JMIR Medical Informatics

Impact Factor (2023): 3.1
Volume 10 (2022), Issue 12 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Original Papers

Comparison of Methods for Estimating Temporal Topic Models From Primary Care Clinical Text Data: Retrospective Closed Cohort Study (e40102)
Christopher Meaney, Michael Escobar, Therese Stukel, Peter Austin, Liisa Jaakkimainen. 3

State-of-the-Art Evidence Retriever for Precision Medicine: Algorithm Development and Validation (e40743)
Qiao Jin, Chuanqi Tan, Mosha Chen, Ming Yan, Ningyu Zhang, Songfang Huang, Xiaozhong Liu. 20

Pan-Canadian Electronic Medical Record Diagnostic and Unstructured Text Data for Capturing PTSD: Retrospective Observational Study (e41312)
Leanne Kosowan, Alexander Singer, Farhana Zulkernine, Hasan Zafari, Marcello Nesca, Dhasni Muthumuni. 34

Natural Language Processing and Graph Theory: Making Sense of Imaging Records in a Novel Representation Frame (e40534)
Laurent Binsfeld Gonçalves, Ivan Nesic, Marko Obradovic, Bram Stieltjes, Thomas Weikert, Jens Bremerich. 45

Telehealth System Based on the Ontology Design of a Diabetes Management Pathway Model in China: Development and Usability Study (e42664)
ZhiYuan Fan, LiYuan Cui, Ying Ye, ShouCheng Li, Ning Deng. 62

Construction of Cohorts of Similar Patients From Automatic Extraction of Medical Concepts: Phenotype Extraction Study (e42379)
Christel Gérardin, Arthur Mageau, Arsène Mékinian, Xavier Tannier, Fabrice Carrat. 79

Synchronous Teleconsultation and Monitoring Service Targeting COVID-19: Leveraging Insights for Postpandemic Health Care (e37591)
Milena Marcolino, Clara Diniz, Bruno Chagas, Mayara Mendes, Raquel Prates, Adriana Pagano, Thiago Ferreira, Maria Alkmim, Clara Oliveira, Isabela Borges, Magda Raposo, Zilma Reis, Maria Paixão, Leonardo Ribeiro, Gustavo Rocha, Clareci Cardoso, Antonio Ribeiro. 93

A Framework for Modeling and Interpreting Patient Subgroups Applied to Hospital Readmission: Visual Analytical Approach (e37239)
Suresh Bhavnani, Weibin Zhang, Shyam Visweswaran, Mukaila Raji, Yong-Fang Kuo. 105

Identifying Patterns of Clinical Interest in Clinicians' Treatment Preferences: Hypothesis-free Data Science Approach to Prioritizing Prescribing Outliers for Clinical Review (e41200)
Brian MacKenna, Helen Curtis, Lisa Hopcroft, Alex Walker, Richard Croker, Orla Macdonald, Stephen Evans, Peter Inglesby, David Evans, Jessica Morley, Sebastian Bacon, Ben Goldacre. 124

A Privacy-Preserving Distributed Medical Data Integration Security System for Accuracy Assessment of Cancer Screening: Development Study of Novel Data Integration System (e38922)

Atsuko Miyaji, Kaname Watanabe, Yuuki Takano, Kazuhisa Nakasho, Sho Nakamura, Yuntao Wang, Hiroto Narimatsu. 134

Implementation of Machine Learning Pipelines for Clinical Practice: Development and Validation Study (e37833)

Lara Kanbar, Benjamin Wissel, Yizhao Ni, Nathan Pajor, Tracy Glauser, John Pestian, Judith Dexheimer. 151

Boosting Delirium Identification Accuracy With Sentiment-Based Natural Language Processing: Mixed Methods Study (e38161)

Lu Wang, Yilun Zhang, Mark Chignell, Baizun Shan, Kathleen Sheehan, Fahad Razak, Amol Verma. 160

Corrigenda and Addenda

Correction: The Application of Graph Theoretical Analysis to Complex Networks in Medical Malpractice in China: Qualitative Study (e44374)

Shengjie Dong, Chenshu Shi, Wu Zeng, Zhiying Jia, Minye Dong, Yuyin Xiao, Guohong Li. 132

Original Paper

Comparison of Methods for Estimating Temporal Topic Models From Primary Care Clinical Text Data: Retrospective Closed Cohort Study

Christopher Meaney[1,2], MSc; Michael Escobar[1], PhD; Therese A Stukel[3,4], PhD; Peter C Austin[3,4], PhD; Liisa Jaakkimainen[2,3,4], MSc, MD

[1]Dalla Lana School of Public Health, Division of Biostatistics, University of Toronto, Toronto, ON, Canada

[2]Department of Family and Community Medicine, University of Toronto, Toronto, ON, Canada

[3]ICES, Toronto, ON, Canada

[4]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Christopher Meaney, MSc
Dalla Lana School of Public Health, Division of Biostatistics
University of Toronto
155 College Street
Toronto, ON, M5G1V7
Canada
Phone: 1 4169785602
Email: christopher.meaney@utoronto.ca

Abstract

Background: Health care organizations are collecting increasing volumes of clinical text data. Topic models are a class of unsupervised machine learning algorithms for discovering latent thematic patterns in these large unstructured document collections.

Objective: We aimed to comparatively evaluate several methods for estimating temporal topic models using clinical notes obtained from primary care electronic medical records from Ontario, Canada.

Methods: We used a retrospective closed cohort design. The study spanned from January 01, 2011, through December 31, 2015, discretized into 20 quarterly periods. Patients were included in the study if they generated at least 1 primary care clinical note in each of the 20 quarterly periods. These patients represented a unique cohort of individuals engaging in high-frequency use of the primary care system. The following temporal topic modeling algorithms were fitted to the clinical note corpus: nonnegative matrix factorization, latent Dirichlet allocation, the structural topic model, and the BERTopic model.

Results: Temporal topic models consistently identified latent topical patterns in the clinical note corpus. The learned topical bases identified meaningful activities conducted by the primary health care system. Latent topics displaying near-constant temporal dynamics were consistently estimated across models (eg, pain, hypertension, diabetes, sleep, mood, anxiety, and depression). Several topics displayed predictable seasonal patterns over the study period (eg, respiratory disease and influenza immunization programs).

Conclusions: Nonnegative matrix factorization, latent Dirichlet allocation, structural topic model, and BERTopic are based on different underlying statistical frameworks (eg, linear algebra and optimization, Bayesian graphical models, and neural embeddings), require tuning unique hyperparameters (optimizers, priors, etc), and have distinct computational requirements (data structures, computational hardware, etc). Despite the heterogeneity in statistical methodology, the learned latent topical summarizations and their temporal evolution over the study period were consistently estimated. Temporal topic models represent an interesting class of models for characterizing and monitoring the primary health care system.

(JMIR Med Inform 2022;10(12):e40102) doi:10.2196/40102

KEYWORDS

clinical text data; temporal topic model; nonnegative matrix factorization; latent Dirichlet allocation; structural topic model; BERTopic; text mining

Introduction

Primary Care Text Data

Electronic medical record (EMR) systems are increasingly being adopted in clinical settings across the globe [1]. As a result, health care organizations are generating, collecting, and digitally storing large volumes of routinely collected clinical information. In this study, we focused on clinical text data commonly collected in primary care EMR systems. We compared a class of unsupervised machine learning models—temporal topic models—used to characterize the latent thematic content of large document corpora and summarize latent topical dynamics over time. Temporal topic models have the potential to be applied to large unstructured clinical document collections, routinely captured in modern EMR systems, to passively characterize the primary health care system.

Topic Models

Several methods can be used to estimate a topic model, given a document collection, and to characterize the evolution of latent topical bases over time. Latent Dirichlet allocation (LDA) [2,3] uses a Bayesian probabilistic graphical modeling framework to define a topic model. Learned topical vectors describe the affinity of a word (v=1...V) in the corpus for a particular topic (k=1...K). A latent admixing vector describes the affinity of a specific document (d=1...D) for a specific topic (k=1...K). The latent matrices in the LDA model are learned from document-word co-occurrence statistics empirically collected from the clinical note corpus. The traditional LDA model is not intended for modeling temporal document collections; however, Griffiths et al [4,5] demonstrated how simple time-stratified estimators can be used to illustrate the evolution of latent topical vectors over time. The structural topic model (STM), extends the classical LDA model, allowing either (1) the matrix of per-document topical prevalence weights or (2) the matrix of per-topic word probabilities to deterministically vary according to covariate information parameterized using a generalized linear model [6]. Several parameterizations of time can be incorporated into the generalized linear model (eg, discrete, continuous, or spline effects), allowing the STM to flexibly model the evolution of topical prevalence vectors over time. Nonnegative matrix factorization (NMF) [7-9] uses a linear algebraic framework and principles from constrained optimization for topic modeling. NMF directly estimates the parameter matrices of a topic model by factorizing an observed document term matrix (DTM) into 2 latent nonnegative matrices. One of the latent parameter matrices describes the affinity of a document (d=1...D) to a topic (k=1...K), and the other latent matrix describes the affinity of a word (v=1...V) to a topic (k=1...K). Post hoc multivariate transformations of the NMF latent parameter matrices can be used to generate estimates of topical evolution over time. Recently, neural frameworks have been developed for topic modeling, such as top2vec [10] and BERTopic [11]. The BERTopic neural topic models begin by embedding documents into a latent vector space. A finite number of clusters (k=1...K) of semantically similar documents are identified in the embedding space. For each document cluster (k), the most relevant words describing the cluster or topic are extracted using a cluster-specific term-frequency inverse-document frequency (TF-IDF) weighting technique [11].

Study Objectives

The objective of this study was to compare the performance of several temporal topic modeling methodologies fitted to a corpus of primary care clinical notes. We compared the following temporal topic modeling methodologies: NMF, LDA, STM, and BERTopic. We examined (1) the overall matrix of per-topic word probabilities estimated over the corpus and (2) the multivariate time series structures describing the evolution of latent topical prevalence weights (k=1...K) over discrete times (t=1...T). We compared the methods using a data set of longitudinal primary care clinical notes collected over 5 years (2011-2015) in Ontario, Canada.

Methods

Mathematically Representing and Computationally Processing Our Clinical Text Corpus

Topic models use statistical information regarding document-word co-occurrence frequencies to learn meaningful latent variable representations from a corpus. Each document in the collection (d=1...D) is represented as a high-dimensional length-V vector (v=1...V), where each element is a count of the number of times a particular word or token (v) in an empirical vocabulary is observed in a particular document (d). We represented the collection of document-specific term-frequency vectors into a matrix X of dimension D*V, called the DTM. The DTM is a large, sparse matrix. However, the matrix is overdetermined because many of the rows (representing document-specific term-frequency vectors) and columns (representing word or token occurrence frequency over all documents in the corpus) demonstrate strong intercorrelations. Dimension-reduction techniques, such as topic models, use intercorrelated statistical semantic information to estimate meaningful thematic representations from document collections. Topic models learn (1) clusters of intercorrelated words describing the topical content of the corpus and (2) clusters of correlated documents sharing latent topical concepts.

The most challenging and subjective aspect associated with construction of the DTM involves specification of the vocabulary or dictionary (v=1...V) encoding the column space of the matrix. A priori constructed lexicons or dictionaries (of dimension V) can be used to determine the study vocabulary. Specification of appropriate domain-specific dictionaries would be tasked with subject matter experts on the research team. Alternatively, an entirely computational approach could specify a text tokenization or normalization pipeline and computationally parse the input character sequences into a finite number of tokens.

In this study, we adopted a hybrid approach to vocabulary or dictionary specification. We began by tokenizing the clinical notes on whitespace boundaries (spaces, tabs, newlines, carriage returns, etc). We normalized tokens using lower-case conversion and removed all nonalphabetic characters. We removed tokens with a character length ≤1. Finally, we sorted the list of tokens or words by decreasing occurrence frequency and manually reviewed the sorted list of tokens. Our manual review identified V=2930 distinct tokens for inclusion in our final vocabulary. The total number of tokens in the corpus was 3,003,583. The tokens chosen for inclusion in our final dictionary or vocabulary were mainly medical terms with precise semantic meanings (disease names, disease symptoms, drug names, medical procedures, medical specialties, anatomical locations, etc). We excluded stop words or tokens (ie, syntactic or functional tokens with little clinical semantic meaning). Words with low occurrence frequency were excluded for computational considerations. All text processing was conducted using R (R Foundation for Statistical Computing; version 3.6).

