJ Qual Saf 2020 Oct 07. [doi:
       10.1136/bmjqs-2020-011604] [Medline: 33028659]
32.    Sasangohar F, Bradshaw MR, Carlson MM, Flack JN, Fowler JC, Freeland D, et al. Adapting an outpatient psychiatric
       clinic to telehealth during the COVID-19 pandemic: a practice perspective. J Med Internet Res 2020 Oct 01;22(10):e22523
       [FREE Full text] [doi: 10.2196/22523] [Medline: 32936768]

## Abbreviations

**COR:** Center for Outcomes Research
**CT:** computerized tomography
**CURATOR:** COVID-19 Surveillance and Outcomes Registry
**DIP:** data infrastructure and pipeline
**EHR:** electronic health record
**ETL:** extract, transform, and load
**ICU:** intensive care unit
**INSPIRE:** Innovative Support for Patients With SARS-CoV-2 Infections Registry
**LHC:** learning health care system
**RRTF:** retrospective research task force
**SSoT:** single source of truth
**vICU:** virtual intensive care unit

XSL•FO

**RenderX**

Original Paper

# Utilization of Telehealth Services in Libya in Response to the COVID-19 Pandemic: Cross-sectional Analysis

Muhammed Elhadi[1], MBBCh; Ahmed Msherghi[1], MBBCh; Ahmed Elhadi[1], MBBCh; Aimen Ashini[1], MD; Ahmed Alsoufi[1], MBBCh; Fatimah Bin Alshiteewi[1], MBBCh; Amna Elmabrouk[1], MBBCh; Ali Alsuyihili[1], MBBCh; Alsafa Elgherwi[1], MBBCh; Fatimah Elkhafeefi[2], MBBCh; Sarah Abdulrazik[1], MD; Ahmed Tarek[1], MBBCh

[1]Faculty of Medicine, University of Tripoli, Tripoli, Libyan Arab Jamahiriya

[2]Faculty of Medicine, University of Benghazi, Benghazi, Libyan Arab Jamahiriya

**Corresponding Author:**
Muhammed Elhadi, MBBCh
Faculty of Medicine
University of Tripoli
Furnaj, University Road
Tripoli, 13275
Libyan Arab Jamahiriya
Phone: 218 945196407
Email: muhammed.elhadi.uot@gmail.com

## Abstract

**Background:** Health care systems in transitional countries have witnessed unprecedented challenges related to adequate and continuous health care provision during the COVID-19 pandemic. In many countries, including Libya, institutions and organizations have begun to implement telehealth technology for the first time. This serves to establish an alternative modality for direct physician-patient interviews to reduce the risk of COVID-19 transmission.

**Objective:** This study aimed to assess the usability of telehealth services in Libya and to provide an overview of the current COVID-19 scenario.

**Methods:** In this cross-sectional study, an anonymous web-based survey was administered to Libyan residents between April and May 2020. Participants were contacted through text messaging, emails, and social media. The survey items yielded information on the sociodemographic characteristics, availability and accessibility of health care services, effects of the COVID-19 pandemic on health care services, mental health status, and the feasibility and application of the telehealth system.

**Results:** We obtained 2512 valid responses, of which 1721 (68.5%) were from females. The participants were aged 28.2 (SD 7.6) years, of whom 2333 (92.9%) were aged <40 years, and 1463 (58.2%) were single. Regarding the health care services and their accessibility, 786 (31.1%) participants reported having a poor health status in general, and 492 (19.6%) reported having a confirmed diagnosis of at least one chronic disease. Furthermore, 498 (19.9%) participants reported varying degrees of difficulty in accessing health care centers, and 1558 (62.0%) could not access their medical records. Additionally, 1546 (61.6%) participants experienced problems in covering medical costs, and 1429 (56.9%) avoided seeking medical care owing to financial concerns. Regarding the feasibility of the telehealth system, approximately half of the participants reported that telehealth services were useful during the COVID-19 pandemic, and 1545 (61.5%) reported that the system was an effective means of communication and of obtaining health care services. Furthermore, 1435 (57.1%) participants felt comfortable using the telehealth system, and 1129 (44.9%) felt that they were able to express themselves effectively. Moreover, 1389 (55.3%) participants found the system easy to understand, and 1354 (53.9%) reported having excellent communication with physicians through the telehealth system. However, only 1018 (40.5%) participants reported that communication was better with the telehealth system than with traditional methods.

**Conclusions:** Our study revealed high levels of usability and willingness to use the telemedicine system as an alternative modality to in-person consultations among the Libyan residents in this study. This system is advantageous because it helps overcome health care costs, increases access to prompt medical care and follow-up evaluation, and reduces the risk of COVID-19 transmission. However, internet connectivity and electricity issues could be a substantial barrier for many resource-limited communities, and further studies should address such obstacles.

## Introduction

Since December 2019, COVID-19 has disrupted health care systems in many countries [1,2]. SARS-CoV-2, the causative virus of the severe viral pneumonia COVID-19, has infected more than 14,000,000 individuals and resulted in approximately 600,000 deaths worldwide [3,4].

The emergence of the COVID-19 pandemic has posed unprecedented challenges on health care systems in many countries, resulting in the cancellation of surgical procedures, closure of clinics, and an increase in the burden on health care services from patients with COVID-19 who require careful consideration and more care [5-7]. In response to this disruption, many countries have started implementing strategies to reduce the risk of disease transmission and to provide adequate health care services, especially among individuals with chronic diseases who need more care [8-11].

Therefore, many countries have started using telehealth services; that is, advanced technologies such as video, audio, and other means to provide health care services and promote the well-being of individuals while being physically distant from the health care provider [12-14]. Telehealth services offer several advantages to both patients and health care providers. The benefits of this system among patients include a shorter waiting time, a reduced need to travel long distances from their homes for health consultations, total avoidance of transport, and a reduced risk of infection, especially during the COVID-19 pandemic. The telehealth system offers several advantages to health care workers, including a marked reduction in patient interaction, the ability to assess patients with different diseases, the ability to perform follow-up evaluation and record outcomes, and a reduction in stress related to the COVID-19 pandemic.

The Libyan health care system operates primarily through public health centers and primary care centers, at which health services are provided at no cost. However, this manner of health care provision lacks sophisticated and advanced health services. In addition, a large number of private centers provide paid health services that may confer more advantages over traditional public health care services. However, these expensive private care services might represent a barrier to care provision among many Libyan residents. Although some large companies provide health insurance to their employees, such as policies that cover the main services provided by private health care centers, this provision is limited to a very specific portion of the population.

The Libyan health care system was not prepared for the COVID-19 pandemic. Hospitals and health care workers in Libya need better training and more resources to combat the pandemic, and there have been shortages of personal protective equipment, a lack of health services, and few prepared intensive care units [15,16]. These difficulties, along with the current ongoing civil war, pose challenges on the Libyan health care system, resulting in a high demand for urgent and determined action [17].

Many institutions and organizations have initiated telehealth programs for the first time during the COVID-19 pandemic in Libya to help patients avail of health care services and consultations without the need to visit hospitals and to reduce the risk of COVID-19 transmission. Since this is the first time telehealth services have been introduced in Libya, their usability by the Libyan population and their health-related effects need to be assessed, and the challenges faced by health care authorities and organizations in implementing these new strategies need to be determined. In this study, we aimed to assess the usability of telehealth services in Libya and to provide an overview of the current COVID-19 scenario.

## Methods

### Participant Recruitment

In this cross-sectional study, we distributed an anonymous web-based survey among Libyan residents between April and May 2020 via text message, email, and social media platforms. The inclusion criteria were that participants should be Libyan residents aged over 18 years. Incomplete questionnaires with >30% missing data were excluded from the analysis. We marked essential questions with the "required" function to ensure high quality of the data collected.

### Measurement Tool

The web-based survey was conducted in accordance with the Checklist for Reporting Results of Internet E-Surveys [18]. The first section of the survey included questions on demographic characteristics such as age, gender, social status, economic stability, residential status, smoking status, and general health. The second section of the survey comprised questions on the availability and accessibility of health care services, including the distance travelled to access nearest health care facility and the time taken to reach it, mode of transport available, and ability to cover medical costs. The third section comprised questions related to the effects of the COVID-19 pandemic on health care services. A subsequent section contained questions on mental health status, screening symptoms of anxiety and depression using the self-administered 9-item Patient Health Questionnaire (PHQ-9) [19,20], wherein a score of ≥15 was considered the threshold for depression. The 7-item Generalized Anxiety Disorder scale [21] was used to assess anxiety, with a score of ≥15 as the threshold for anxiety. The final section of the questionnaire assessed the usability of telehealth in accordance with the Telehealth Usability Questionnaire devised by Parmanto et al [22] and validated by Zhou et al [23]. The telehealth questionnaire was translated to Arabic by two independent translators using the forward-backward translation method. Any discrepancy in the translated versions was resolved through further discussion until consensus was reached. The translated

XSL•FO
**RenderX**

questionnaire was piloted with 30 participants, and the internal consistency was determined from a Cronbach α of .74 in the translated Arabic version. The telehealth usability survey included 3 items on the usefulness of telehealth, 3 items on the ease of use and learnability, 4 items on interface quality, 4 items on the quality of interaction, 3 items on reliability, and 4 items on satisfaction. Each of these items was scored on a 3-point Likert scale (0=agree, 1=neutral, 2=disagree).

## Statistical Analysis

Descriptive statistics are presented as frequency and percentage values. Continuous variables are presented as mean (SD) values. We performed a chi-square test to investigate associations between basic study characteristics and gender. Statistical analyses were performed using SPSS software (version 25.0, IBM Corp). A $P$ value of ≤.05 was considered significant

## Ethical Considerations

The study was approved by the Bioethics Committee of the Biotechnology Research Centre (Tripoli, Libya; 109.3-2020).

Participants provided written informed consent and their anonymity was maintained.

## Results

### Participant Characteristics

We obtained 2512 valid responses. In total, 1721 (68.5%) respondents were female and 791 (31.5%) were male. The study participants were aged 28.2 (SD 7.6) years. Most participants (n=2333, 92.9%) were aged <40 years and 1463 (58.2%) were single. In total, 766 (30.5%) participants were students and constituted the majority of the cohort by occupation. A total of 492 (19.6%) respondents reported having a chronic disease, 1137 (45.3%) reported having an excellent health status, and 786 (31.1%) reported having a poor health status. Table 1 summarizes the baseline demographic characteristics of the study population. We observed a significant association between gender and age ($P<.001$), marital status ($P=.002$), employment status ($P<.001$), stable income ($P=.03$), smoking status ($P<.001$), chronic disease ($P=.02$), and living arrangement ($P=.03$).