Review of Methods for Temporal Topic Modeling

NMF Model

NMF estimates latent topical matrices using the document-word co-occurrence statistics contained in the empirical DTM. NMF factorizes the D*V dimensional DTM into 2 latent submatrices of dimensions D*K (θ) and K*V (Φ). The DTM (X) consists of nonnegative integers (ie, word frequency counts), whereas the learned matrices (θ,Φ) consist of nonnegative real values. Mathematically, the NMF objective involves learning optimal values of the latent matrices (θ,Φ) that best approximate the input data set ($X \approx \theta\Phi$), subject to the constraint that the learned matrices contain nonnegative values.

We selected a least square loss function to train the NMF model. The objective function specifies that the observed data elements are approximated in a K-dimensional bilinear form . The analyst must specify the dimensions of the latent space: K (the number of topics). Seminal articles on NMF include Paatero and Tapper [7] and Lee and Seung [8,9]. Surveys of NMF and low-rank models are provided by Berry et al [12] and Udell et al [13].

Post hoc, the row vectors constituting both θ and Φ, can be normalized by dividing by their respective row sums. The resulting normalized vectors can be interpreted as compositional or probability vectors (ie, each normalized row of θ and Φ contains nonnegative entries that sum to 1, row-wise). The row vectors of the matrix Φ encode a set of k=1...K per-topic word probabilities or proportions (estimated over a discrete set of v=1...V words in the empirical corpus vocabulary). The row vectors of the matrix θ encode a set of d=1...D per-document topic proportions (estimated over a discrete set of k=1...K latent dimensions), encoding the affinity a given document has for a particular topic.

For each document d=1...D, assume we observe a time stamp that allows us to associate each document (and latent embedding) with a T-dimensional indicator variable denoting the observation time (t=1...T). We estimated a K-dimensional multivariate mean topical prevalence vector for each design point, t=1...T. This resulted in a multivariate time series structure (a T*K dimensional matrix). Each column (k=1...K) of the matrix is a length T time series that described the evolution of a latent topical vector.

The sklearn.decomposition.NMF() function in the Python SKLearn package (version 0.24.2) was used to fit the NMF topic model.

LDA Model

LDA is a probabilistic topic model. Probabilistic topic models assume that a document comprises a mixture of topics. These (latent) topics represent a probability distribution over a finite vocabulary of words or tokens. Topic models can also be described as admixture models. Each document is a soft mixture of topics (k=1...K), where a topic is itself a probability distribution over words in the vocabulary (v=1...V). A graphical model describing LDA is shown in Figure 1 [2].

The LDA graphical model also describes a generative process for creating a single document in the corpus. This can be succinctly described using the following sampling notation [14,15].

To generate a document, we begin by sampling the per-topic word distributions from a Dirichlet distribution parameterized by a V dimensional prior concentration parameter (β). Topical vectors (k=1...K) are shared over the collection of documents.

Next, for each document d=1...D in the collection, we sample the per-document topic distribution from a Dirichlet distribution parameterized according to a K-dimensional prior concentration parameter (α).

For each word in each document, we sample a topical indicator variable, $z_{d,n}$. This variable takes an integer value between 1 and K and signifies the per-topic word distribution from which a specific word, $w_{d,n}$, is chosen. The index n denotes the n^{th} word in a variable length document ($n=1...N_d$).

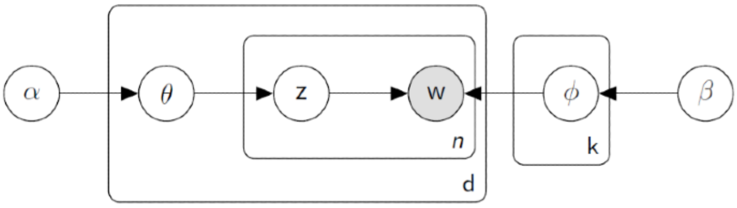
Finally, we draw a single word token, $w_{d,n}$, from the topical distribution associated with $z_{d,n}$. The word indicator is an element v=1...V in our empirical dictionary or vocabulary.

The statistical inference problem associated with probabilistic topic modeling involves inverting the sampling process and learning model-defined latent parameters given the observed text data. The latent variables indicate which words are assigned to which topical indicators (z), which documents have an affinity for which topics (θ), and which words co-occur with high likelihood under which topics (Φ). The latent parameters associated with an LDA topic model are typically estimated using Bayesian statistical machinery (Gibbs sampling [14], variational inference [2], and other methods).

A multivariate transformation of the matrix of per-document topical prevalence weights generates a multivariate time series data structure. This object is of dimension T*K, where each column k=1…K represented a univariate topical time series of length T. This series describes the evolution of latent topical vectors over our study period.

The sklearn.decomposition.LatentDirichletAllocation() function in Python SKLearn (version 0.24.2) was used to fit the LDA topic model.

Figure 1. Graphical model representation of the latent Dirichlet allocation topic model.



STM Model

The STM is another type of probabilistic topic model. The STM extends the LDA topic model, allowing latent matrices of (1) per-document topical prevalence weights or (2) per-topic word proportions to vary according to a generalized linear model parameterization [6]. Covariate effects on the latent matrix of per-document topical prevalence weights are incorporated into the model using a logistic-normal prior distribution over per-document topical prevalence vectors, similar to the correlated topic model [16]. Covariate effects on the latent matrix of per-topic word proportions are incorporated into the model using a type of multinomial logit prior. In this study, we modeled covariate effects (in our study, discrete time effects, t=1...T) on the matrix of per-document topic prevalence weights. We did not assume that the matrix of per-topic word proportions varied according to covariates. The plate notation of STM is shown in Figure 2. Variational methods are used for posterior inference in STM [6].

To generate a document under STM, we begin by sampling the per-topic word distributions from an (intercept-only) multinomial logit model (where multinomial logit regression parameters are given sparse “gamma-lasso” prior) [6].

Next, we sample the per-document topic distribution from a logistic-normal distribution parameterized in terms of a mean vector and covariance matrix. γ represents a D*T dimensional design matrix encoding the time point (t=1...T) under which the document (d=1...D) was observed. The vector γ is a matrix of dimension T*K and encodes discrete time effects on each of the per-document topical prevalence weights (a length K vector for each document d=1...D). Finally, Σ is a K*K dimensional covariance matrix that encodes correlations between topical prevalence vectors (parameterized under a logistic-normal model).

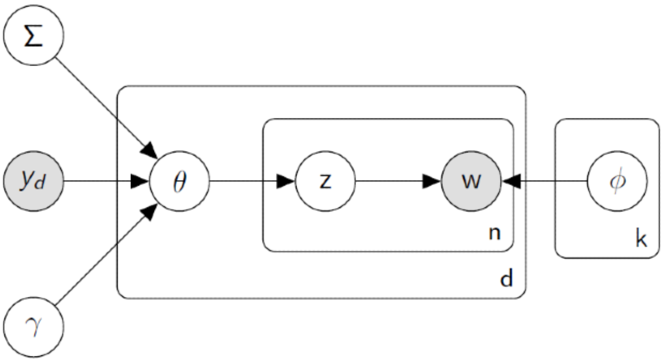
For each word (n=1...N_d) in each document (d=1...D), we sample a topical indicator variable $z_{d,n}$. This variable takes an integer value between 1 and K and signifies the per-topic word distribution from which a specific word, $w_{d,n}$, is chosen. It must be noted that the upper limit N_d suggests that the number of words used for any given document (d) can vary.

Finally, we draw a single word or token, $w_{d,n}$, from the topical distribution associated with $z_{d,n}$. The word indicator is an element v=1...V in our empirical dictionary or vocabulary.

The framework for STM naturally allows for the estimation of temporal effects on topical prevalence weights. In our study, discrete time effects on topical prevalence can be interpreted using the coefficient matrix (γ) from the fitted logistic-normal model. As the temporal effects are encoded in a Bayesian regression modeling framework, we can also compute inferential measures (posterior means, highest posterior density intervals, etc). The single-stage inferential mechanism encoded in STM is a clear strength over earlier NMF and LDA models.

We used the stm() function in the STM package in R to fit the STM to our study data.

Figure 2. Graphical model representation of the structural topic model.



Neural Topic Modeling via BERTopic

Recently, researchers have developed topic models that integrate neural architectures and related techniques for model specification and learning. These neural topic models represent a different class of topic models compared with those introduced previously. Examples of recently developed neural topic models include top2vec [10] and BERTopic [11]. In this study, we focused on the BERTopic model.

BERTopic begins with embedding documents empirically observed in the study corpus into a latent embedding space. Many methods exist for embedding discrete linguistic units (words, sentences, paragraphs, documents, etc) into an embedding space. For example, words can be embedded in a vector space using word2vec [17-19], GloVe [20], FastText [21], ELMO [22], Flair [23], and transformer models [24]. Sentences and documents can be embedded using methods such as doc2vec [25], universal sentence encoders [26], and transformers [24]. The BERTopic model used in this study relies on sentence transformers [27], particularly the MPNet sentence transformer model [28]. The neural embedding model is a discrete "hyperparameter" in the BERTopic modeling pipeline. Different choices of neural embedding models are associated with their own model-specific hyperparameters (embedding dimension, context window width, model training or optimization arguments, etc).

Each document (d=1...D) is embedded in a vector space, typically of a few hundred dimensions. The uniform manifold approximation and projection (UMAP) algorithm [29] was used as a further nonlinear dimension-reduction technique to assist in the visualization and clustering of document vectors. Clustering was accomplished in the UMAP-reduced space using the hierarchical density-based spatial clustering algorithm of applications with noise (HDBSCAN) [30].

Clusters (k=1...K) of semantically related documents were identified. Scores over words v=1...V in the vocabulary were computed using cluster-specific TF-IDF weights. If a cluster consisted of semantically focused documents, and hence words, we expect to observe coherent and meaningful words identified via TF-IDF scoring. The proportion of documents assigned to each cluster during a specific period (t=1...T) can be used to generate a T*K dimensional multivariate time series structure, depicting the evolution of latent topic over our study period.

We fitted the BERTopic model using default hyperparameter settings. The BERTopic pipeline requires (1) specification of a document embedding algorithm (in our case, the MPNet sentence transformer model [28]), (2) the UMAP nonlinear dimension-reduction algorithm, (3) the HDBSCAN algorithm for cluster identification, and (4) cluster-specific TF-IDF scoring. The individual components of the pipeline could involve substantive hyperparameter optimization. In this study, we used the default model hyperparameter settings.

We used the Python package bertopic to fit BERTopic models.

Statistical Methods for Corpus Description and Evaluation of Learned Temporal Topic Models

We used simple counts and percentages to describe the characteristics of our study sample. We described the number of unique patients and number of unique clinical notes. Each patient in our sample was a "high-user" of the primary care system, in the sense they generated at least one encounter/note for each of the twenty quarterly time periods between 2011-2015. We described the distribution of the number of notes per patients. We described demographic characteristics of the sample (age/sex distributions).

When fitting the NMF, LDA, and STM models, we constructed a DTM whose row dimension corresponded to the number of unique patients in the sample (ie, 1727 unique patients) multiplied by the number of distinct time periods (t=20; 1727×20=34,540). Each term-frequency vector observed in the DTM was length V (V=2930), and an individual element counted the number of times a given word was observed for a given patient in each quarterly period. Across the DTM, we counted the total number of words and the number of unique words. We described the counts and percentages of the top 25 most prevalent words in our clinical note corpus. We also described the sparsity of the DTM.

For each of the NMF, LDA, STM, and BERTopic models, we constructed a K*T dimensional multivariate time series matrix (this is the transpose of the T*K data structure described earlier). Each row corresponds to a latent topic vector and each column corresponds to a specific quarterly time period. A row vector is a length T time series describing the evolution of a latent topical vector across the study periods. Each column corresponds to a distribution over topics at a particular period (ie, described which topics are most important at a given period). For each row k=1...K, we report the top 5 words loading most strongly on a given topic. The cluster of words was semantically correlated and described the essence of the latent topical vector. A heatmap was used to visualize this high-dimensional multivariate time series structure; and we hierarchically clustered the rows of the matrix using a Euclidean distance metric and Ward agglomeration method (a dendrogram was used to visualize the cluster structure of the topical series).

The topical structure of each of the NMF, LDA, STM, and BERTopic model fits was described in terms of the top 5 words loading most strongly on each of the k=1...K latent topics. In other words, the topical structure of each model can be described in terms of a "bag" of 250 words or tokens. We investigated the topical diversity of the model fits. Topical diversity was calculated in terms of the number of unique words in the bag of 250 total words. Furthermore, we investigated the top 5 most frequently occurring words in the "bag" describing each model fit. The redundantly occurring words in the topical summaries provided a rough approximation of the semantic concepts that the models repeatedly identified as important.

We investigated several measures of topical coherence for the NMF, LDA, STM, and BERTopic models. We considered the "UMASS," "UCI," and normalized pointwise mutual information ("NPMI") metrics described in the surveys of Roder et al [31] and Rosner et al [32]. These metrics assessed the internal consistency of the collection of word clusters describing the topical structure of the NMF, LDA, STM, and BERTopic models. The theoretical minima or maxima of each coherence measure varies; however, larger values indicate models that generated more coherent topical characterizations. Mathematical details related to the calculation of the aforementioned topical coherence metrics are provided later and further outlined in the studies by Roder et al [31] and Rosner et al [32]. In all the equations used, we assumed that a topical vector is described in terms of its top-L most probable words or tokens; $\{w_i,w_j\}$ represented distinct words from the top-L set, ε is a small positive constant to avoid potential numerical issues in computation; and δ is a weighting term (used in the normalized NPMI estimates, compared with the unnormalized pointwise mutual information estimates used in the UCI coherence measure).

We used a set-based measure of concordance, the Jaccard coefficient, to assess similarities or differences in the topical structure describing the NMF, LDA, STM, and BERTopic models. Each model was described in terms of a "bag" of 250 words or tokens (ie, k=50 topics, described in terms of their top 5 most probable words); consider 2 models generating bags of words or tokens, b_0 and b_1. The Jaccard coefficient is defined as the cardinality of the intersection of b_0 and b_1 divided by the cardinality of the union of b_0 and b_1. In mathematical notation, the Jaccard coefficient is expressed as follows:

Finally, we described the wall time (in seconds or minutes) required to fit each of the NMF, LDA, STM, and BERTopic models. We also discussed the computational issues associated with hyperparameter tuning of each of the models.

Study Design, Setting, Data Sources, and Inclusion or Exclusion Criteria

This study used a retrospective closed cohort design. Clinical notes were obtained from primary care EMR systems geographically distributed across Ontario, Canada. We included all clinical notes written by the patient's primary care provider between January 01, 2011, and December 31, 2015. We discretized time into quarterly strata (January-March; April-June; July-September; and October-December). Patients were excluded if they did not have at least one clinical note in each of the 20 quarterly strata over the study period. Hence, the selected sample of patients reflects a unique set of individuals who frequently engaged with the primary health care system.

Results

Description of Corpus and Study Sample

Our document collection contained 160,478 clinical notes from 1727 patients. The 1727 patients received primary care services from 1066 unique primary care physicians at 40 unique primary care clinics (geographically distributed across Ontario, Canada). The median age of the patients was 68 (IQR 55-80) years and ranged from 20 to 103 years (age statistics were calculated using study baseline as a reference date, January 1, 2011). Female patients were observed more frequently than male patients (1157/1727, 67% vs 570/1727, 33%). Table 1 describes the characteristics of the study sample (in terms of both note-level and patient-level units of analysis).

The initial note-level DTM had dimensions of 160,478 rows (one row for each clinical note in the corpus) by 2930 columns (one column for each unique word or token in the corpus). The corpus comprised 3,003,583 tokens. The DTM was >99% sparse (ie, it contained almost all zero elements). We also constructed a patient-quarter–level DTM by aggregating notes observed on the same patient within a quarter. This DTM had dimensions of 1727×20=34,540 rows by 2930 columns and was >98% sparse. The top 25 most frequently occurring words in the analytic corpus are listed in Table 2.

Table 1. Descriptive statistics for study sample, at note-level and patient-level unit of analysis.

Characteristic	Unique notes (n=160,478), n (%)	Unique patients (n=1727), n (%)
Age (years)		
20-40	9713 (6.1)	107 (6.1)
40-65	63,588 (39.6)	675 (39.1)
65-85	63,839 (39.8)	704 (40.8)
>85	23,338 (14.5)	241 (14)
Sex		
Male	51,530 (32.1)	570 (33)
Female	108,948 (67.9)	1157 (67)
Year		
2011	28,012 (17.5)	—[a]
2012	31,220 (19.5)	—
2013	33,676 (21)	—
2014	33,756 (21)	—
2015	33,814 (21)	—

[a] Not applicable.

Table 2. Top 25 most frequently occurring tokens or words in the final analytic primary care clinical note corpora (N=3,003,583).

Token or word	Occurrence frequency, n (%)
pain	88,132 (2.93)
mg	65,612 (2.18)
inr	52,970 (1.76)
bp	50,751 (1.69)
back	43,556 (1.45)
dose	29,861 (0.99)
feels	24,736 (0.82)
rx	23,211 (0.77)
chest	22,256 (0.74)
meds	20,914 (0.7)
referral	19,409 (0.65)
work	19,398 (0.65)
wt	19,322 (0.64)
feeling	17,415 (0.58)
blood	16,121 (0.54)
symptoms	15,905 (0.53)
prn	15,706 (0.52)
urine	14,633 (0.49)
bw	13,779 (0.46)
lab	13,543 (0.45)
clear	13,271 (0.44)
knee	12,677 (0.42)
pharmacy	12,503 (0.42)
sleep	12,331 (0.41)
prescription	11,945 (0.4)

Comparing Temporal Topic Models Estimated With NMF, LDA, STM, and BERTopic Models

We comparatively evaluated inferences obtained from fitting the NMF, LDA, STM, and BERTopic models to our primary care clinical note corpus. For each model, we varied the number of topics (K={25,40,45,50,55,60,75}) and observed similar inferences at various levels of the model complexity parameter (K). When K was too small, distinct semantic topics tended to be grouped together, whereas when K was too large, semantically similar topics tended to be split into arbitrary clusters (resulting in an overclustering effect). Using human judgment evaluation, we determined that a model complexity of K=50 topics balanced a parsimonious, while simultaneously expressive, characterization of the clinical document corpus. For each of the NMF, LDA, STM, and BERTopic models, we reported the results assuming K=50 latent topics.

A summary of the distribution of words over the k=1...50 latent topics (for each of the 4 models under comparison) is given in Figures 3-6, respectively. The y-axis in each figure lists the top 5 words loading most strongly on a given topic. For NMF, LDA, and STM, we reported topical prevalence weights associated with each word or token (which is approximately the probability of observing the word or token under a given latent topic). For the BERTopic model, we reported normalized cluster-specific TF-IDF scores associated with words under topics (which can be interpreted similarly to the outputs of the NMF, LDA, and STM models). The x-axis of these plots represents t=1...20 quarterly periods. A column in the plot represents a topical prevalence distribution over latent topics at a given time point. A row in the plot illustrates the evolution of a latent topic over the study period.

Each of the 4 latent temporal topic models learned a meaningful representation of the primary care clinical notes corpus. In the following paragraphs, we discuss (1) topics consistently estimated across models that demonstrated constant trends in topical prevalence across quarterly periods and (2) topics consistently estimated across quarterly periods that demonstrated interesting seasonal patterns.

Each of the fitted models consistently identified the following latent primary care topical constructs (and these topics show constant patterns across quarterly periods): sleep

(NMF=Topic−45; LDA=Topic-2 or Topic-31; STM=Topic-11; BERTopic=not applicable); mental health, for example, mood, anxiety, and depression, (NMF=Topic-33; LDA=Topic-22; STM=Topic-19; BERTopic=Topic-16); pain (NMF=Topic-1; LDA=Topic-39, Topic-36, Topic-14, Topic-49, Topic-34, or Topic-37; STM=Topic-8; BERTopic=Topic-9 or Topic-39); blood pressure control and monitoring (NMF=Topic-36; LDA=Topic-9; STM=Topic-21; BERTopic=Topic-31); respiratory disease, for example, cough, throat, chest, fever, etc (NMF=Topic-46; LDA=Topic-13; STM=Topic-46; BERTopic=Topic-1), smoking (NMF=Topic-31; LDA=Topic-32; STM=Topic-44; BERTopic=Topic-38); diabetes, for example, blood, sugar, insulin, fbs, etc (NMF=Topic-5; LDA=Topic-43; STM=Topic-42; BERTopic=Topic-8); pharmaceutical prescription management (NMF=Topic-26; LDA=Topic-40; STM=Topic-9; BERTopic=Topic-36 or Topic-5); and annual influenza vaccination programs (NMF=Topic-6; LDA=Topic-29; STM=Topic-36; BERTopic=Topic-50). These thematic areas represented archetypical patients, conditions, or roles encountered in the primary health care system. The consistent extraction of latent themes (represented as semantically correlated word clusters) suggests that each model can leverage information regarding word-context co-occurrence to learn meaningful patterns from a large unstructured clinical document corpus.

Figures 3-6 illustrate 4 different temporal topic model multivariate time series structures. For a given plot, the x-axis represents time (t=1...20 quarterly periods from 2011-2015), and the y-axis represents a topical vector (k=1...50). The intensity of color in the cell (t,k) indicates the extent to which an encounter at time (t) is related to a latent topic (k). Topical labels are exchangeable and clustered along the y-axis, according to the similarity of the topical time series (a dendrogram describing the similarity or differences across topical clusters is illustrated in Figure 7). Figure 3-6 represent different multivariate time series structures estimated with NMF (Figure 3), LDA (Figure 4), STM (Figure 5), and BERTopic (Figure 6).

For certain learned topics, seasonal harmonic patterns were stably estimated over the study period. For example, the annual influenza vaccination program consistently occurred in the fall or winter months of the study (NMF=Topic-6; LDA=Topic-29; STM=Topic-36; BERTopic=Topic-50). Similarly, annual spikes in respiratory diseases (cough, cold, influenza, etc) are identified as achieving peaks in the winter months and lows in the summer months (NMF=Topic-46; LDA=Topic-13; STM=Topic-46; BERTopic=Topic-1). These findings are illustrated in Figures 3-6; however, we also present individual time series plots of these topics in Figures 8 and 9, so the reader can better appreciate the ability of the different temporal topical models to extract consistent seasonal patterns from the primary care clinical document corpus. Findings regarding consistent seasonal variation in primary care roles over time have strong face validity and are corroborated by complementary data sources (eg, administrative data). Furthermore, the consistency by which these patterns are extracted from our large clinical document collection helps build trust in the opportunity to use word-context co-occurrence statistics (and topic models) to characterize and monitor primary care practices and systems.

Figure 3. A heat map of the multivariate time series structure associated with the nonnegative matrix factorization temporal topic model.

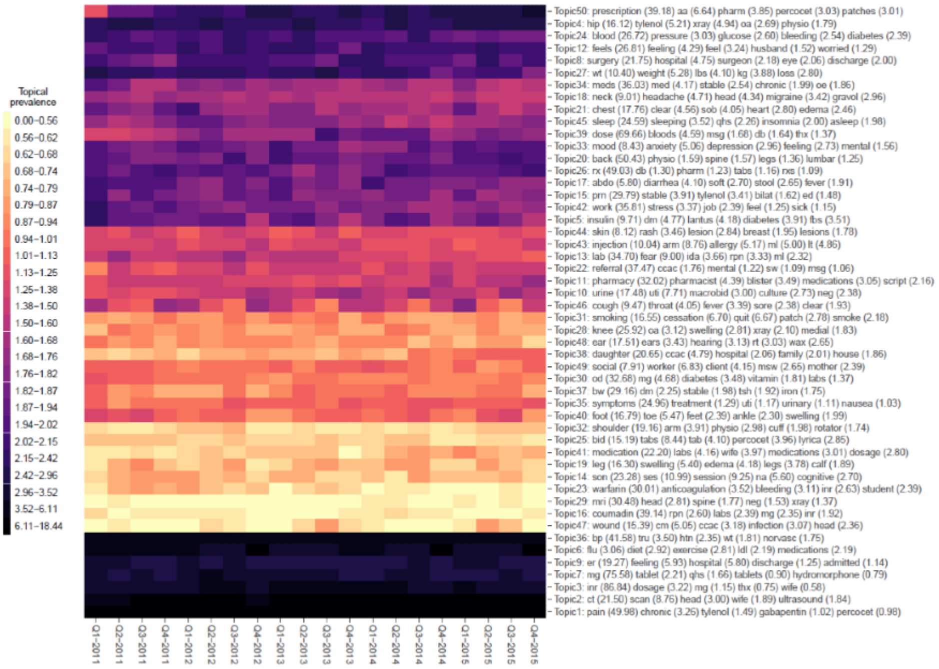


Figure 4. A heat map of the multivariate time series structure associated with the latent Dirichlet allocation temporal topic model.