**Table 1.** Baseline demographic characteristics of the study population (N=2512).

| Variables | Total, n (%) | Female (1721), n (%) | Male (n=791), n (%) | P value |
|---|---|---|---|---|
| **Age range (years)** | | | | <.001[a] |
| 18-25 | 1062 (42.3) | 748 (43.5) | 314 (39.7) | |
| 26-40 | 1271 (50.6) | 878 (51.0) | 393 (49.7) | |
| >40 | 179 (7.1) | 95 (5.5) | 84 (10.6) | |
| **Marital status** | | | | .002[b] |
| Married | 971 (38.7) | 707 (41.1) | 264 (33.4) | |
| Single | 1463 (58.2) | 960 (55.8) | 503 (63.6) | |
| Divorced | 58 (2.3) | 39 (2.3) | 19 (2.4) | |
| Widowed/widower | 20 (0.8) | 15 (0.9) | 5 (0.6) | |
| **Employment status** | | | | <.001[a] |
| Student | 766 (30.5) | 527 (30.6) | 239 (30.2) | |
| Business and management | 180 (7.2) | 119 (6.9) | 61 (7.7) | |
| Engineering and manufacturing | 68 (2.7) | 42 (2.4) | 26 (3.3) | |
| Health care | 315 (12.5) | 213 (12.4) | 102 (12.9) | |
| Teaching and education | 266 (10.6) | 205 (11.9) | 61 (7.7) | |
| Science and pharmaceutical | 65 (2.6) | 47 (2.7) | 18 (2.3) | |
| Retails and sales | 12 (0.5) | 3 (0.2) | 9 (1.1) | |
| Homemaker | 377 (15.0) | 327 (19.0) | 50 (6.3) | |
| Laborer | 24 (1.0) | 17 (1.0) | 7 (0.9) | |
| Freelance | 54 (2.1) | 10 (0.6) | 44 (5.6) | |
| Retired | 4 (0.2) | 2 (0.1) | 2 (0.3) | |
| Unemployed | 316 (12.6) | 177 (10.3) | 139 (17.6) | |
| Other | 65 (2.6) | 32 (1.9) | 33 (4.2) | |
| Stable income | 846 (33.7) | 556 (32.3) | 290 (36.7) | .03[b] |
| Smoking | 104 (4.1) | 35 (2.0) | 69 (8.7) | <.001[a] |
| Having chronic disease | 492 (19.6) | 329 (19.1) | 163 (20.6) | .02[b] |
| **Living arrangement** | | | | .03[b] |
| With family | 2065 (82.2) | 1396 (81.1) | 669 (84.6) | |
| Alone | 447 (17.8) | 325 (18.9) | 122 (15.4) | |
| **General health status** | | | | .40 |
| Excellent | 1137 (45.3) | 766 (44.5) | 371 (46.9) | |
| Very good | 61 (2.4) | 44 (2.6) | 17 (2.1) | |
| Good | 528 (21.0) | 376 (21.8) | 152 (19.2) | |
| Bad | 786 (31.1) | 535 (31.1) | 251 (31.7) | |

[a]P<.001.

[b]P<.05.

## Accessibility of Health Care Services

Most participants reported an adequate level of ease in accessing health care centers, with 715 (28.5%) responding with "very easy" and 1299 (51.7%) responding with "easy." Regarding their mode of transport, almost half of the participants (n=1140, 45.4%) reported driving their own car, and 1097 (43.7%) hired a taxi or a private driver. In addition, 989 (39.4%) respondents reported that it took them 15-30 minutes to reach a health care facility, whereas 804 (32%) reported that they needed less than 15 minutes to reach their nearest facility. Furthermore, 1407 (56%) participants reported that they could consult with a

specialist physician within 2 days, whereas 1558 (62%) reported that they could not access their medical records or files.

Most participants (n=1546, 61.6%) found it difficult to cover the costs of medical care. More than half of the participants (n=1429, 56.9%) reported that they avoided seeking medical care for the fear of being financially burdened. Some participants (n=1225, 48.7%) reported difficulties in availing of emergency health care services. Most participants (n=2171, 86.4%) agreed that the working hours of clinics needed to be extended; however, approximately 1264 (50.4%) reported that it was easy to interact with physicians and nurses during health care counseling, and the remaining 1248 (49.6%) participants responded with "neutral" or disagreed with the statement, "Interactions with nurses and doctors are easy." The findings of the assessment of the accessibility of health care services are summarized in Table 2.

**Table 2.** Accessibility of health care services among the study participants (N=2512).

| Component | Total, n (%) |
| --- | --- |
| **Ease of traveling to the health care center** | |
| Very easy | 715 (28.5) |
| Somewhat easy | 1299 (51.7) |
| Slightly difficult | 441 (17.6) |
| Very difficult | 57 (2.3) |
| **Mode of transport to the health care center** | |
| Private car | 1140 (45.4) |
| Friend's or relative's car | 83 (3.3) |
| Public transport | 93 (3.7) |
| Taxi or private driver | 1097 (43.7) |
| Walking | 99 (3.9) |
| **How long does it take to reach the health care facility?** | |
| <15 minutes | 804 (32.0) |
| 15-30 minutes | 989 (39.4) |
| 31-45 minutes | 362 (14.4) |
| 46-60 minutes | 129 (5.1) |
| 1-2 hours | 132 (5.3) |
| >2 hours | 96 (3.8) |
| **If you get sick and need to see a specialist physician, how long does it take to get an appointment?** | |
| ≤2 days | 1407 (56.0) |
| 3 days to 1 week | 704 (28.0) |
| 1-2 weeks | 228 (9.1) |
| 3-4 weeks | 70 (2.8) |
| >4 weeks | 103 (4.1) |
| **The health care center provides me with the results of all my laboratory tests** | |
| No | 615 (24.5) |
| Rarely | 193 (7.7) |
| Sometimes | 536 (21.3) |
| Usually | 568 (22.6) |
| Always | 600 (23.9) |
| **If I want to see my medical records, the health care center allows me to have them** | |
| No | 1558 (62.0) |
| Rarely | 172 (6.8) |
| Sometimes | 276 (11.0) |
| Usually | 207 (8.2) |
| Always | 299 (11.9) |
| **I find it difficult to cover my medical care costs** | |
| Strongly disagree | 42 (1.7) |
| Disagree | 220 (8.8) |
| Neutral | 704 (28.0) |
| Agree | 899 (35.8) |
| Strongly agree | 647 (25.8) |

| Component | Total, n (%) |
|---|---|
| **I avoid seeking medical care owing to financial concerns** | |
| Strongly disagree | 106 (4.2) |
| Disagree | 396 (15.8) |
| Neutral | 581 (23.1) |
| Agree | 746 (29.7) |
| Strongly agree | 683 (27.2) |
| **I find it challenging to obtain prompt medical counseling in emergencies** | |
| Strongly disagree | 83 (3.3) |
| Disagree | 392 (15.6) |
| Neutral | 812 (32.3) |
| Agree | 797 (31.7) |
| Strongly agree | 428 (17.0) |
| **Hospital care is always accessible without difficulty** | |
| Strongly disagree | 119 (4.7) |
| Disagree | 479 (19.1) |
| Neutral | 777 (30.9) |
| Agree | 849 (33.8) |
| Strongly agree | 288 (11.5) |
| **I can get medical advice whenever I need it** | |
| Strongly disagree | 70 (2.8) |
| Disagree | 465 (18.5) |
| Neutral | 842 (33.5) |
| Agree | 879 (35.0) |
| Strongly agree | 256 (10.2) |
| **Working hours of clinics should be further extended** | |
| Strongly disagree | 0 (0) |
| Disagree | 49 (2.0) |
| Neutral | 292 (11.6) |
| Agree | 1001 (39.8) |
| Strongly agree | 1170 (46.6) |
| **Interactions with nurses and physicians are easy** | |
| Strongly disagree | 130 (5.2) |
| Disagree | 516 (20.5) |
| Neutral | 602 (24.0) |
| Agree | 853 (34.0) |
| Strongly agree | 411 (16.4) |

## Impact of the COVID-19 Pandemic on the Participants' Health

Only 90 (3.6%) participants agreed that the lockdown had affected their health, while most participants (n=2207, 87.9%) disagreed with that statement. Most participants (n=1264, 86.8%) disagreed with the statement that the lockdown had psychological effects. However, 1531 (60.9%) participants were worried about being afflicted with COVID-19, and 1974 (78.6%) were worried about their family members becoming infected. Table 3 provides an overview of the effects of COVID-19 on the survey participants.

**Table 3.** Effects of the COVID-19 pandemic on the health of the study participants (N=2512).

| Component | Total, n (%) |
|---|---|
| **The lockdown affected my health** | |
| Strongly disagree | 1241 (49.4) |
| Disagree | 966 (38.5) |
| Neutral | 215 (8.6) |
| Agree | 90 (3.6) |
| Strongly agree | 0 (0) |
| **The lockdown markedly affected my psychological status** | |
| Strongly disagree | 1155 (46.0) |
| Disagree | 1025 (40.8) |
| Neutral | 138 (5.5) |
| Agree | 194 (7.7) |
| Strongly agree | 0 (0) |
| **Are you worried about getting COVID-19?** | |
| Yes | 1531 (60.9) |
| No | 981 (39.1) |
| **Are you worried about the current government policies and responses of the society members to prevent and control the spread of COVID-19?** | |
| Yes | 2004 (79.8) |
| No | 508 (20.2) |

## Mental Health Outcomes

In total, 454 (18.1%) participants had scores of ≥15 on the PHQ-9, which indicates that they had symptoms of depression. Furthermore, 382 (15.2%) participants had scores of 20-27 on the PHQ-9, which indicated a high likelihood among these participants to experience symptoms of depression. In addition, 372 (14.8%) participants had scores of ≥15 on the 7-item Generalized Anxiety Disorder scale, which indicates that they had anxiety symptoms.

## Telehealth Usability

Approximately half of the survey respondents agreed on the usefulness of telehealth services during the COVID-19 pandemic. In addition, more than half of the study participants reported that it was easy to avail of telehealth services and trusted this type of health care service. More than half of the study participants (n=1435, 57.1%) reported that they felt comfortable seeking the telehealth services, and 1389 (55.3%)

reported that the system was easy to understand. Furthermore, 1354 (53.9%) participants reported that they could easily communicate with physicians through the telehealth system. However, only 1018 (40.5%) participants reported that communication with physicians was easier through the telehealth system than through other traditional methods. Regarding the reliability of the system, 1034 (41.2%) participants reported that the efficiency of the system was similar to that of in-person physician consultations, while 1175 (46.8%) participants responded with "neutral" to this statement. Additionally, 1439 (57.3%) participants felt comfortable communicating with physicians through the telehealth system, while 1545 (61.5%) reported that the system provides an effective means of communication when availing health care services. More than half of the participants (n=1383, 55.1%) felt satisfied with the telehealth system and were willing to use it in future. Table 4 provides an overview of the participants' perception of the usability of telehealth services during the COVID-19 pandemic.