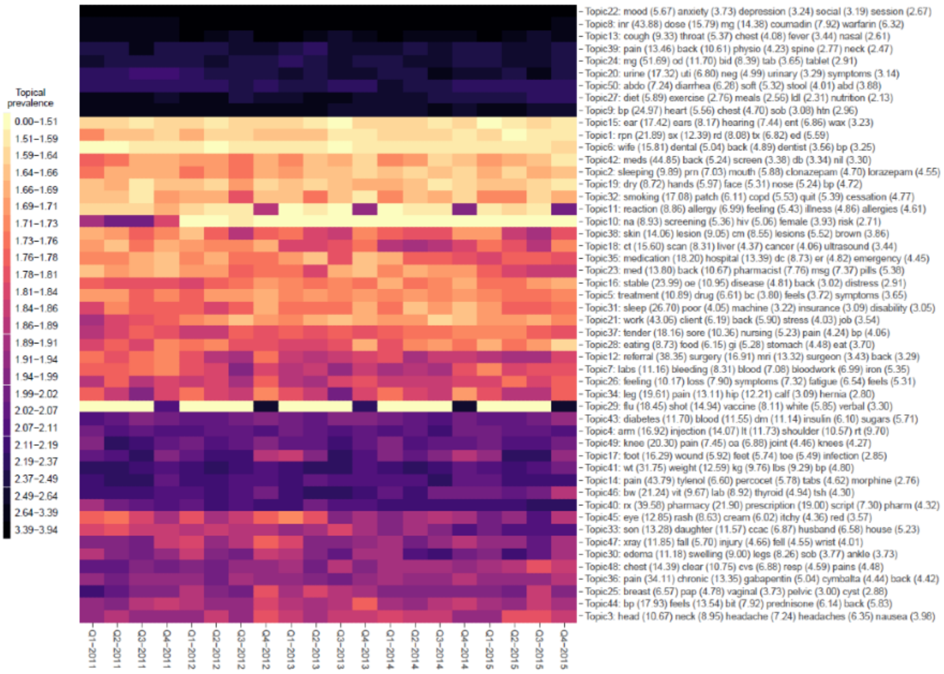


Figure 5. A heat map of the multivariate time series structure associated with the structural topic model temporal topic model.

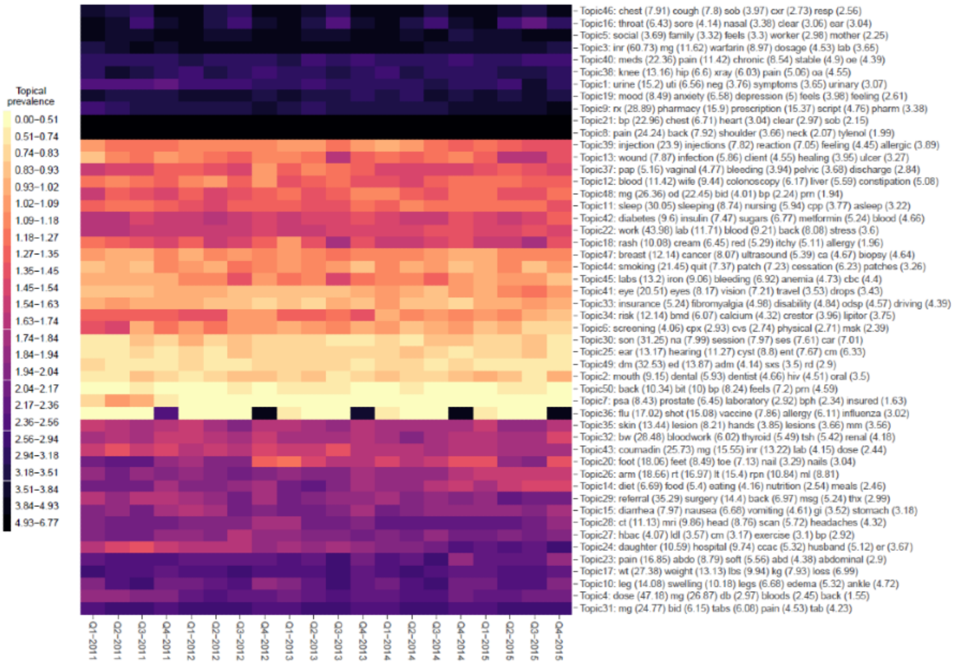


Figure 6. A heat map of the multivariate time series structure associated with the BERTopic temporal topic model.

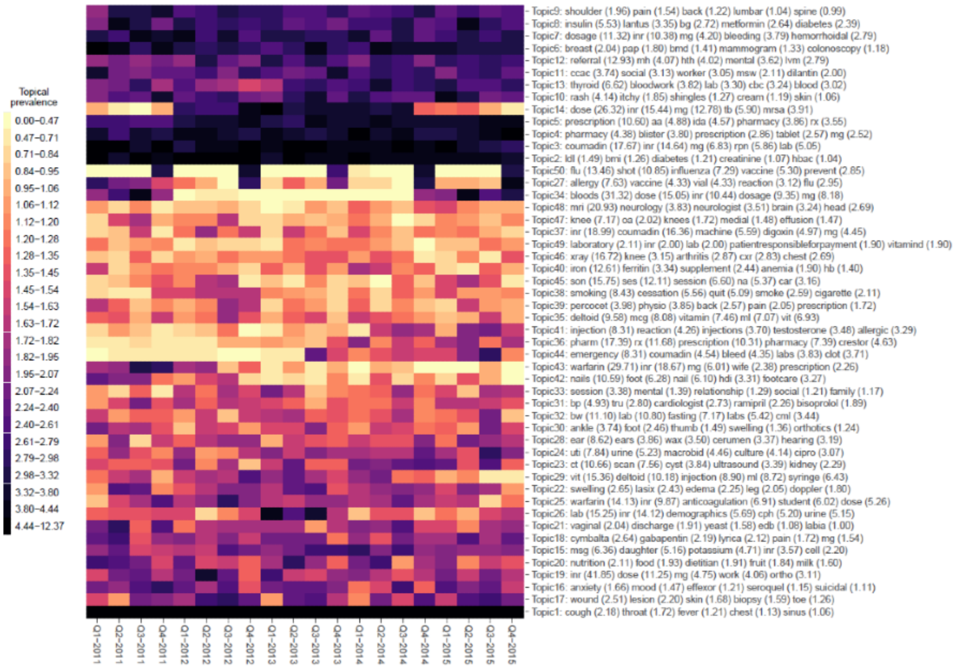


Figure 7. Dendrograms displaying the clustering structure of the latent multivariate time series objects learned from nonnegative matrix factorization model (A), latent Dirichlet allocation model (B), structural topic model (C) and BERTopic model (D).

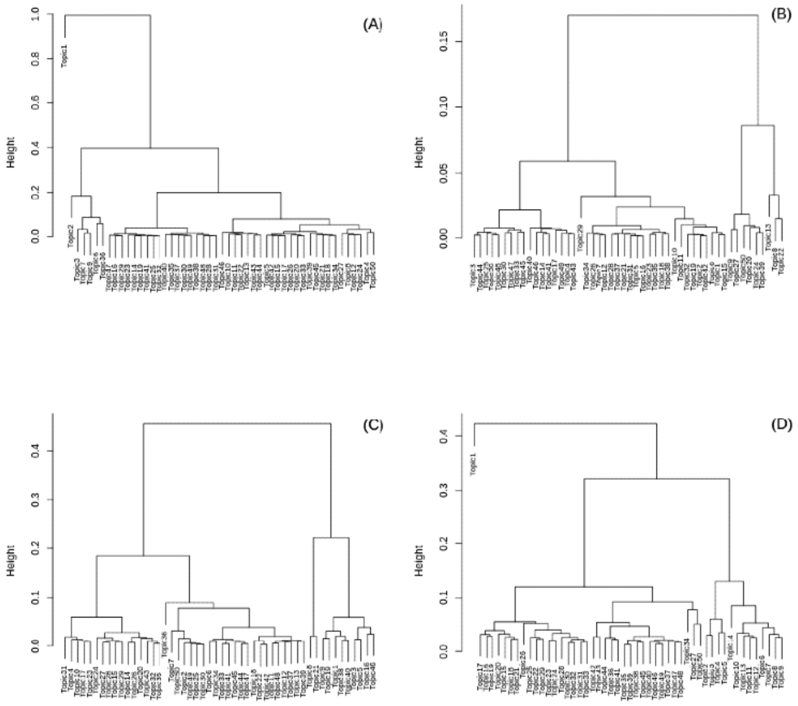


Figure 8. Descriptive time series plots characterizing the seasonal evolution of annual influenza program topic, as estimated by nonnegative matrix factorization model (A), latent Dirichlet allocation model (B), structural topic model (C) and BERTopic-models (D).

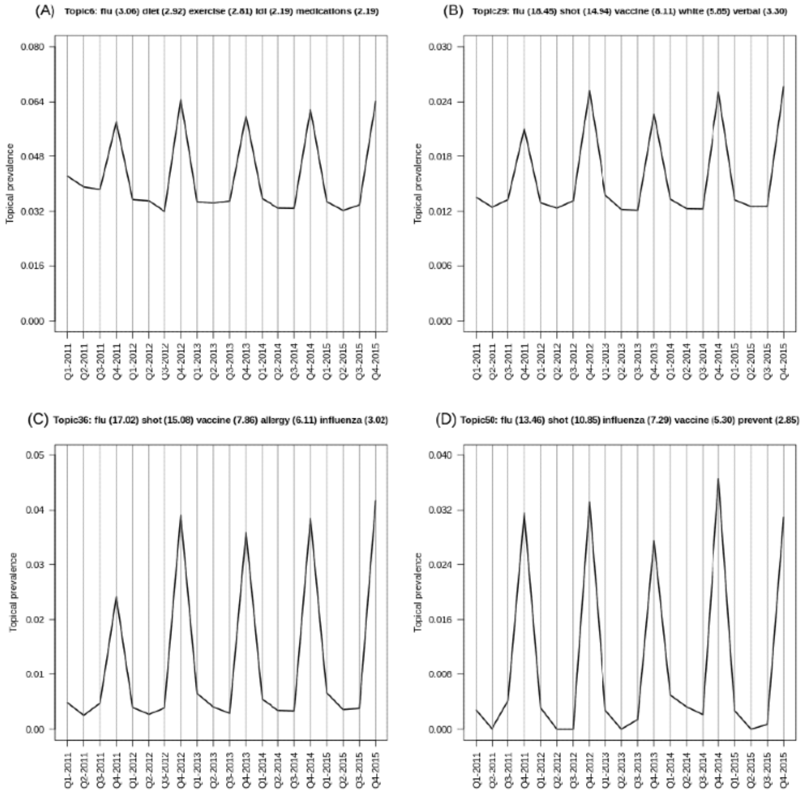
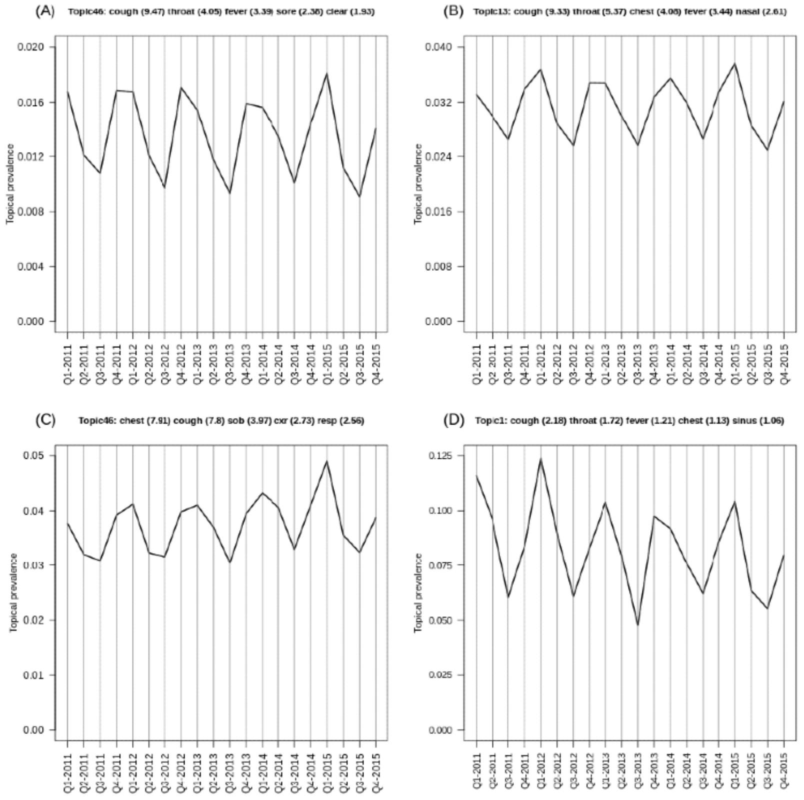


Figure 9. Descriptive time series plots characterizing the seasonal evolution of the respiratory disease topic, as estimated by nonnegative matrix factorization model (A), latent Dirichlet allocation model (B), structural topic model (C) and BERTopic-models (D).