**Table 4.** Usability of telehealth services among the study participants (N=2512).

| Component | Agree, n (%) | Neutral, n (%) | Disagree, n (%) |
|---|---|---|---|
| **Usefulness** | | | |
| The telehealth system has improved my access to health care services | 1134 (45.1) | 1058 (42.1) | 320 (12.7) |
| The telehealth system saves me the time I spend going to a hospital, clinic, or specialist | 1407 (56.0) | 959 (38.2) | 146 (5.8) |
| The telehealth system satisfies all of my health care needs | 1331 (53.0) | 978 (38.9) | 203 (8.1) |
| **Ease of use and learnability** | | | |
| The system is simple and easy to use | 1479 (58.9) | 877 (34.9) | 156 (6.2) |
| It was easy to learn how to use the system | 1348 (53.7) | 1024 (40.8) | 140 (5.6) |
| I trust that I can quickly get results with this system | 1367 (54.4) | 979 (39.0) | 166 (6.6) |
| **Interface quality** | | | |
| Handling this system is very comfortable to me | 1435 (57.1) | 959 (38.2) | 118 (4.7) |
| I like to use this system | 1224 (48.7) | 1073 (42.7) | 215 (8.6) |
| The system is simple and easy to understand | 1389 (55.3) | 974 (38.8) | 149 (5.9) |
| The system is able to perform all the tasks that I want it to perform | 1158 (46.1) | 1142 (45.5) | 212 (8.4) |
| **Interaction quality** | | | |
| I can easily talk to a physician when using the telehealth system | 1354 (53.9) | 942 (37.5) | 216 (8.6) |
| I can clearly hear the doctor when using the telehealth system | 1048 (41.7) | 1194 (47.5) | 270 (10.7) |
| I felt able to express myself effectively | 1129 (44.9) | 898 (35.7) | 485 (19.3) |
| Seeing a physician through the telehealth system was as easy as having an in-person consultation | 1018 (40.5) | 1247 (49.6) | 247 (9.8) |
| **Reliability** | | | |
| Physician consultations using the telehealth system were as efficient as in-person consultations | 1034 (41.2) | 1175 (46.8) | 303 (12.1) |
| If any error occurs when using the system, I can correct the error quickly and easily | 1225 (48.8) | 1072 (42.7) | 215 (8.6) |
| The system sends messages when errors occur, and that precisely guided me on how to correct errors | 1444 (57.5) | 902 (35.9) | 166 (6.6) |
| **Satisfaction and future use** | | | |
| I felt comfortable when communicating with a physician through the telehealth system | 1439 (57.3) | 855 (35.2) | 188 (7.5) |
| The telehealth system is an acceptable way to receive health care services | 1545 (61.5) | 785 (31.3) | 182 (7.2) |
| I will use the telehealth service again in the future | 1383 (55.1) | 949 (37.8) | 180 (7.2) |
| In general, I am completely satisfied with the telehealth system | 1451 (57.8) | 861 (34.3) | 200 (8.0) |

## Discussion

### Principal Findings

This study aimed to assess the accessibility, applicability, and feasibility of telehealth services in Libya during the COVID-19 pandemic. Most study participants reported being able to access health care facilities. However, most participants found it difficult to cover the cost of the health care services, and more than half of them avoided seeking these services owing to the difficulty in covering their costs. Our study participants reported high levels of telehealth usability and applicability. More than half of the participants (n=1435, 57.1%) reported feeling comfortable using telehealth services, 1389 (55.3%) reported that the system was easy to use, and 1354 (53.9%) reported an enhanced level of interaction with health care providers. Furthermore, 1018 (40.5%) participants reported that communication with physicians was better through the telehealth system than through traditional health care methods, while 1034 (41.2%) reported that the telehealth services were equally as efficient as in-person health care services. In addition, our participants experienced moderate levels of symptoms of anxiety (n=372, 14.8%) and depression (n=382, 15.2%).

Regarding the accessibility of the telehealth services, most participants did not report any difficulty in in reaching health care facilities or arranging an appointment with a specialist. However, the participants reported difficulties in accessing their laboratory findings and hospital records, which may be

influenced by the unavailability of electronic health records in most hospitals and the unavailability of laboratory facilities in several major hospitals, where patients have to visit another facility to obtain the results of laboratory tests.

## Comparison With Previous Studies

Our results suggest that the telehealth system potentially provides an alternative to traditional health care services, especially during the COVID-19 pandemic, concurrent with Mclean et al [24] who reported no difference in outcomes between traditional and telehealth services. Telehealth services offer more flexibility in securing appointments and are more comfortable for individuals living far from health care facilities, especially cancer patients or individuals with chronic diseases who might be at increased risk of severe COVID-19 [25,26]. However, further development of telehealth services is needed to supply the demand for several specialties that require a specific eHealth-based health care infrastructure [27]. A recent study from China [28] reported that telemedicine and telehealth services were effective and feasible and allowed for substantial improvement in health care services for patients with COVID-19. Another study from China [29] on the role of internet hospitals in the COVID-19 pandemic reported that telehealth services and internet hospitals can offer essential support in controlling the pandemic by reducing the cross-infection risk, thus curbing the spread of COVID-19. However, several barriers and challenges regarding the implementation of internet hospitals remain, and further studies with large samples are required to focus on testing specific prototype internet hospitals [30].

Another example of the efficiency of telehealth services is based on individuals with diabetes; telemedicine has been reported to substantially improve health-related outcomes among individuals with diabetes [31]. Another study that used big data and artificial intelligence approaches for diabetes care supported this finding [32].

Our study provides an overview of the mental health status of the Libyan population. Mental health services constitute a major application of telehealth, with several smartphone apps yielding positive outcomes during self-management of depression [33]. The demand for these interventions is high, given the current state of mental health during the COVID-19 pandemic, and telehealth facilitates follow-up evaluation of patients with psychiatric disorders without the need for hospital visits. On-demand mental health telehealth systems have shown promising outcomes in reducing the symptoms of depression [34].

Several challenges are associated with the implementation of telehealth services in Libya. The first issue relates to finances. The government and health care authorities need to provide support for telehealth services and provide funding opportunities to help new organizations and health care centers change their traditional approaches for care provision to a remote system to reduce the risk of COVID-19. This is critical because Libya is faced with a financial crisis, and it would be difficult for Libyan residents to cover the cost of such expensive health care services. Second, Libyan residents have been facing problems with electricity provision and internet access over the last few years.

It would be majorly challenging to implement these services during the civil war when internet connectivity is troublesome.

Another issue that needs to be considered in Libya is the protection of patient confidentiality and the ability to devise methods that protect the personal and clinical data of patients. Our study clearly highlights this issue; 1558 (62.0%) participants reported that they could not access their health reports. However, it is also concerning that several private telehealth services are providing care without any formal agreements or specific policies to protect patient data or let patients access their own health records. Therefore, further studies should address this issue and focus on data protection and on providing patients easy access to their health data.

## Strengths and Limitations

This study has several strengths. First, to our knowledge, this study is the first to validate the Telehealth Usability Questionnaire as an instrument in Arabic. Second, this study is the first to provide an overview of the usability of telehealth and the health status of the Libyan population. Third, our study included a large sample of 2512 individuals with complete data and assessed several health-related issues with a special emphasis on accessibility and the mental health status of the Libyan population. Furthermore, to our knowledge, this study is the first to assess the usability of telehealth services during the COVID-19 pandemic in a transitional, resource-limited country. Our study furthers the current understanding of the usability of telemedicine in transitional, resource-limited countries, and our results support the implementation of telehealth programs despite certain potential challenges associated with them.

The study has several limitations that should be addressed in future studies. First, it was a cross-sectional survey carried out during the COVID-19 pandemic simultaneously with the introduction of telehealth services, and some participants may not have used telehealth services during this period. Second, we did not assess internet connectivity and electricity provision, which are potential factors affecting the ability of some communities to access telehealth services; these factors need to be addressed in future studies. Third, this study was an online cross-sectional survey, and most participants were young and without major comorbidities or chronic diseases; therefore, we may have overestimated the usability of these services because older individuals may face more challenges in using these technologies and may require more assistance. Therefore, further studies are required to assess elderly individuals and those with special needs and to investigate how telehealth systems could be implemented for these specific groups of patients. Fourth, we used self-reported scales, which may have increased the risk of bias. We also used a 3-point Likert scale for the Telehealth Usability Questionnaire; this may have influenced our findings.

## Conclusions

Telehealth services can substantially replace in-person consultations and prevent nosocomial infections in Libya during the COVID-19 pandemic. Our study provides an overview of the accessibility and usability of telehealth services, and our results suggest that telehealth services help reduce the workload

of physicians and direct contact with patients during the COVID-19 pandemic. Our results support the implementation of telehealth services; however, further studies are required to focus on the protection of patient confidentiality and personal data.

## Acknowledgments

## Authors' Contributions

ME analyzed and interpreted the data and drafted the manuscript. All authors helped design the study, collect the data, and write the manuscript. All authors read and approved the final manuscript.

## Conflicts of Interest

None declared.