Post Hoc Internal Evaluation of Fitted Temporal Topic Models

When investigating the top-ranked words associated with per-word topic distributions in Figures 3-6 we note that each model can describe the corpus using a "bag" of up to 250 unique words (K=50 topics multiplied by top 5 words being presented for each latent topical representation). The number of unique words—also known as the topic diversity—observed in NMF, LDA, STM, and BERTopic model fits was 76.4% (191/250), 88.4% (221/250), 87.6% (219/250), and 77.2% (193/250), respectively. The top 5 most frequently recurring words or tokens describing the topical structure of each of the NMF, LDA, STM, and BERTopic models are listed in Table 3. Recurring words for LDA and STM are similar, suggesting that primary care issues related to back pain (and other musculoskeletal pain) are important, as are issues related to hypertension and feelings (eg, mood disorders). Conversely, the BERTopic model seems to prioritize primary care issues related to prescription drugs and laboratory ordering or management.

We explored the semantic coherence of NMF, LDA, STM, and BERTopic models using the following metrics: "UMASS," "UCI," and "NPMI" (Table 4) [31,32]. Larger coherence metrics indicated increasingly internally consistent latent topical characterizations. The "UMASS" metric favored the STM model, whereas, the "UCI" and "NPMI" metrics favored the BERTopic model.

To investigate the differences and similarities in the fitted topic model, we used the Jaccard coefficient (Table 5). Using the Jaccard measure of concordance, the Bayesian models (LDA or STM) were identified as resulting in the most similar fit. The BERTopic model generated the most distinct topical representation compared with the other models.

The time required to train each model was reported. For NMF, LDA, and STM models, we used a single central processing unit (although Python SKLearn implementations of decomposition models can be parallelized). For the BERTopic model, we used a single graphics processing unit for embedding documents and a single central processing unit for dimensionality reduction (UMAP) and clustering (HDBSCAN). Under these settings, the time required to fit the NMF, LDA, STM, and BERTopic models was 237 seconds, 67 seconds, 879 seconds (14.7 minutes), and 2624 seconds (43.7 minutes), respectively. The computational requirements of the BERTopic model exceeded those of the other models, particularly the highly optimized NMF or LDA implementations in Python SKLearn.

Table 3. The most frequently occurring tokens observed in each of the bags of 250 words describing the topical structure of latent Dirichlet allocation (LDA), nonnegative matrix factorization (NMF), structural topic model (STM) and BERTopic model fits (and their occurrence counts in the bag).

Word or token	Topic model			
	NMF (n)	LDA (n)	STM (n)	BERTopic (n)
Word or token-1	head (4)	back (9)	back (5)	inr (11)
Word or token-2	mg (4)	bp (6)	mg (5)	mg (9)
Word or token-3	ccac (3)	pain (6)	pain (5)	lab (5)
Word or token-4	diabetes (3)	chest (3)	bp (4)	prescription (5)
Word or token-5	feeling (3)	feels (3)	feels (3)	dose (4)

Table 4. Topical coherence measures ("UMASS," "UCI," and normalized pointwise mutual information ["NPMI"]) estimated on each of the nonnegative matrix factorization (NMF), latent Dirichlet allocation (LDA), structural topic model (STM) and BERTopic models.

Topical coherence measure	Topic model			
	NMF	LDA	STM	BERTopic
UMASS	−2.522	−2.488	−2.372	−2.591
UCI	1.220	0.987	1.192	1.405
NPMI	0.183	0.149	0.190	0.230

Table 5. Jaccard coefficient metrics of set-based concordance between fitted topic models: nonnegative matrix factorization (NMF), latent Dirichlet allocation (LDA), structural topic model (STM), and BERTopic.

	NMF	LDA	STM	BERTopic
NMF	—[a]	—	—	—
LDA	0.526	—	—	—
STM	0.491	0.577	—	—
BERTopic	0.343	0.286	0.329	—

[a] Not applicable.

Discussion

Principal Findings

In this study, we compared several distinct methodologies (ie, NMF, LDA, STM, and BERTopic) to estimate temporal topic models from a large collection of primary care clinical notes. Despite differences in the underlying statistical methodology, models often converged on a consistent latent characterization of the corpus. Furthermore, the temporal evolution of latent topics was reliably extracted from each of the NMF, LDA, STM, and BERTopic models.

Clinically, our data set represented high-users of the primary care system. Many of the latent topics emerging from this analysis are consistent with a high-user archetype, for example, family counseling or social work, mood disorders, anxiety or depression, chronic pain, arthritis and musculoskeletal disorders, neurological conditions, cardiovascular disease and hypertension, diabetes, cancer screening (breast, cervical, colorectal, and prostate), laboratory requisitions and blood work, diagnostic imaging, and pharmaceutical or prescription management. Topic models also identified numerous acute health conditions as important latent themes, such as cough, cold and other respiratory infections, urinary tract infections, skin conditions, and wound care. NMF, LDA, STM, and BERTopic models each consistently captured (1) annual primary care influenza programs and (2) seasonal respiratory conditions, demonstrating predictable seasonal variation. Findings regarding primary care use patterns, extracted solely from clinical text data, were largely corroborated by provincial reporting based on structured administrative data [33].

We observed that disparate statistical methodologies for estimating temporal topic models generated a concordant or consistent latent representation. We interpreted this to mean that as the signal-to-noise ratio increases in a given clinical text data set, the subtle choice of statistical methodology seems to matter less, and any of these methods would extract a meaningful latent representation of the primary care corpus. For smaller corpora, where word-document co-occurrence statistics are less certain, this hypothesis may not hold.

Furthermore, subtle or nuanced differences in model representations emerged, which may lead analysts to favor specific modeling strategies in particular settings. For example, consider Figure 8 for the annual influenza vaccination program. Models such as NMF and LDA are purely unsupervised and do not consider external covariate information when formulating the model objective function. For NMF or LDA models we noticed that the “grand mean” topical prevalence over time centers at approximately 2% (ie, 1/50 topics). Conversely, an STM intentionally incorporates covariate information in the Bayesian graphical models’ prior structure, and we observed that for STM, the lows for annual influenza topic are much closer to 0%, whereas the fall or winter peaks are more pronounced. The BERTopic model does not intentionally incorporate covariate information into its objective function(s) either; however, it adopts a more “local averaging” principle to estimate topical distributions over time and, as such, demonstrated similar seasonal harmonic patterns as STM in the context of the annual influenza program. Similar patterns can be observed in Figure 9 for seasonal respiratory diseases. This suggests that different topic models may perform more or less optimally in certain scientific settings (ie, may be dependent on the research question, available data, and how these foundational aspects of a study interplay with model choice). A priori, should the analyst or researcher expect topical prevalence to vary about select observable covariates, it may make sense to adopt a more flexible model that can adequately incorporate this anticipated behavior. If there is no a priori rationale to believe that topical prevalence varies as a function of covariates (eg, time in this study), then the choice of model may become less relevant, as all models may perform similarly well.

Because of the different statistical principles associated with each temporal topic modeling methodology, each method is associated with its own strengths and weaknesses. We have elaborated on the methodological and computational issues associated with each class of models.

First, NMF is the most mature and seemingly parsimonious methodology for topic modeling. NMF is strongly rooted in linear algebraic principles and is fundamentally based on the constrained optimization of a simple least squares objective function. Vanilla NMF is a well-studied statistical methodology and many efficient computational routines exist for estimating NMF models. NMF is flexible and can be readily extended. Possible model extensions can be viewed as discrete tunable hyperparameters in the model fitting process. Berry et al [12] and Cichocki et al [34] discussed distinct algorithmic techniques for estimating the latent parameters of an NMF model, such as gradient descent, multiplicative updates, and alternating nonnegative least squares. The choice of algorithm can be conceived as a discrete tunable hyperparameter. Furthermore, analysts are often confronted with the choice of whether to regularize the latent parameter matrices [35]. Ridge, lasso, and elastic net regularization are commonly encountered, although more complex regularization can be used to encourage latent representations with smoothness, minimal volume, and other characteristics. Furthermore, many researchers have attempted to introduce coherent generalizations of NMF and related techniques [13]. For example, generalized low-rank models that flexibly incorporate different loss functions, functional forms, weighting of data points, and regularization have been discussed by Udell et al [13].

LDA and STM are Bayesian topic models. LDA was developed as a fully Bayesian extension of existing linear algebraic-based (eg, latent semantic analysis) and maximum likelihood-based (eg, probabilistic latent semantic indexing) techniques for topic modeling [2]. LDA has been extended in various ways, illustrating the flexibility of Bayesian probabilistic graphical models. For example, STM is a direct extension of LDA, which allows latent parameter matrices to vary as a function of observed covariates [6]. Efficient computational fitting routines have been developed for LDA, and STM to a certain extent. Analysts face several decisions when fitting LDA and STM models to empirical data sets, including Bayesian inferential or computational methods (eg, Gibbs sampling vs variational inference) and prior distribution specifications.

BERTopic represents the most novel approach to topic modeling [11]. The BERTopic model is a pipeline: (1) deep neural networks (eg, sentence transformer models) embed documents in a vector space; (2) nonlinear dimension reduction is applied to latent document vectors (UMAP); (3) document clusters are identified (HDBSCAN); and (4) representative topics (collections of semantically correlated words) are extracted from document clusters using a cluster-specific TF-IDF scoring method. A disadvantage of the BERTopic pipeline is related to computational requirements. For large corpora, a graphics processing unit is required to learn document embeddings within a reasonable time. In our study, we randomly down-sampled our data set (3/8 documents were included, whereas 5/8 documents were excluded), even with a graphics processing unit. That said, the BERTopic model's strength is related to its modularity. We observed that the BERTopic model generates meaningfully coherent topics, and as neural embedding methods continue to evolve, we anticipate that state-of-the-art document embedding techniques can be dropped into this pipeline.

Limitations and Future Work

We attempted to be transparent with respect to how our final vocabulary of words or tokens was selected and accordingly the DTMs were constructed for this study. Different computational pipelines could have been used to preprocess our clinical text corpus. For instance, we could have used different strategies for tokenization, lemmatization, stemming, stop-word removal, and frequency-based word or token removal. Different text preprocessing pipelines would ultimately lead to different DTM structures (with different vocabularies). Further research is needed to better understand the implications of these text preprocessing decisions on downstream study inferences.

Each topic model considered in this study requires specification of hyperparameters that govern the aspects of model fitting. Fitting these topic models is computationally intensive for large input data sets. We focused mainly on the stability and robustness of inferences with respect to model complexity (K), a common hyperparameter across all models. We did not explore the stability of the inferences across other model-specific hyperparameters.