## References

1. Zhou P, Yang X, Wang X, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020 Mar;579(7798):270-273 [FREE Full text] [doi: 10.1038/s41586-020-2012-7] [Medline: 32015507]

2. Guo Y, Cao Q, Hong Z, Tan Y, Chen S, Jin H, et al. The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak - an update on the status. Mil Med Res 2020 Mar 13;7(1):11 [FREE Full text] [doi: 10.1186/s40779-020-00240-0] [Medline: 32169119]

3. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020 May;20(5):533-534 [FREE Full text] [doi: 10.1016/S1473-3099(20)30120-1] [Medline: 32087114]

4. Guan W, Ni Z, Hu Y, Liang W, Ou C, He J, China Medical Treatment Expert Group for Covid-19. Clinical Characteristics of Coronavirus Disease 2019 in China. N Engl J Med 2020 Apr 30;382(18):1708-1720 [FREE Full text] [doi: 10.1056/NEJMoa2002032] [Medline: 32109013]

5. COVIDSurg Collaborative. Elective surgery cancellations due to the COVID-19 pandemic: global predictive modelling to inform surgical recovery plans. Br J Surg 2020 Oct;107(11):1440-1449 [FREE Full text] [doi: 10.1002/bjs.11746] [Medline: 32395848]

6. Truog RD, Mitchell C, Daley GQ. The Toughest Triage - Allocating Ventilators in a Pandemic. N Engl J Med 2020 May 21;382(21):1973-1975. [doi: 10.1056/NEJMp2005689] [Medline: 32202721]

7. Mehrotra A, Ray K, Brockmeyer DM, Barnett ML, Bender JA. Rapidly Converting to "Virtual Practices": Outpatient Care in the Era of Covid-19. NEJM Catalyst 2020;1(2) [FREE Full text] [doi: 10.1056/CAT.20.0091]

8. Jones J, Sullivan PS, Sanchez TH, Guest JL, Hall EW, Luisi N, et al. Similarities and Differences in COVID-19 Awareness, Concern, and Symptoms by Race and Ethnicity in the United States: Cross-Sectional Survey. J Med Internet Res 2020 Jul 10;22(7):e20001 [FREE Full text] [doi: 10.2196/20001] [Medline: 32614778]

9. Vandekerckhove P, Vandekerckhove Y, Tavernier R, De Jaegher K, de Mul M. Leveraging User Experience to Improve Video Consultations in a Cardiology Practice During the COVID-19 Pandemic: Initial Insights. J Med Internet Res 2020 Jun 25;22(6):e19771 [FREE Full text] [doi: 10.2196/19771] [Medline: 32519964]

10. Lin C, Tseng W, Wu J, Tay J, Cheng M, Ong H, et al. A Double Triage and Telemedicine Protocol to Optimize Infection Control in an Emergency Department in Taiwan During the COVID-19 Pandemic: Retrospective Feasibility Study. J Med Internet Res 2020 Jun 23;22(6):e20586 [FREE Full text] [doi: 10.2196/20586] [Medline: 32544072]

11. Fisk M, Livingstone A, Pit SW. Telehealth in the Context of COVID-19: Changing Perspectives in Australia, the United Kingdom, and the United States. J Med Internet Res 2020 Jun 09;22(6):e19264 [FREE Full text] [doi: 10.2196/19264] [Medline: 32463377]

12. Author. Telemedicine: fad or future? Lancet 1995 Jan 14;345(8942):73-74. [doi: 10.1016/S0140-6736(95)90054-3] [Medline: 7815881]

13. Ohannessian R, Duong TA, Odone A. Global Telemedicine Implementation and Integration Within Health Systems to Fight the COVID-19 Pandemic: A Call to Action. JMIR Public Health Surveill 2020 Apr 02;6(2):e18810 [FREE Full text] [doi: 10.2196/18810] [Medline: 32238336]

14. Dorsey ER, Topol EJ. State of Telehealth. N Engl J Med 2016 Jul 14;375(2):154-161. [doi: 10.1056/NEJMra1601705] [Medline: 27410924]

15. Elhadi M, Msherghi A, Alkeelani M, Alsuyihili A, Khaled A, Buzreg A, et al. Concerns for low-resource countries, with under-prepared intensive care units, facing the COVID-19 pandemic. Infect Dis Health 2020 Nov;25(4):227-232 [FREE Full text] [doi: 10.1016/j.idh.2020.05.008] [Medline: 32631682]

16.    Elhadi M, Msherghi A, Alkeelani M, Zorgani A, Zaid A, Alsuyihili A, et al. Assessment of Healthcare Workers' Levels of
       Preparedness and Awareness Regarding COVID-19 Infection in Low-Resource Settings. Am J Trop Med Hyg 2020
       Aug;103(2):828-833 [FREE Full text] [doi: 10.4269/ajtmh.20-0330] [Medline: 32563273]

17.    Elhadi M, Msherghi A. COVID-19 and civil war in Libya: the current situation. Pathog Glob Health 2020 Jul;114(5):230-231.
       [doi: 10.1080/15575330.2020.1769292] [Medline: 32432976]

18.    Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES).
       J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]

19.    Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001
       Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]

20.    Levis BBTB, Benedetti A, Thombs BD, DEPRESsion Screening Data (DEPRESSD) Collaboration. Accuracy of Patient
       Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. BMJ
       2019 Apr 09;365:l1476 [FREE Full text] [doi: 10.1136/bmj.l1476] [Medline: 30967483]

21.    Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7.
       Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]

22.    Parmanto B, Lewis AN, Graham KM, Bertolet MH. Development of the Telehealth Usability Questionnaire (TUQ). Int J
       Telerehabil 2016;8(1):3-10 [FREE Full text] [doi: 10.5195/ijt.2016.6196] [Medline: 27563386]

23.    Zhou L, Bao J, Setiawan IMA, Saptono A, Parmanto B. The mHealth App Usability Questionnaire (MAUQ): Development
       and Validation Study. JMIR Mhealth Uhealth 2019 Apr 11;7(4):e11500 [FREE Full text] [doi: 10.2196/11500] [Medline:
       30973342]

24.    McLean S, Sheikh A, Cresswell K, Nurmatov U, Mukherjee M, Hemmi A, et al. The impact of telehealthcare on the quality
       and safety of care: a systematic overview. PLoS One 2013;8(8):e71238 [FREE Full text] [doi: 10.1371/journal.pone.0071238]
       [Medline: 23977001]

25.    Tashkandi E, Zeeneldin A, AlAbdulwahab A, Elemam O, Elsamany S, Jastaniah W, et al. Virtual Management of Patients
       With Cancer During the COVID-19 Pandemic: Web-Based Questionnaire Study. J Med Internet Res 2020 Jun
       24;22(6):e19691 [FREE Full text] [doi: 10.2196/19691] [Medline: 32501807]

26.    Khairat S, Meng C, Xu Y, Edson B, Gianforcaro R. Interpreting COVID-19 and Virtual Care Trends: Cohort Study. JMIR
       Public Health Surveill 2020 Apr 15;6(2):e18811 [FREE Full text] [doi: 10.2196/18811] [Medline: 32252023]

27.    Cruz-Martínez RR, Wentzel J, Asbjørnsen RA, Noort PD, van Niekerk JM, Sanderman R, et al. Supporting Self-Management
       of Cardiovascular Diseases Through Remote Monitoring Technologies: Metaethnography Review of Frameworks, Models,
       and Theories Used in Research and Development. J Med Internet Res 2020 May 21;22(5):e16157 [FREE Full text] [doi:
       10.2196/16157] [Medline: 32436852]

28.    Hong Z, Li N, Li D, Li J, Li B, Xiong W, et al. Telemedicine During the COVID-19 Pandemic: Experiences From Western
       China. J Med Internet Res 2020 May 08;22(5):e19577 [FREE Full text] [doi: 10.2196/19577] [Medline: 32349962]

29.    Gong K, Xu Z, Cai Z, Chen Y, Wang Z. Internet Hospitals Help Prevent and Control the Epidemic of COVID-19 in China:
       Multicenter User Profiling Study. J Med Internet Res 2020 Apr 14;22(4):e18908 [FREE Full text] [doi: 10.2196/18908]
       [Medline: 32250962]

30.    Li P, Luo Y, Yu X, Wen J, Mason E, Li W, et al. Patients' Perceptions of Barriers and Facilitators to the Adoption of
       E-Hospitals: Cross-Sectional Study in Western China. J Med Internet Res 2020 Jun 11;22(6):e17221 [FREE Full text] [doi:
       10.2196/17221] [Medline: 32525483]

31.    Timpel P, Oswald S, Schwarz PEH, Harst L. Mapping the Evidence on the Effectiveness of Telemedicine Interventions in
       Diabetes, Dyslipidemia, and Hypertension: An Umbrella Review of Systematic Reviews and Meta-Analyses. J Med Internet
       Res 2020 Mar 18;22(3):e16791 [FREE Full text] [doi: 10.2196/16791] [Medline: 32186516]

32.    Musacchio N, Giancaterini A, Guaita G, Ozzello A, Pellegrini MA, Ponzani P, et al. Artificial Intelligence and Big Data
       in Diabetes Care: A Position Statement of the Italian Association of Medical Diabetologists. J Med Internet Res 2020 Jun
       22;22(6):e16922 [FREE Full text] [doi: 10.2196/16922] [Medline: 32568088]

33.    Firth J, Torous J, Nicholas J, Carney R, Pratap A, Rosenbaum S, et al. The efficacy of smartphone-based mental health
       interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry 2017
       Oct;16(3):287-298 [FREE Full text] [doi: 10.1002/wps.20472] [Medline: 28941113]

34.    Kunkle S, Yip M, Ξ W, Hunt J. Evaluation of an On-Demand Mental Health System for Depression Symptoms: Retrospective
       Observational Study. J Med Internet Res 2020 Jun 18;22(6):e17902 [FREE Full text] [doi: 10.2196/17902] [Medline:
       32554387]

## Abbreviations

**PHQ-9:** 9-item Patient Health Questionnaire

Original Paper

# An Automated Patient Self-Monitoring System to Reduce Health Care System Burden During the COVID-19 Pandemic in Malaysia: Development and Implementation Study

Hooi Min Lim[1,2*], MBBS, MMed; Chin Hai Teo[2,3*], BMedImag, PhD; Chirk Jenn Ng[2,4*], MBBS, MMed, PhD; Thiam Kian Chiew[2,5*], BCompSc, MCompSc, PhD; Wei Leik Ng[4*], MBBS, MMed; Adina Abdullah[2,4*], BMBS, BMedSc, MMed, PhD; Haireen Abdul Hadi[4*], MB BCH BAO, MMed; Chee Sun Liew[2,6*], BCompSc, MSc, PhD; Chee Seng Chan[2,7*], BEng, MSc, PhD

[1]Department of Primary Care Medicine, University of Malaya Medical Centre, Kuala Lumpur, Malaysia

[2]eHealth Unit, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia

[3]Dean's Office, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia

[4]Department of Primary Care Medicine, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia

[5]Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia

[6]Department of Computer System and Technology, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia

[7]Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia

[*]all authors contributed equally

**Corresponding Author:**
Chirk Jenn Ng, MBBS, MMed, PhD
eHealth Unit
Faculty of Medicine
University of Malaya
Kuala Lumpur, 50603
Malaysia
Phone: 60 142204126
Email: ngcj@um.edu.my

## Abstract

**Background:** During the COVID-19 pandemic, there was an urgent need to develop an automated COVID-19 symptom monitoring system to reduce the burden on the health care system and to provide better self-monitoring at home.

**Objective:** This paper aimed to describe the development process of the COVID-19 Symptom Monitoring System (CoSMoS), which consists of a self-monitoring, algorithm-based Telegram bot and a teleconsultation system. We describe all the essential steps from the clinical perspective and our technical approach in designing, developing, and integrating the system into clinical practice during the COVID-19 pandemic as well as lessons learned from this development process.

**Methods:** CoSMoS was developed in three phases: (1) requirement formation to identify clinical problems and to draft the clinical algorithm, (2) development testing iteration using the agile software development method, and (3) integration into clinical practice to design an effective clinical workflow using repeated simulations and role-playing.

**Results:** We completed the development of CoSMoS in 19 days. In Phase 1 (ie, requirement formation), we identified three main functions: a daily automated reminder system for patients to self-check their symptoms, a safe patient risk assessment to guide patients in clinical decision making, and an active telemonitoring system with real-time phone consultations. The system architecture of CoSMoS involved five components: Telegram instant messaging, a clinician dashboard, system administration (ie, back end), a database, and development and operations infrastructure. The integration of CoSMoS into clinical practice involved the consideration of COVID-19 infectivity and patient safety.

**Conclusions:** This study demonstrated that developing a COVID-19 symptom monitoring system within a short time during a pandemic is feasible using the agile development method. Time factors and communication between the technical and clinical teams were the main challenges in the development process. The development process and lessons learned from this study can guide the future development of digital monitoring systems during the next pandemic, especially in developing countries.

XSL•FO

**RenderX**

## Introduction

To combat the COVID-19 pandemic, digital technology has been used extensively for health information dissemination, contact tracing [1], population surveillance, and forecasting modeling [2]. It provides a platform for real-time updates of COVID-19 cases, performs modeling studies to forecast COVID-19 disease activity, and disseminates public health education [3]. However, there are issues related to technology use during a pandemic; there are occasions when urgency supersedes accuracy, which may compromise patient and public safety. The accuracy and security of a technology must be considered carefully and scrutinized systematically to ensure patients' safety. Also, false information may cause unnecessary alarm and chaos in the population or may fail in the detection or monitoring of those at risk of infection [4]. Privacy and data protection concerns have been raised with large-scale digital data collection during the pandemic [5]. Therefore, there is a need for research, risk assessments, and pilot studies before technologies roll out to the public in order to avoid harms.

During the COVID-19 pandemic, several COVID-19 remote monitoring tools were developed and rolled out to reduce the burden on health care systems and to free up hospital spaces. In France, Covidom, a web-based application, was used as a remote surveillance tool for COVID-19 patients with mild to moderate symptoms in order to preserve medical resources for more severe patients and limit in-person interactions during the pandemic [6]. COVID-19 symptom monitoring tools could potentially reduce the burden on the health care system. A practical and systematic structure is required to link digital monitoring tools with teleconsultation and local health care service systems. However, a detailed description of the developmental process and practical operational steps in developing a digital monitoring system for urgent use in a pandemic was lacking in the literature.

COVID-19 cases in Malaysia have increased since March 2020, with a total number of 70,236 confirmed COVID-19 cases as of December 5, 2020. It is estimated that for every positive COVID-19 patient diagnosed, there are 16 other suspected COVID-19 patients (ie, patients under investigation [PUIs]) and asymptomatic close contacts to be taken care of in Malaysia. These patients were quarantined and required daily follow-up [7]. In Malaysia, PUIs and close contacts of a confirmed case, who was fit to be discharged home, were given a home surveillance tool that included a checkbox list of COVID-19 symptoms [8]. Public health officers from the district health offices would call each patient via phone to monitor their symptoms daily for up to 14 days. This manual surveillance work is time-consuming and labor-intensive. Hence, there is an urgent need to use digital technology to monitor PUIs and close contacts during COVID-19 to reduce the burden on the health care system in Malaysia.

Conventionally, digital health systems are developed based on software development life cycle models, such as the waterfall, spiral, V-shaped, and rapid application development models [9]. In developing our COVID-19 Symptom Monitoring System (CoSMoS), an automated self-monitoring system that is the focus of this paper, the agile model was preferred because it uses an adaptive approach that easily adapts to changing requirements. This characteristic is important because the guidelines and clinical evidence of COVID-19 were changing rapidly, especially in the early phase of the pandemic, which would change the requirements, expressed as user stories, during the development. The agile method is more appropriate and feasible for developing a digital health app within a short period during a pandemic.

The development of CoSMoS was a time-intensive process requiring commitment on the part of the experts; in this context, that includes the expert in digital software development and the health care professionals who were actively involved in providing care during the pandemic. This paper aims to describe the development process of a COVID-19 monitoring system, CoSMoS, using the agile model within a short period during the COVID-19 pandemic as well as the lessons learned.

## Methods

### Overarching Development Framework and Phases

The agile software development life cycle was used to develop CoSMoS because it was urgently needed for patients suspected of having COVID-19. The development of CoSMoS was divided into three phases: requirement formation, development testing iteration, and integration into clinical practice. All requirements, development, and testing work were conducted online via Zoom, WhatsApp, Slack, Trello, and email. No physical meetings were held due to the COVID-19 movement control order (ie, lockdown) in Malaysia. The whole development process took 19 days, from study inception on March 21, 2020, to the launch of CoSMoS on April 9, 2020. The detailed timeline is shown in Multimedia Appendix 1. This study was approved by the Medical Research Ethics Committee of the University of Malaya Medical Centre (UMMC) (MECID No. 202043-8434).

### Development and Research Team

An interdisciplinary team-based approach, which has been commonly used for mobile health solutions in the literature, was adapted for the development of CoSMoS [10]. The core team, which was in charge of the overall study inception, development, and execution, comprised primary care physicians (n=6), computer science academicians (n=3), a patient advocate (n=1), and an eHealth research fellow (n=1). The clinical team (ie, six primary care physicians and an eHealth research fellow) was responsible for the content of CoSMoS and the integration of CoSMoS into clinical practice, including the workflow, monitoring, and teleconsultation. The technical team, which

was in charge of the system development, consisted of the three computer science academicians, an eHealth research fellow, 26 students and graduates from the Faculty of Computer Science and Information Technology, and one user experience designer. The technical development process was led by an associate professor in computer science, who is an expert in software development and the agile model; this team member oversaw the development process; designed the system architecture, task, and workforce management; and supervised the developers. A high number of technical team members volunteered to develop CoSMoS so they could play a part contributing to combating the pandemic. The eHealth research fellow straddled across all teams and acted as the project coordinator between the clinical and technical teams. There were two postgraduate clinical master's students in primary care medicine involved in integrating CoSMoS into clinical practice.

## Phase 1: Requirement Formation

### Problem Identification and Requirements

The core team first met via Zoom to explore the problems in depth. This discussion was led by the clinical team, which consisted of a panel of seven experts in primary care health services, infectious disease, and health informatics. The panel obtained expert input from the district health officials and infectious control team from UMMC to have a clearer picture of the current COVID-19 situation.

The objectives of this discussion were to (1) identify the major problems in the process of monitoring PUIs and close contacts at home and the limitations of the existing system, (2) identify areas in which digital technology could help to reduce the burden on health care workers (HCWs) and provide a better home monitoring service, and (3) identify the target users of CoSMoS.

The problems identified were then converted to the requirements of the system, including the functions and features. The integration of CoSMoS into the existing clinical workflow was also discussed.

### Algorithm and Content Drafting

With the list of requirements at hand, the clinical team started to draft the content and algorithm of CoSMoS. The clinical expert panel developed the clinical decision-making algorithm based on the World Health Organization and the Malaysian Ministry of Health guidelines of COVID-19 monitoring [7,8]. A decision tree method was used to develop the clinical algorithm, establish whether a patient was safe to continue monitoring at home, and establish whether a patient should call HCWs for teleconsultation [11]. The panel selected the variables to be included in the decision tree and assessed the importance of each variable in clinical decision making. We also considered the selection of clinical relevance variables when developing the algorithm. After the variables were selected, a decision node was determined, representing a choice that would divide the patients into asymptomatic or symptomatic. An *if-then-else* rule was used to create a decision tree pathway. The clinical team tested the algorithm using different clinical scenarios and revised the algorithm iteratively.

## Phase 2: Development Testing Iteration

### Overview

The requirements and the contents drafted were presented to the technical team. The technical team explored potential solutions to achieve the requirements. The technical team reviewed several software development platforms and tools and made selections based on their suitability to deliver the solution; availability of development resources, including the team's existing skill sets; security; affordability for the end users; and cost.

### Continuous Integration and Continuous Deployment

The development of the CoSMoS system involved several short cycles comprising the following activities: elicit requirements, develop and test, review and feedback, and revise requirements. The continuous integration and continuous deployment (CI/CD) pipeline was used to enable swift flow from one activity to the next within a cycle, uninterrupted integration of deliverables (ie, software codes) of a cycle with deliverables of previous cycles, continuous testing by the core team, and deployment of the tested system into the production environment. CI/CD has two environments: staging (ie, nonproduction) and production. Any changes requested are first pushed to the staging environment for testing. Once the revision has been tested and finalized, it is swiftly pushed to the production environment.

### Iterative Development Testing Cycles

Among the technical team members, there were nine developers focusing on testing before CoSMoS was tested by the core team members in a daily manner. The core team members reviewed CoSMoS (ie, the system's Telegram bot and dashboard) in terms of its utility (eg, accuracy of the algorithm, the wording of the content, timing of daily ping messages, and correctness of patient categorization in the dashboard) and usability (eg, button size and position, layout, and the flow of content). The change requests with screenshots were compiled in a Word document and sent to the technical team, who would implement the changes in the evening. Managing change requests was crucial in ensuring that the requirements of CoSMoS were properly documented and updated.

## Phase 3: Integration Into Clinical Practice

CoSMoS was presented to the stakeholders, including the primary care clinic management team and clinicians, the hospital's COVID-19 task force committee, and the emergency department. Live demonstrations of the system were used during these presentations, and the stakeholders could test out the CoSMoS Telegram bot. Feedback from the stakeholders on the utility, usability, and feasibility of the system was collected to further improve the system. Medicolegal aspects of the clinical implication and teleconsultation were discussed.

Repeated simulations with different clinical workflows were tested to identify the most effective and suitable clinical workflow for CoSMoS integration. Role-playing with clinic staff was done to test out the CoSMoS workflow in the actual clinic setting. Barriers and weaknesses in the workflow were identified. Careful consideration was done to reduce the physical contact between HCWs and patients with suspected COVID-19.