We did not consider all possible methods for estimating temporal topic models in this study. Bespoke NMF and LDA variants exist that are applicable for estimating temporal topic models. Sequential NMF [36] and dynamic LDA [37] are 2 extensions which are relevant for estimating temporal topic models. Tensor factorization models such as the canonical polyadic decomposition or Tucker decomposition, which factorize a D*V*T tensor into meaningful latent parameter matrices, may also be applicable [34,38]. Additional surveys related to topic modeling are provided in the studies by Churchill and Singh [39], Zhao et al [40], and Boyd-Graber et al [41].

These works have led us to consider several possible ways of extending different topic modeling frameworks, including Bayesian NMF with document-level covariates (similar to the STM extension of LDA), neural matrix factorization with (nontemporal) covariates, LDA or STM extensions that allow per-document topical prevalence weights to vary according to a flexible generalized linear mixed model or multilevel model (for modeling dependencies introduced because of the complex design or sampling mechanism by which documents are created), and computational methods for improving statistical inference (eg, interval estimation and hypothesis testing) when engaging with temporal topic models (eg, resampling methods, bootstrap, and multiple outputation).

Conclusions

In this study, we compared several statistical techniques for estimating temporal topic models from primary care clinical text data. Different temporal topic models have unique strengths and weaknesses owing to their underlying statistical properties. Nonetheless, each model consistently estimated a latent variable representation of a primary care document collection, which meaningfully characterized high-use primary care patients and their longitudinal interactions with the primary health care system. As the adoption of EMRs increases and health care organizations amass increasingly large volumes of clinical text data, temporal topic models may offer a mechanism for leveraging unstructured clinical text data for characterization and monitoring of primary care practices and systems.

Acknowledgments

This study was supported by funding provided by a Foundation Grant (FDN 143303) from the Canadian Institutes of Health Research. The funding agency had no role in the study design; collection, analysis, or interpretation of data; writing of the report; or decision to submit the report for publication. Dr Austin is supported by a Mid-Career Investigator Award from the Heart and Stroke Foundation.

Conflicts of Interest

None declared.

References

1. Mossialos E, Djordjevic A, Osborn R, Sarnak D. International profiles of health care systems. The Commonwealth Fund. 2017 May 31. URL: https://www.commonwealthfund.org/publications/fund-reports/2017/may/international-profiles-health-care-systems [accessed 2022-09-30]
2. Blei D, Ng A, Jordan M. Latent dirichlet allocation. J Mach Learn Res 2003 Jan 3;3:993-1022. [doi: 10.5555/944919.944937]
3. Blei DM. Probabilistic topic models. Commun ACM 2012 Apr;55(4):77-84. [doi: 10.1145/2133806.2133826]

4. Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci U S A 2004 Apr 06;101 Suppl 1(suppl_1):5228-5235 [FREE Full text] [doi: 10.1073/pnas.0307752101] [Medline: 14872004]
5. Griffiths T, Steyvers M. Probabilistic topic models. In: Landauer TK, McNamara DS, Dennis S, Kintsch W, editors. Handbook of Latent Semantic Analysis. New York, NY, USA: Psychology Press; 2007.
6. Roberts ME, Stewart BM, Airoldi EM. A model of text for experimentation in the social sciences. J Am Stat Assoc 2016 Oct 18;111(515):988-1003. [doi: 10.1080/01621459.2016.1141684]
7. Paatero P, Tapper U. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 1994 Jun;5(2):111-126. [doi: 10.1002/env.3170050203]
8. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999 Oct 21;401(6755):788-791. [doi: 10.1038/44565] [Medline: 10548103]
9. Lee D, Seung S. Algorithms for non-negative matrix factorization. In: Proceedings of the 13th International Conference on Neural Information Processing Systems. 2000 Presented at: NeurIPS '00; January 1, 2000; Denver, CO, USA.
10. Angelov D. TOP2VEC: distributed representations of topics. arXiv 2020 [FREE Full text]
11. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv 2022 [FREE Full text]
12. Berry MW, Browne M, Langville AN, Pauca VP, Plemmons RJ. Algorithms and applications for approximate nonnegative matrix factorization. Computation Stat Data Analysis 2007 Sep;52(1):155-173. [doi: 10.1016/j.csda.2006.11.006]
13. Udell M, Horn C, Zadeh R, Boyd S. Generalized low rank models. FNT Mach Learn 2016;9(1):1-118. [doi: 10.1561/2200000055]
14. Griffiths M. Gibbs sampling in the generative model of latent dirichlet allocation. CiteSeerX. URL: https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.7.8022 [accessed 2022-09-30]
15. Heinrich G. Parameter Estimation for Text Analysis: Technical Report. University of Leipzig. 2008. URL: http://www.arbylon.net/publications/text-est.pdf [accessed 2022-11-07]
16. Blei DM, Lafferty JD. A correlated topic model of Science. Ann Appl Stat 2007 Jun 1;1(1):17-35. [doi: 10.1214/07-aoas114]
17. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. 2013 Presented at: NeurIPS '13; December 5-10, 2013; Lake Tahoe, NV, USA.
18. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv 2013 [FREE Full text] [doi: 10.3126/jiee.v3i1.34327]
19. Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013 Presented at: NAACL '13; June 9-14, 2013; Atlanta, GA, USA.
20. 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 26–28, 2014; Doha, Qatar. [doi: 10.3115/v1/d14-1162]
21. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. arXiv 2016 Aug 6 [FREE Full text] [doi: 10.18653/v1/e17-2068]
22. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. arXiv 2018 [FREE Full text] [doi: 10.18653/v1/n18-1202]
23. Akbik A, Bergmann T, Blythe D, Rasul K, Schweter S, Vollgraf R. FLAIR: an easy to use framework for state of the art NLP. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). 2019 Presented at: NAACL '19; June, 2019; Minneapolis, MN, USA.
24. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv 2018 [FREE Full text]
25. Le Q, Mikolov T. Distributed representations of sentences and documents. arXiv 2014 [FREE Full text]
26. Cer D, Yang Y, Kong S, Hua N, Limtiaco N, St. John R, et al. Universal sentence encoder. arXiv 2018 [FREE Full text]
27. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. arXiv 2019 [FREE Full text] [doi: 10.18653/v1/d19-1410]
28. Song K, Tan X, Qin T, Lu J, Liu TY. MPNet: masked and permuted pre-training for language understanding. arXiv 2020 [FREE Full text]
29. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv 2018 [FREE Full text] [doi: 10.21105/joss.00861]
30. Campello RJ, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates. In: Proceedings of the 17th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. 2013 Presented at: PAKDD '13; April 14-17, 2013; Gold Coast, Australia p. 160-172 URL: https://link.springer.com/chapter/10.1007/978-3-642-37456-2_14 [doi: 10.1007/978-3-642-37456-2_14]
31. Roder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. 2015 Presented at: WSDM '15; February 2-6, 2015; Shanghai, China p. 399-408. [doi: 10.1145/2684822.2685324]
32. Rosner F, Hinneburg A, Roder M, Nettling M, Both A. Evaluating topic coherence measures. arXiv 2014.

33. Jaakkimainen L, Upshur RE, Klein-Geltink JE, Leong A, Maaten S, Schultz SE, et al. Primary Care in Ontario: ICES Atlas. Institute for Clinical Evaluative Sciences. Toronto, Canada: Institute for Clinical Evaluative Sciences; 2006 Nov. URL: https://www.ices.on.ca/~/media/Files/Atlases-Reports/2006/Primary-care-in-Ontario/Full-report.ashx [accessed 2022-11-07]
34. Cichocki A, Zdunek R, Phan A, Amari SI. Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation. Hoboken, NJ, USA: Wiley Online Library; 2009.
35. Hoyer P. Non-negative matrix factorization with sparseness constraints. J Mach Learn Res 2004 Jan 12;5:1457-1469. [doi: 10.5555/1005332.1044709]
36. Mackevicius E, Bahle A, Williams A, Gu S, Denisenko NI, Denisenko MS, et al. Unsupervised discovery of temporal sequences in high-dimensional datasets, with applications to neuroscience. eLife 2019;8:e38471. [doi: 10.7554/elife.38471]
37. Blei D, Lafferty J. Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning. 2006 Presented at: ICML '13; June 25-29, 2006; Pittsburgh, PA, USA.
38. Kolda TG, Bader BW. Tensor decompositions and applications. SIAM Rev 2009 Aug 06;51(3):455-500. [doi: 10.1137/07070111x]
39. Churchill R, Singh L. The evolution of topic modeling. ACM Comput Surv (forthcoming) 2022 Jan 12:2021. [doi: 10.1145/3507900]
40. Zhao H, Phung D, Huynh V, Jin Y, Du L, Buntine W. Topic modelling meets deep neural networks: a survey. arXiv 2021 [FREE Full text] [doi: 10.24963/ijcai.2021/638]
41. Boyd-Graber J, Hu Y, Mimno D. Applications of topic models. FNT Inf Retrieval 2017;11(2-3):143-296. [doi: 10.1561/1500000030]

Abbreviations

DTM: document term matrix
EMR: electronic medical record
HDBSCAN: hierarchical density-based spatial clustering algorithm of applications with noise
LDA: latent Dirichlet allocation
NMF: nonnegative matrix factorization
NPMI: normalized pointwise mutual information
STM: structural topic model
TF-IDF: term-frequency inverse-document frequency
UMAP: uniform manifold approximation and projection

Edited by C Lovis; submitted 06.06.22; peer-reviewed by I Mircheva, A King; comments to author 31.07.22; revised version received 01.09.22; accepted 18.09.22; published 19.12.22.

Please cite as:
Meaney C, Escobar M, Stukel TA, Austin PC, Jaakkimainen L
Comparison of Methods for Estimating Temporal Topic Models From Primary Care Clinical Text Data: Retrospective Closed Cohort Study
JMIR Med Inform 2022;10(12):e40102
URL: https://medinform.jmir.org/2022/12/e40102
doi:10.2196/40102
PMID:36534443

Original Paper

State-of-the-Art Evidence Retriever for Precision Medicine: Algorithm Development and Validation

Qiao Jin[1], MD; Chuanqi Tan[1], PhD; Mosha Chen[1], PhD; Ming Yan[1], PhD; Ningyu Zhang[2], PhD; Songfang Huang[1], PhD; Xiaozhong Liu[3], PhD

[1]Alibaba Group, Hangzhou, China
[2]Zhejiang University, Zhejiang, China
[3]Indiana University Bloomington, Bloomington, IN, United States

Corresponding Author:
Chuanqi Tan, PhD
Alibaba Group
No. 969 West Wen Yi Road, Yuhang District
Hangzhou, 311121
China
Phone: 86 15201162567
Email: chuanqi.tcq@alibaba-inc.com

Abstract

Background: Under the paradigm of precision medicine (PM), patients with the same disease can receive different personalized therapies according to their clinical and genetic features. These therapies are determined by the totality of all available clinical evidence, including results from case reports, clinical trials, and systematic reviews. However, it is increasingly difficult for physicians to find such evidence from scientific publications, whose size is growing at an unprecedented pace.

Objective: In this work, we propose the PM-Search system to facilitate the retrieval of clinical literature that contains critical evidence for or against giving specific therapies to certain cancer patients.

Methods: The PM-Search system combines a baseline retriever that selects document candidates at a large scale and an evidence reranker that finely reorders the candidates based on their evidence quality. The baseline retriever uses query expansion and keyword matching with the ElasticSearch retrieval engine, and the evidence reranker fits pretrained language models to expert annotations that are derived from an active learning strategy.

Results: The PM-Search system achieved the best performance in the retrieval of high-quality clinical evidence at the Text Retrieval Conference PM Track 2020, outperforming the second-ranking systems by large margins (0.4780 vs 0.4238 for standard normalized discounted cumulative gain at rank 30 and 0.4519 vs 0.4193 for exponential normalized discounted cumulative gain at rank 30).

Conclusions: We present PM-Search, a state-of-the-art search engine to assist the practicing of evidence-based PM. PM-Search uses a novel Bidirectional Encoder Representations from Transformers for Biomedical Text Mining–based active learning strategy that models evidence quality and improves the model performance. Our analyses show that evidence quality is a distinct aspect from general relevance, and specific modeling of evidence quality beyond general relevance is required for a PM search engine.