We finalized the clinical workflow for the integration of CoSMoS after several rounds of testing. Training was conducted for all clinic staff, including the system's objectives, functions, features, and operational guides; how to integrate CoSMoS into the existing workflow; the teleconsultation guide; and possible issues. Simulations were done during training to ensure familiarity with the new workflow. We installed the CoSMoS dashboard onto the desktop computers in the clinic. The doctors used designated smartphones with the CoSMoS Telegram bot installed for monitoring and teleconsultation. A technical help desk was set up to assist the doctors should any technical issues arise.

Records of patients recruited for CoSMoS will be kept in the hospital electronic medical record system and the CoSMoS system. These data will be kept confidential. Only the researchers, research assistants, and CoSMoS doctors have access to the medical record system, using designated log-in credentials. Data are transmitted over the https secure communication protocol and stored on cloud storage with adequate security protection.

## Results

### Phase 1: Requirement Formation

#### *Problem Identification and Requirements*

The expert panel identified the target users and sites for CoSMoS. The target users were the PUIs and close contacts who attended the primary care clinic and emergency department in the UMMC and the health care providers who used CoSMoS to monitor the patients. The expert panel identified four main problems in the existing home monitoring system for PUIs and close contacts that could be solved by digital innovation (see Table 1).

**Table 1.** Problems identified with the existing home monitoring system and proposed requirements for the COVID-19 Symptom Monitoring System (CoSMoS).

| Problems with the existing home monitoring system | Requirements for CoSMoS |
| --- | --- |
| High workload for the district health officials in calling the patients to assess their daily symptoms | A daily automated messaging and reminder system sends active reminders for patients to key in their symptoms every day |
| Issues of surveillance using phone calls (ie, patients did not pick up the calls, patients worried about scam calls, and long phone conversations to assess each symptom) | Development of a simple system that is easy to use, secure, interactive, and trustworthy |
| Lack of coordinating care and continuity of care for the patients under investigation and close contacts, especially coordinating care between COVID-19 centers and district health officials | Development of a system to monitor patients with continuity of care |
| Unnecessary clinic or emergency department visits due to change of symptoms or administrative issues (ie, letter of quarantine) expose the patients and health care workers (HCWs) to COVID-19 infection risk | An automated system performs patients' risk assessments safely and guides patients in clinical decision making, whether to continue home monitoring or to seek medical attention; a dashboard system provides active monitoring and facilitates the HCW in identifying patients who require teleconsultation and unreported patients |

Besides the three main functionalities, the additional requirements proposed included (1) materials to be sent to patients (ie, privacy policy, user guide, and educational materials), (2) additional notes in the dashboard so that CoSMoS can be used independently without relying on electronic medical records, and (3) a data export function in the dashboard for external data analysis. A storyboard on how CoSMoS will be used from enrollment to completion of monitoring was used to illustrate the system more clearly and thoroughly (see Figure 1).

**Figure 1.** COVID-19 Symptom Monitoring System (CoSMoS) storyboard. DR: doctor; HCW: health care worker; PT: patient; UMMC: University of Malaya Medical Centre.
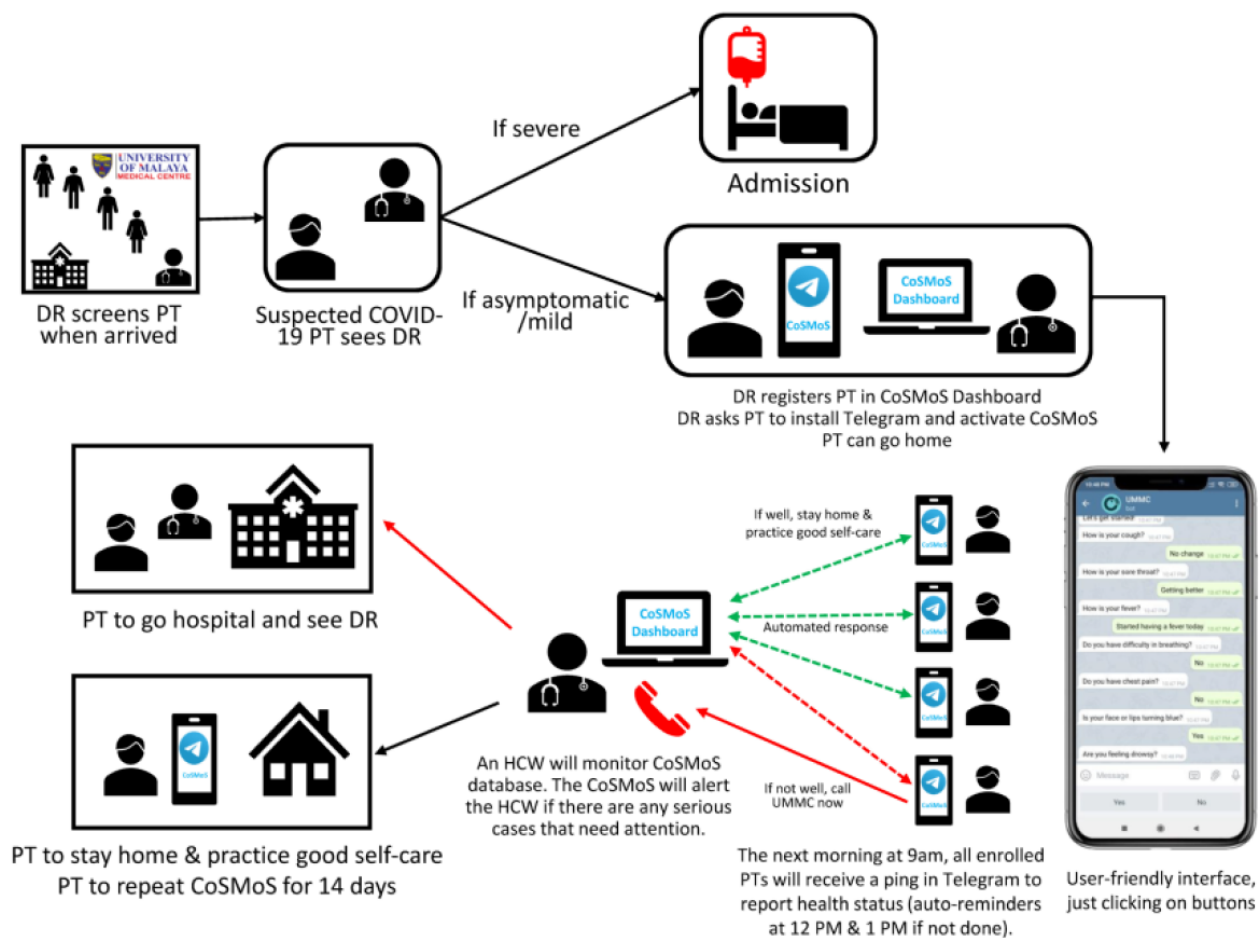


### Algorithm and Content Drafting

A total of 13 iterative cycles, including major and minor changes, were done to refine the clinical algorithm (see Multimedia Appendix 2). The text for questions and answer options were refined so that they were easily understandable by the layperson. An end message would be sent to patients automatically based on the algorithm to inform patients whether they could stay at home or call the HCW for teleconsultation. A safety net message advised the patients to attend nearby health care facilities if they were unable to reach the HCW for teleconsultation. The algorithm had a lower threshold to prompt a patient to call the HCW because patient safety was taken into significant consideration. A free-text remark field was created as another safety net measure for the patients. The algorithm was translated into Malay and Chinese languages to suit local needs.

### Phase 2: Development Testing Iteration

### Development Platform

The technical team first proposed to employ WhatsApp as the platform to deliver the symptom assessment function, as WhatsApp is widely used in Malaysia. However, the plan was halted, as obtaining the WhatsApp Business application programming interface required a long approval time. The team then proceeded with Telegram, as it could provide the solution without complicated approval.

CoSMoS is hosted on the DigitalOcean cloud infrastructure. GitHub was used for code repository, Trello was used for task management, Slack was used for discussions on technical issues, and WhatsApp was used for communication between core and technical teams. Development tools and languages included the React JavaScript library for the clinician dashboard and Go for the back end. Firebase was first chosen for the database and Algolia for the search engine, but it was switched to MongoDB due to performance and cost considerations. Docker Hub was used for container images, and Sketch software was used to create CoSMoS's logo. In addition, Freshdesk was used to set up a technical help desk, through which the technical team received issues and reports from the clinical team. Kibana was used for logging server activities and for auditing purposes. The CI/CD pipeline enabled the seamless integration of multiple tools used in the development CoSMoS.

### System Architecture

The architecture of CoSMoS is shown in Figure 2. From the user's end, the Telegram bot retrieves PUIs' records from the database through the back end. Ping messages are sent to the PUIs at 9 AM, 12 PM, and 1 PM daily to remind the patients to report their symptoms. The PUIs use the Telegram bot to

report their daily symptoms. The symptoms are sent to the back end before being stored in the database; the HCWs monitor PUIs' symptoms through the clinician dashboard at 2 PM every day. PUIs' daily statuses are grouped into four main categories:

1. Category A. Reported: unstable and called the HCW.
2. Category B. Reported: unstable, but yet to call the HCW.
3. Category C. Reported: stable.
4. Category D. Not reported.

**Figure 2.** System architecture of the COVID-19 Symptom Monitoring System (CoSMoS). HCW: health care worker.
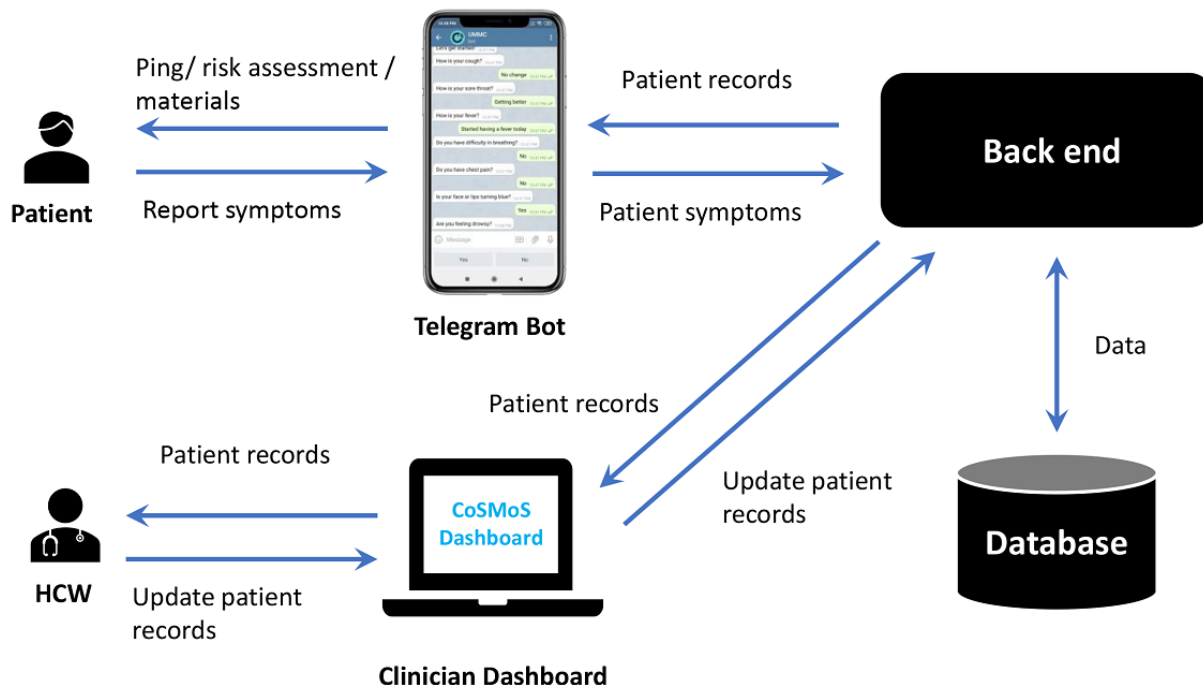


For category A, the patient would be assessed immediately over the phone (ie, teleconsultation) as they called the HCW. For categories B and D, the HCW would contact patients at 2 PM during the clinical review. There is no need to contact patients for category C unless the HCW identified serious comments in the open-ended remark box during the review. The HCWs may update the PUIs' records, such as changing a PUI's status from *asymptomatic* to *symptomatic* or *completed monitoring*, entering a PUI's swab test result, and adding a note to a PUI's record. The HCWs can also export PUIs' records in comma-separated vector format. Retrieving, updating, and storing PUIs' records between the dashboard and database would also be done through the back end, preserving data security and integrity with mechanisms such as authentication, authorization, and validation.

### *Iterative Development Testing Cycles*

There were many cycles of specify-develop-test-revise of CoSMoS, as this study used the agile development method. The tester team, which was independent from the developers, documented test cases that were used and reused to ensure quality and efficiency of testing; see example test cases in

Multimedia Appendix 3. The core team continuously gave feedback to the technical team after testing new deployments to the staging environment daily. However, there were three main prototype checkpoints—CoSMoS prototype 1, CoSMoS prototype 2, and finalized CoSMoS—to present major changes made to the system. Thus, it can be considered that the development process comprised three major agile cycles.

The clinicians were the main help desk contact people when patients had any technical issues. Patients could call the hotline to the clinician directly for technical issues. If the clinicians could not solve an issue, the technical issue would be channeled to the technical help desk system.

### Phase 3: Integration Into Clinical Practice

Textbox 1 shows the technical and clinical implementation barriers identified during the simulation and role-playing. The workflow was improved by introducing new strategies, such as refining the patients' recruitment criteria for CoSMoS and creating a communication script for HCWs to explain CoSMoS to patients. We also prepared a patient's and doctor's guide to CoSMoS and Telegram bot installation.

**Textbox 1.** Technical and clinical implementation barriers of the COVID-19 Symptom Monitoring System (CoSMoS).

---

**Technical barriers:**

- Poor internet connectivity

- Insufficient digital storage to download the Telegram app by suspected COVID-19 patients

- Inadequate mobile data quota to download an app

**Clinical implementation barriers:**

- Ineffective communication between health care workers (HCWs) and patients (ie, enrollment in CoSMoS)

- CoSMoS prolonged the consultation time (ie, longer contact time between HCWs and suspected COVID-19 patients) for enrollment and the installation of CoSMoS

- Patient factors (ie, unfamiliarity with digital technology)

---

Because of the high infectivity of COVID-19, necessary precautions were taken to ensure the safety of both patients and HCWs. Patient enrollment procedures, informed consent–taking, and CoSMoS installation were done remotely by a CoSMoS doctor via phone call to reduce the physical contact time. Figure 3 shows the clinical workflow of integrating CoSMoS into

clinical practice. The teleconsultation of CoSMoS was done following the Malaysian Medical Council advisory on virtual consultation [12]. The CoSMoS app served as a complementary service to the existing manual home monitoring system by the public health offices, not as a replacement, to ensure that the safety of patients was not compromised.

**Figure 3.** Integration of the COVID-19 Symptom Monitoring System (CoSMoS) into the clinical setting. ED: emergency department; PPE: personal protective equipment.
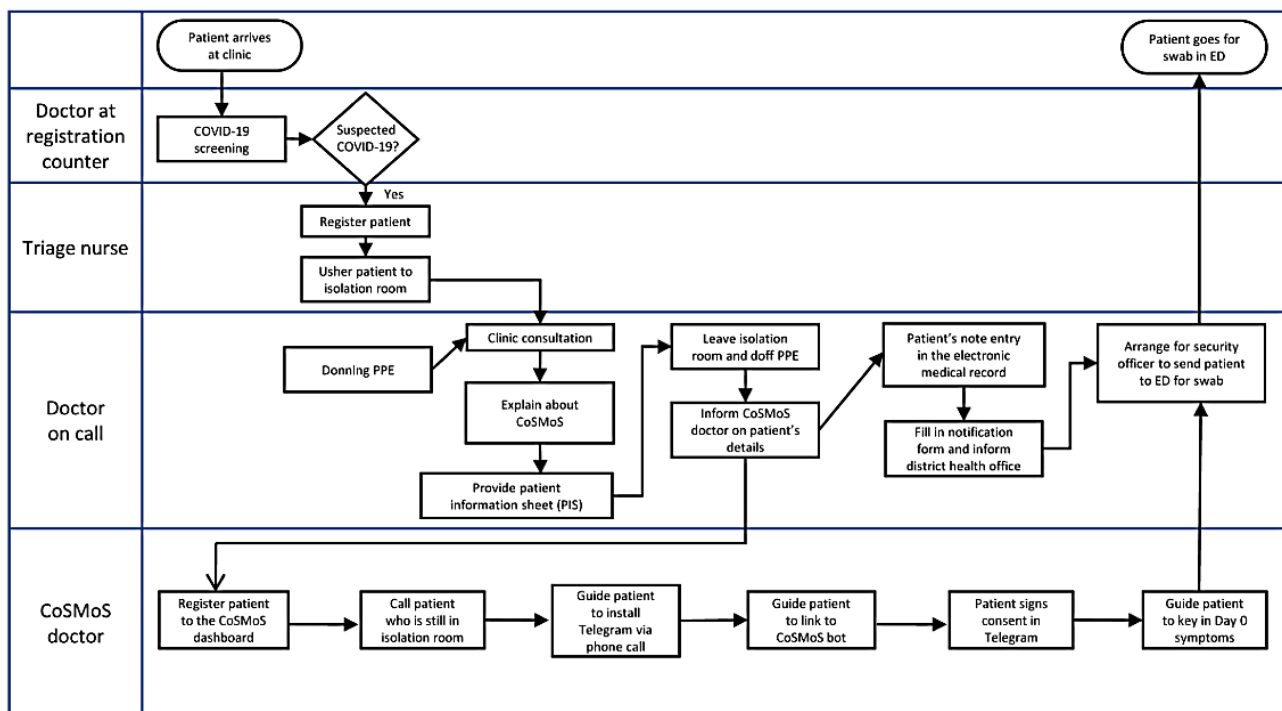


## Discussion

### Principal Findings

This paper presents the agile development process of a digital health innovation during the COVID-19 pandemic within a short period. CoSMoS was the first automated monitoring system aimed at the Malaysian population to provide home monitoring and phone consultation for patients with suspected COVID-19. There was an urgent need for CoSMoS to reduce the burden on the health care system while providing a better home monitoring service to patients in Malaysia. This paper reports on a systematic approach to its development involving

clinical experts, evidence-based clinical guidelines, data security and privacy considerations, and real-life clinical integration of an eHealth app during a pandemic.

Several COVID-19 symptom monitoring systems were developed and deployed in developed countries, such as France, China, and Korea [6,13,14]. Covidom, a web-based application developed in France, was used on a large scale for home monitoring of patients with COVID-19 [6]. A similar approach was used in China, where the telemedicine system was developed based on a popular social media platform, WeChat, to monitor those who needed to be quarantined at home [13]. In this study, the CoSMoS symptom monitoring bot was

delivered via the Telegram app; this app is commonly used among Malaysians, as government ministries and agencies disseminate COVID-19 news and health awareness via Telegram. Using the existing digital platform would be more user friendly for patients as they are familiar with the app. In the Korean military hospitals, a similar COVID-19 symptom monitoring app was developed with an additional model prediction programming interface to predict whether a patient requires attention from health care professionals [14]. This function was not developed in CoSMoS because of limited time, data, and resources required to accurately predict patient outcomes. The automated CoSMoS was useful in alleviating the burden on the health care system, reducing physical contact, and reducing the risk of delayed hospitalization [6,13]. The detailed developmental process described in this paper could be transposed to other developing countries where resources are limited to engaging commercial developers. Using existing local expertise and available resources, developing a digital solution for a COVID-19 symptom monitoring system is feasible and applicable for urgent use in the local setting.

The time factor was the most critical challenge to the development of CoSMoS because an automated COVID-19 home monitoring system was urgently needed in the health care system to combat the COVID-19 pandemic. Compared to the usual developmental process of a monitoring system for chronic diseases [12,15,16], conducting a detailed patient needs assessment was not feasible during the COVID-19 pandemic, due to Malaysia's movement control restriction situation. Hence, the initial requirement formation from Phase 1 was done with input from clinical experts who would be the users for the system and with a review of the existing guidelines for COVID-19 home monitoring. We were able to develop CoSMoS in 19 days because many experienced developers were organized into specialized teams according to their expertise (ie, front end, back end, Telegram bot, and infrastructure) and were led by computer science academicians who worked long hours. Each development task was systematically delegated to the designated team using a task management tool (ie, Trello). Each task was granular enough to be executed by a team member for accountability. Progression of the task to completion was also monitored using Trello. Thus, the development project was adequately managed amid changing requirements. Using an existing app, such as the Telegram app that was used in this study, instead of developing a new bot system reduced the development time. CoSMoS is an independent system that does not require integration into the existing electronic medical record system; its development and deployment into clinical practice were simpler.