(JMIR Med Inform 2022;10(12):e40743) doi:10.2196/40743

KEYWORDS

precision medicine; evidence-based medicine; information retrieval; active learning; pretrained language models; digital health intervention; data retrieval; big data; algorithm development

Introduction

Traditionally, patients with the same diseases are treated with the same therapies. However, the treatment effects can be highly heterogeneous, that is, the benefits and risks may differ substantially among patient subgroups [1]. The precision medicine (PM) research initiative [2] takes into account individual differences in people's genes, environments, and lifestyles when tailoring their treatment and prevention strategies. Under the ideal paradigm of PM, patients of the same diseases are divided into several subgroups, and different patient subgroups receive different treatments that are the most suitable for them. PM is now widely applied in oncology, since sequencing techniques can identify considerable genetic variations in patients with cancer. For example, patients with non–small cell lung cancer with epidermal growth factor receptor gene mutations are sensitive to gefitinib therapy [3], and patients with breast cancer who have human epidermal growth factor receptor 2 mutations are sensitive to trastuzumab therapy [4].

PM practices should be guided by the principles of evidence-based medicine [5], where treatments are based on high-quality clinical evidence, such as systematic reviews and randomized controlled trials, instead of individual experiences. However, as the number of scientific publications is growing rapidly (eg, about 2700 articles are added to PubMed each day in 2019), it is difficult for physicians to find clinical evidence in the literature that supports or reject specific treatment options for certain patients. Information retrieval (IR) is aimed at automatically finding relevant documents for users' queries. IR has been successfully applied to the general consumer and biomedical research domain with search engines such as Google and PubMed. However, most current search engines cannot process PM queries that contain structured information about patients and therapies and neither do they rank the documents based on their significance as clinical evidence.

To facilitate IR research for PM, the Text Retrieval Conference (TREC) holds the PM Track annually since 2017. From 2017 to 2019, the TREC PM focused on finding relevant academic papers or clinical trials of patient topics specified by their demographics, diseases, and gene mutations [6-8]. In 2020, the TREC PM focus was changed to retrieve academic papers that report critical clinical evidence for or against a given treatment in a population specified by its disease and gene mutation [9]. Both supporting and opposing clinical evidence are important, because they provide valuable guidance to clinical decision making regarding whether or not to use the treatment. To assist the practices of PM, such as in the case of the TREC PM task, the most vital property of a retriever is to rank the relevant papers by their evidence quality, that is, to what extent they can assist clinical decision-making. The objective of this work was to develop a retrieval model that can rank relevant papers by their evidence quality to a given PM topic.

Traditional IR systems are mostly based on term frequency–inverse document frequency and its derivatives that basically rank the documents by their bag-of-word similarities with the input query. However, biomedical concepts are often referred to by various synonyms, and multiple studies have shown the importance of expanding query concepts to their synonyms before sending them to IR systems [10-12]. To further model for domain-specific relevance, such as evidence quality in our case, rerankers are often added to finely rerank the candidates returned by retrieval systems. However, such rerankers are typically based on deep learning, and training them requires a large number of labeled instances [13], which are prohibitively expensive to collect in the biomedical domain. Recent large-scale pretrained language models such as Embeddings from Language Models [14] and Bidirectional Encoder Representations from Transformers (BERT) [15] show significant performance improvement over several natural language processing benchmarks such as General Language Understanding Evaluation [16]. BERT is basically a transformer [17] encoder that is pretrained to predict a randomly masked token in the original input. BERT can be effectively used to rank documents given a specific query [18].

In this work, we propose the PM-Search model that tackles the aforementioned problems of traditional search engines to assist the practice of PM. The PM-Search system has two main components: (1) a baseline retriever using query expansion and keyword matching with the ElasticSearch engine; and (2) an evidence reranker that ranks the initial documents returned by ElasticSearch based on their evidence quality. The reranking uses article features as well as pretrained language models under an expert-in-the-loop active learning strategy, where a biomedical language model BERT for Biomedical Text Mining (BioBERT) [19] is fine-tuned interactively with the experts. Our models participated in the TREC PM 2020 as the ALIBABA team and ranked the highest in the evidence quality assessment: PM-Search achieved standard normalized discounted cumulative gain (NDCG) at rank 30 (NDCG@30) of 47.80% and exponential NDCG@30 of 45.19%, outperforming the second-ranking system by large margins.

In summary, our contributions of this work are three-fold:

1. We present PM-Search, which is an integrated IR system specifically designed to assist precision medicine. PM-Search achieved state-of-the-art performance in the TREC PM Track.
2. We used an expert-in-the-loop active learning strategy based on BioBERT to efficiently derive annotations and improve model performance. To the best of our knowledge, this is the first precision medicine search engine that combines active learning and pretrained language models.
3. We thoroughly analyzed the importance of each system feature with a full set of ablation studies, where we found that the most important features included publication types and active learning. We hope the experiments can provide some insights into the potential future directions of PM search engines.

Methods

Data and Materials

The TREC 2020 PM Track provided 40 topics for evaluation. Each topic represented a PM query that contains three key elements of a specific patient population: (1) the disease, that is, the type of cancer; (2) the genetic variant, that is, the gene mutation; and (3) the tentative treatment. The topics were synthetically generated by biomedical experts and several examples are shown in (Table 1). The task used the 2019 PubMed baseline as the official corpus, which contains over 29 million biomedical citations. Each citation is composed of the title, authors, abstract, etc, of the article. For each topic, we denoted its disease as , the genetic variant as and the treatment as . The returned articles were denoted as . Each retrieval result was a query-article pair that contained , , and . We also used the publication type and citation count information extracted in PubMed as additional data sources.

The evaluation of the task followed standard TREC procedures of ad hoc retrieval, where participants submitted a maximum number of 1000 ranked articles and up to 5 different runs for each topic. The assessments were divided into 2 phases, where phase 1 was "Relevance Assessment," judging the relevance of each article, and phase 2 was "Evidence Assessment," judging the evidence quality provided by the article.

Phase 1 assessment was a general IR assessment that only considered relevance, where the assessors first judged whether the returned article *a* is generally related to PM. For the PM papers, the assessors then assessed whether the *d*, *g*, and *t* were exact, partially matching, or missing in *a*. Finally, the results were classified as "Definitely Relevant," "Partially Relevant," or "Not Relevant" based on a predefined rules of how the *d*, *g*, and *t* matched. The evaluation metrics used in phase 1 include precision at rank 10 (P@10), inferred NDCG (infNDCG), and R-precision (R-prec). P@10 and R-prec are precisions at different ranks:

where is the number of relevant articles for the query. NDCG is computed by:

where

rel_i is the relevance score of article *i* and $|\mathrm{REL}_n|$ denotes the number of relevant articles ordered by the relevance up to position *n*. Since not all submitted articles would be judged by the organizers, there cannot be an exact value of NDCG. To deal with this issue, a sample set of all articles in the top 30 ranks and a 25% sample of articles in ranks 31-100 was used to compute the NDCG, that is, infNDCG.

In the phase 2 assessment, the assessors scored the relevant papers from the phase 1 assessment using a 5-point scale. For example, the tier 4 results should be "randomized controlled trial with >200 patients and single drug, or meta-analysis" and tier 0 should be "Not Relevant" for topic 16. The scale was tailored for each topic to adjust for the differences in the disease, genetic variant, and treatment. The main evaluation metric for phase 2 assessment was NDCG@30. NDCG values at this phase are exact since all articles in the top 30 ranks are judged. Two sets of relevance values were used to compute NDCG, the standard gains (std-gains) and the exponential gains (exp-gains). Standard gains have scores (ie, rel_i) of 0, 1, 2, 3, and 4 corresponding to the 5 tiers, whereas exponential gains have scores of 0, 1, 2, 4, and 8 corresponding to 5 tiers.

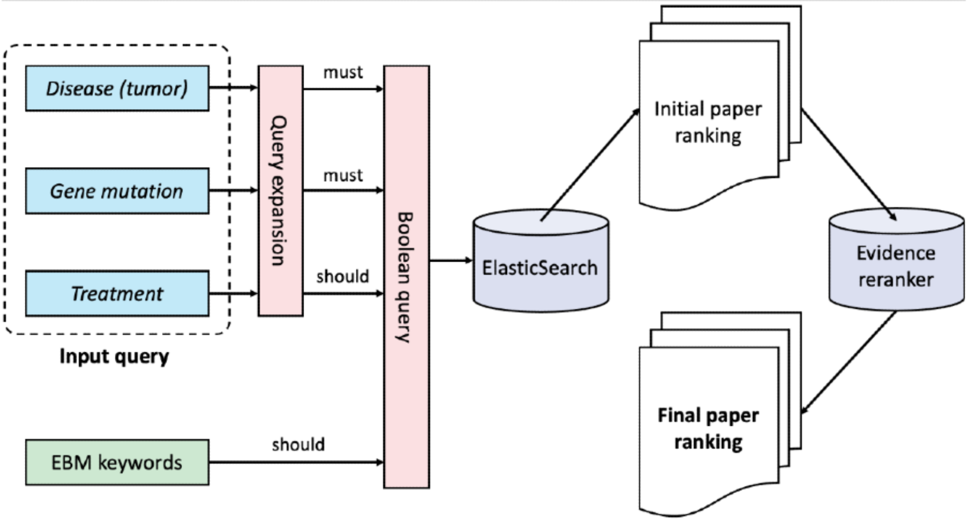
Table 1. Examples of the Text Retrieval Conference Precision Medicine 2020 topics.

Topic	Disease	Gene	Treatment
1	Colorectal cancer	ABL proto-oncogene 1	Regorafenib
11	Breast cancer	Cyclin dependent kinase 4	Abemaciclib
21	Differentiated thyroid carcinoma	Fibroblast growth factor receptor 2	Lenvatinib
31	Hepatocellular carcinoma	Neurotrophic receptor tyrosine kinase 2	Sorafenib

PM-Search Overview

As shown in (Figure 1), PM-Search uses a 2-step approach to retrieve relevant articles for each given PM topic: (1) a *baseline retriever* that is fast and scalable, generating a relatively small number (eg, thousands) of candidates out of millions of PubMed articles—the baseline retriever is based on ElasticSearch (reference) where the original queries are expanded by a list of weighted synonyms; and (2) an *evidence reranker* that finely reranks the retrieved documents based on their evidence quality—the evidence reranker combines the predictions from a BioBERT fine-tuned by an expert-in-the-loop active learning strategy and a feature-based linear regressor.

Figure 1. The architecture of PM-Search. EBM: evidence-based medicine; PM: Precision Medicine.



Baseline Retriever

We indexed the titles and abstracts of all articles from the PubMed 2019 baseline provided by the TREC organizers using ElasticSearch, a Lucene-based search engine. The synonyms of the disease *d* and gene variant *g* were found via the National Library of Medicine's web application programming interface in MedlinePlus [20,21]. We denoted the retrieved synonyms of *d* and *g* as $\{d_1, d_2, \dots, d_m\}$ and $\{g_1, g_2, \dots, g_m\}$, where $d_1 = d$ and $g_1 = g$. We did not expand the treatment because the provided term either had no synonym or was used in almost all articles.

For each synonym d_1 and g_1, we counted their document frequency $df(d_i)$ and $df(g_i)$ in the baseline corpus and calculated the weights of each synonym used in ElasticSearch:

where

We used the normalized document frequency to lower the ranks of rare terms.

We performed the retrieval in ElasticSearch, which ranks the documents based on their word-level relevance with the input query using the Okapi BM25 algorithm [22]. At the highest level, we queried the ElasticSearch indices using a Boolean query that *must match* the disease and treatment query and *should match* the gene query. The disease, treatment, and gene queries were all *dis_max* queries composed of their synonyms with the weights as boost factors. The *tie_breaker* was set to 0.8 and the title field had a 3.0 boost factor, whereas that of the abstract field was 1.0. In addition, the Boolean query *should match* a list of keywords, including words such as "trial" and "patient" that are chosen empirically to serve as a weak classifier for evidence-based PM papers.

TREC PM allowed a maximum number of 1000 documents per topic in the submission. We set the maximum number of retrieved documents for each topic as 10,000. On average, we retrieved 1589 candidates from the baseline retriever for each topic.

Evidence Reranker

Overview

The Evidence Re-ranker scores a given candidate article *a* based on its evidence quality for the query *q* by:

where r_i is the output score, which is a weighted sum of: (1) a linear regressor (LR) using the features of the ElasticSearch score (es), pretrained BioBERT (pb), publication type (ty), and citation count (ct); and (2) a fine-tuned BioBERT (FB). w_{LR} and w_{FB} are the corresponding weights of the LR and FB. The FB is trained by the expert-in-the-loop active learning strategy, and the LR is trained by expert annotations.