This study recognized that the successful development of CoSMoS required effective communication between the technical and clinical teams. Communication and cooperation challenges are common in the development process of an eHealth project, especially where experts from different working cultures and backgrounds are involved [17]. In this CoSMoS project, communication was conducted via virtual meetings (ie, Zoom) and electronic platforms, such as chat groups and emails, due to the movement control restriction in Malaysia. However, communication was effective as both clinical and technical teams understood each other's languages and terminologies, work processes, work practices, and limitations as the team collaborated for a few years. The team leaders and patient advocate in the development team also played a prominent role in coordinating the collaboration. The change requests for CoSMoS were made via virtual meetings during the initial development phase using the agile development method. The communication method had changed from virtual meetings to the help desk method after CoSMoS was developed to enable better management and documentation of system evolution.

The consequential feature of CoSMoS development was the concept of combining a home monitoring app and teleconsultation with doctors. Consideration of patient safety was of the utmost importance, especially in designing the decision-making clinical algorithm and the active telemonitoring system that would help patients be safely monitored at home [18]. There were several challenges in the integration of this interactive and real-time patient monitoring system in clinical practice. Given the high infectivity of COVID-19, research and clinical ethics were considered in the clinical workflow process. The enrollment process was done via phone call to reduce patients' and doctors' exposure times in order to prevent transmission of COVID-19 infection to the health care providers. Several rounds of role-playing between the CoSMoS core members and the clinic staff provided useful feedback and change requests to improve the CoSMoS app and dashboard. The role-playing managed to streamline the integration process, identify technical problems, and reduce the risks. To ensure patients' safety, the clinical workflow of CoSMoS was parallel to the usual care provided by the district health officer so that the care of patients would not be compromised with the introduction of new technology. Patient data privacy and confidentiality is often a main concern, especially during a pandemic [19]. There was social stigma toward patients with confirmed COVID-19 infection [20]. Privacy and data protection were ensured in CoSMoS by applying authentication and authorization mechanisms. Thus, only authorized personnel were granted access to the data according to their rights of access. Each of the health care providers had their own access identity and passwords to access the CoSMoS dashboard. The data from CoSMoS were not shared with other parties or service settings. Data were transmitted over the internet using https secure communication protocol, Telegram messages were encrypted end-to-end, and the database was hosted on a trusted cloud infrastructure provider (ie, DigitalOcean) with clearly defined privacy and security policies. Internally, the technical team had also set a security policy to control access to the patient database.

## Strengths

Strengths of this paper include a full description of the CoSMoS development process, which could be replicated for future health system development during a pandemic. We used a patient-centered care approach in the development process to consider the system's integration into clinical practice. The development process of CoSMoS prioritized the health and safety of patients and health care providers.

## Limitations

A limitation of this development process was that usability testing was only done among computer science students and health care providers. It was a clinical challenge to test the system on real patients given the infectivity of COVID-19 and considering researchers' safety. The usability testing reported in this paper did not represent the target population. The usability and utility evaluation via a qualitative study was conducted separately.

## Conclusions

In view of the urgent need of a COVID-19 symptom monitoring system during the COVID-19 pandemic, we have shown that one can be developed systematically and practically within a short period using the agile model. With effective communication between the technical and clinical teams, developing a digital health care monitoring system is feasible and practicable without compromising patient safety. The development process described in this paper and the lessons learned could guide the development of a digital monitoring system during the next pandemic.

## Authors' Contributions

All the authors participated in the conceptualization and design of the study. HML and CHT contributed to development of the research proposal, conceptualization and writing of this manuscript, and the literature review. CJN and TKC led the research and development of CoSMoS and were involved in all phases of the study. AA, WLN, and HH contributed to writing and critically revising the manuscript. TKC, CSL, and CSC critically revised the manuscript for intellectual content. All authors read and approved the final manuscript.

## Conflicts of Interest

None declared.

Multimedia Appendix 1
Development timeline of the COVID-19 Symptom Monitoring System (CoSMoS).
[PNG File , 107 KB - medinform_v9i2e23427_app1.png ]

Multimedia Appendix 2
Clinical algorithm.
[PDF File (Adobe PDF File), 100 KB - medinform_v9i2e23427_app2.pdf ]

Multimedia Appendix 3
COVID-19 Symptom Monitoring System (CoSMoS) test cases.
[PDF File (Adobe PDF File), 3715 KB - medinform_v9i2e23427_app3.pdf ]

## References

1. Zastrow M. Coronavirus contact-tracing apps: Can they slow the spread of COVID-19? Nature 2020 May 19:1 [FREE Full text] [doi: 10.1038/d41586-020-01514-2] [Medline: 32433633]
2. Gong M, Liu L, Sun X, Yang Y, Wang S, Zhu H. Cloud-based system for effective surveillance and control of COVID-19: Useful experiences from Hubei, China. J Med Internet Res 2020 Apr 22;22(4):e18948 [FREE Full text] [doi: 10.2196/18948] [Medline: 32287040]
3. Ting DSW, Carin L, Dzau V, Wong TY. Digital technology and COVID-19. Nat Med 2020 Apr;26(4):459-461 [FREE Full text] [doi: 10.1038/s41591-020-0824-5] [Medline: 32284618]
4. Editorial. Show evidence that apps for COVID-19 contact-tracing are secure and effective. Nature 2020 Apr;580(7805):563. [doi: 10.1038/d41586-020-01264-1] [Medline: 32350479]
5. Ienca M, Vayena E. On the responsible use of digital data to tackle the COVID-19 pandemic. Nat Med 2020 Apr;26(4):463-464 [FREE Full text] [doi: 10.1038/s41591-020-0832-5] [Medline: 32284619]

6.  Yordanov Y, Dechartres A, Lescure X, Apra C, Villie P, Marchand-Arvier J, AP-HP / Universities / Inserm COVID-19 Research Collaboration. Covidom, a telesurveillance solution for home monitoring patients with COVID-19. J Med Internet Res 2020 Oct 22;22(10):e20748 [FREE Full text] [doi: 10.2196/20748] [Medline: 33006938]

7.  Considerations for quarantine of contacts of COVID-19 cases. World Health Organization. 2020 Aug 19. URL: https://tinyurl.com/4nkjfn2k [accessed 2020-12-05]

8.  COVID-19 management guidelines in Malaysia. Ministry of Health Malaysia. 2020 Mar 25. URL: http://covid-19.moh.gov.my/garis-panduan/garis-panduan-kkm [accessed 2020-12-05]

9.  Lalband N, Kavitha D. Software engineering for smart healthcare applications. Int J Innov Technol Explor Eng 2019 Apr;8(6S4):325-331 [FREE Full text] [doi: 10.35940/ijitee.f1066.0486s419]

10. Matthew-Maich N, Harris L, Ploeg J, Markle-Reid M, Valaitis R, Ibrahim S, et al. Designing, implementing, and evaluating mobile health technologies for managing chronic conditions in older adults: A scoping review. JMIR Mhealth Uhealth 2016 Jun 09;4(2):e29 [FREE Full text] [doi: 10.2196/mhealth.5127] [Medline: 27282195]

11. Song Y, Lu Y. Decision tree methods: Applications for classification and prediction. Shanghai Arch Psychiatry 2015 Apr 25;27(2):130-135 [FREE Full text] [doi: 10.11919/j.issn.1002-0829.215044] [Medline: 26120265]

12. Adu MD, Malabu UH, Malau-Aduli AEO, Malau-Aduli BS. The development of My Care Hub mobile-phone app to support self-management in Australians with type 1 or type 2 diabetes. Sci Rep 2020 Jan 08;10(1):7 [FREE Full text] [doi: 10.1038/s41598-019-56411-0] [Medline: 31913296]

13. Xu H, Huang S, Qiu C, Liu S, Deng J, Jiao B, et al. Monitoring and management of home-quarantined patients with COVID-19 using a WeChat-based telemedicine system: Retrospective cohort study. J Med Internet Res 2020 Jul 02;22(7):e19514 [FREE Full text] [doi: 10.2196/19514] [Medline: 32568727]

14. Heo J, Park JA, Han D, Kim H, Ahn D, Ha B, et al. COVID-19 outcome prediction and monitoring solution for military hospitals in South Korea: Development and evaluation of an application. J Med Internet Res 2020 Nov 04;22(11):e22131 [FREE Full text] [doi: 10.2196/22131] [Medline: 33048824]

15. Sadasivam RS, Delaughter K, Crenshaw K, Sobko HJ, Williams JH, Coley HL, et al. Development of an interactive, web-delivered system to increase provider-patient engagement in smoking cessation. J Med Internet Res 2011 Oct 18;13(4):e87 [FREE Full text] [doi: 10.2196/jmir.1721] [Medline: 22011394]

16. Nota I, Drossaert CHC, Melissant HC, Taal E, Vonkeman HE, Haagsma CJ, et al. Development of a web-based patient decision aid for initiating disease modifying anti-rheumatic drugs using user-centred design methods. BMC Med Inform Decis Mak 2017 Apr 26;17(1):51 [FREE Full text] [doi: 10.1186/s12911-017-0433-5] [Medline: 28441950]

17. Petersen LS, Bertelsen P, Bjørnes C. Cooperation and communication challenges in small-scale eHealth development projects. Int J Med Inform 2013 Dec;82(12):e375-e385. [doi: 10.1016/j.ijmedinf.2013.03.008] [Medline: 23664669]

18. Parimbelli E, Bottalico B, Losiouk E, Tomasi M, Santosuosso A, Lanzola G, et al. Trusting telemedicine: A discussion on risks, safety, legal implications and liability of involved stakeholders. Int J Med Inform 2018 Apr;112:90-98. [doi: 10.1016/j.ijmedinf.2018.01.012] [Medline: 29500027]

19. Li J, Seale H, Ray P, Rawlinson W, Lewis L, Macintyre CR. Issues regarding the implementation of eHealth: Preparing for future influenza pandemics. Interact J Med Res 2012 Dec 06;1(2):e20 [FREE Full text] [doi: 10.2196/ijmr.2357] [Medline: 23611788]

20. Singh R, Subedi M. COVID-19 and stigma: Social discrimination towards frontline healthcare providers and COVID-19 recovered patients in Nepal. Asian J Psychiatr 2020 Oct;53:102222 [FREE Full text] [doi: 10.1016/j.ajp.2020.102222] [Medline: 32570096]

## Abbreviations

**CI/CD:** continuous integration and continuous deployment
**CoSMoS:** COVID-19 Symptom Monitoring System
**HCW:** health care worker
**PUI:** patient under investigation
**UMMC:** University of Malaya Medical Centre

XSL•FO
**RenderX**

XSL•FO

**RenderX**