Expert-in-the-Loop BioBERT

BioBERT [19] is a biomedical version of BERT that is trained on PubMed abstracts and PubMed Central articles. BioBERT achieves state-of-the-art performance on several biomedical natural language processing tasks. We followed the same setting as Nogueira et al [18] to use BioBERT in this task: to predict the evidence quality of a candidate article *a* for the query *q*, we first feed the concatenated *q* and *a* to the BioBERT, getting the pair representation *h*:

where *q* is the concatenated disease *d*, gene variant *g*, and treatment *t* in the query; *a* is the concatenated title and abstract of the article; and [SEP] is a special token in BERT to mark the input segments. A sigmoid layer is applied to the [CLS] representation *h* to predict the evidence quality :

where σ denotes the sigmoid function, *w* and *b* are the layer weights. During fine-tuning, we minimized the mean square loss between the predicted evidence quality and the expert-labeled score *r*. BioBERT fine-tuning is implemented using Huggingface's transformers Python package [23]. We use the Adam optimizer [24] with a learning rate of 4 × 10^{-5}, batch size of 16, and fine-tuning epoch number of 10 in each iteration.

We show the expert-in-the-loop active learning procedure in (Figure 2). At each iteration, a biomedical expert (senior MD candidate) annotates the evidence quality of the highest-ranked unannotated document for the given query based on the criteria shown in (Figure 3). This is similar to the top-1 active feedback setting described in Shen and Zhai [25]. Subsequently, we fine-tuned the original BioBERT with all available annotations at this iteration (ie, the newly annotated instances plus all available annotations from the last iteration) and then used the fine-tuned BioBERT to update the predictions for all documents, leading to the new document rankings. Again, the new document rankings were sent to the expert for annotations. We performed 22 iterations of the expert-in-the-loop active learning, where in most iterations, 40 new annotations were added (1 for each topic), resulting in 950 annotations in total. We also randomly sampled 100 topic-article pairs to be annotated by another medical doctor. The Pearson correlation was 0.853 between the annotation scores of 2 annotators, indicating a high level of interannotator agreement.

Figure 2. The architecture of our expert-in-the-loop active learning strategy. BioBERT: Bidirectional Encoder Representations from Transformers for Biomedical Text Mining; Y: yes; N: no.

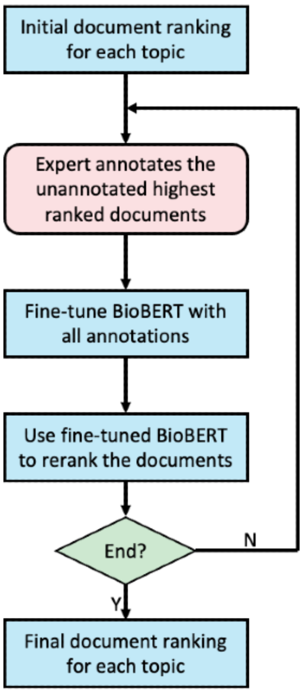
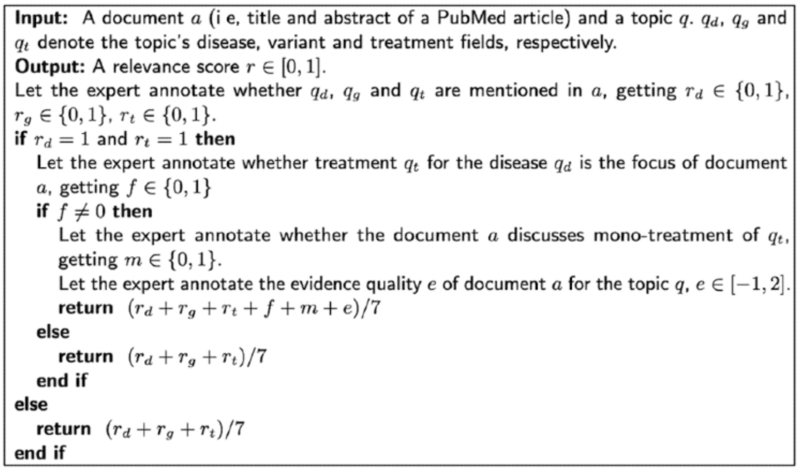

Initial document ranking for each topic
Expert annotates the unannotated highest ranked documents
Fine-tune BioBERT with all annotations
Use fine-tuned BioBERT to rerank the documents
End?
N
Y
Final document ranking for each topic

Figure 3. The expert annotation pipeline.



```
Input: A document $a$ (i e, title and abstract of a PubMed article) and a topic $q$. $q_d$, $q_g$ and $q_t$ denote the topic's disease, variant and treatment fields, respectively.
Output: A relevance score $r \in [0, 1]$.
Let the expert annotate whether $q_d$, $q_g$ and $q_t$ are mentioned in $a$, getting $r_d \in \{0, 1\}$, $r_g \in \{0, 1\}$, $r_t \in \{0, 1\}$.
if $r_d = 1$ and $r_t = 1$ then
  Let the expert annotate whether treatment $q_t$ for the disease $q_d$ is the focus of document $a$, getting $f \in \{0, 1\}$
  if $f \neq 0$ then
    Let the expert annotate whether the document $a$ discusses mono-treatment of $q_t$, getting $m \in \{0, 1\}$.
    Let the expert annotate the evidence quality $e$ of document $a$ for the topic $q$, $e \in [-1, 2]$.
    return $(r_d + r_g + r_t + f + m + e)/7$
  else
    return $(r_d + r_g + r_t)/7$
  end if
else
  return $(r_d + r_g + r_t)/7$
end if
```

Linear Regressor

We used the expert annotations to train a simple linear regression model using the following features:

1. es: the relevance scores returned by the ElasticSearch;
2. pb: the relevance scores predicted by a pretrained BioBERT. We used the annotations from the previous TREC PM challenges to fine-tune the BioBERT. Specifically, we

collected 54,500 topic-document relevance annotations from the *qrel* files of TREC PM 2017-2019, where the queries contained disease, gene variant, and demographics information but not the treatment option. To ensure consistency, we only used the disease and gene variant fields of the queries as input and fine-tuned the BioBERT to predict their normalized relevance in the annotations. We denoted this as “pretrained” BioBERT since the training data were formatted differently from the data of TREC PM 2020;

3. ty: the publication type score. PubMed also indexes each article with a publication type, such as journal article, review, clinical trials, etc. We manually rated the score of each publication type based on the judgments of their evidence quality. Our publication type and score mapping is shown in Table 2;
4. ct: the citation count score. We ranked the citation count of all PubMed articles and used the quantile of a specific article’s citation count as a feature. Similar to but simpler than PageRank [26], this feature was designed to reflect the community-level importance of each article.

The linear regression was implemented using the *sklearn* Python package, which basically minimizes the residual sum of squares between the expert annotations and the predictions from the linear approximation.

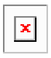
Table 2. Mappings between publication types and clinical evidence quality scores.

Publication type	Score
Comment	–1
Editorial	–1
Published erratum	–2
Retraction of publication	–2
English abstract	0
Journal article	0
Letter	0
Review	0
Case reports	1
Observational study	1
Clinical trial	2
Meta-analysis	2
Systematic review	2

Experiment Settings

We compared our PM-Search submissions to TREC PM 2020 with models submitted by other teams. We used 5 settings in the challenge, namely *PM-Search-auto-1*, *PM-Search-auto-2*, *PM-Search-full-1*, *PM-Search-full-2*, and *PM-Search-full-3*. They use different rerankers to rank the same set of documents retrieved by the baseline retriever. *PM-Search-full-1*, *PM-Search-full-2*, and *PM-Search-full-3* use the evidence reranker. They use the full PM-Search architecture with different combining weights in the evidence reranker.

We also used the *PM-Search-auto-1* and *PM-Search-auto-2* settings that do not use the expert-in-the-loop active learning strategy. Since these settings do not rely on expert annotations, they are considered as the “automatic” runs by the TREC challenge. Specifically, the reranking scores of article *a* for a given query in *PM-Search-auto-1* and *PM-Search-auto-2* are calculated as a weighted sum of the LR features:



where es_a, pb_a, ty_a, ct_a are the features of document *a*; es_{max}, pb_{max}, ty_{max}, ct_{max} are the corresponding maximum feature values among all documents; and w_{es}, w_{pb}, w_{ty}, and w_{ct} are the weights associated with different features and are determined empirically. The feature weights of the submitted systems are shown in Table 3.

Table 3. Feature weights in different systems. Participant denotes the system name submitted to the Text Retrieval Conference (TREC) Precision Medicine (PM).

System	TREC run Id	w_{es}[a]	w_{pb}[b]	w_{ty}[c]	w_{ct}[d]	w_{LR}[e]	w_{FB}[f]
PM-Search runs							
PM-Search-auto-1	damoespb1	1.0	0.5	1.5	0.0	__[g]	—
PM-Search-auto-2	damoespb2	1.0	0.5	1.0	0.0	—	—
PM-Search-full-1	damoespcbh1	–0.465	–0.141	–0.617	–0.005	1.0	1.0
PM-Search-full-2	damoespcbh2	–0.465	–0.141	–0.617	–0.005	1.0	2.0
PM-Search-full-3	damoespcbh3	–0.465	–0.141	–0.617	–0.005	1.0	5.0
Ablations							
Retriever + pb	N/A[h]	1.0	1.0	0.0	0.0	—	—
Retriever + ty	N/A	1.0	0.0	1.0	0.0	—	—
Retriever + ct	N/A	1.0	0.0	0.0	1.0	—	—
LR	N/A	–0.465	–0.141	–0.617	–0.005	1.0	0.0
FB	N/A	–0.465	–0.141	–0.617	–0.005	0.0	1.0

[a] es: ElasticSearch score.

[b] pb: pretrained BioBERT.

[c] ty: publication type.

[d] ct: citation count.

[e] LR: linear regressor.

[f] FB: fine-tuned BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining).

[g] Not available.

[h] N/A: not applicable.

Results

Main Results

The main results of our participating systems in the TREC PM 2020, compared with the other top-ranking systems, are shown in Table 4 [9].

Table 4. Topic-wise averaged performance of different settings in the evaluation. All numbers are percentages. Other top-ranking Text Retrieval Conference (TREC) submissions listed in the table include the systems of BIT.UA [27], CSIROMed [28], and h2oloo [29].

	Evidence quality (phase 2)		General relevance (phase 1)		
	NDCG@30[a], exponential	NDCG@30, standard	infNDCG[b]	P@10[c]	R-prec[d]
All TREC runs					
First	45.19 (ours)	47.80 (ours)	53.25[27]	56.45 [28]	43.58 [28]
Second	41.93* [29]	42.38* [29]	53.03 [28]	55.16 [27]	42.07 [27]
Median	28.57	25.29	43.16	46.45	32.59
PM-Search runs					
PM-Search-full-3	45.19	47.80	44.24	47.42	34.72
PM-Search-full-1	44.97	47.30	43.04	47.42	34.10
PM-Search-full-2	44.95	47.46	43.84	47.10	34.14
PM-Search-auto-1	42.55	44.17*	45.33	47.42	35.93
PM-Search-auto-2	42.54	44.60*	41.12	44.52	32.37
Ablations					
Retriever + pb[e]	32.36*	37.04*	52.26	53.87	41.21
Retriever + ty[f]	41.46*	43.26*	37.80	40.32	29.37
Retriever + ct[g]	35.55*	38.40*	42.20	44.84	32.52
Linear regressor	42.86*	44.86*	37.65	46.13	30.74
Linear regressor, leave-one-out	42.08*	43.81*	37.06	46.45	30.58
Fine-tuned BioBERT[h]	44.40*	47.01*	44.59	47.42	34.87
Fine-tuned BioBERT, leave-one-out	44.15*	46.58*	43.83*	46.45*	33.81*

[a] NDCG@30: normalized discounted cumulative gain NDCG at rank 30.

[b] infNDCG: inferred NDCG.

[c] P@10: precision at rank 10.

[d] R-prec: R-precision.

[e] pb: pretrained BioBERT.

[f] ty: publication type.

[g] ct: citation count.

[h] BioBERT: Bidirectional Encoder Representations from Transformers for Biomedical Text Mining.

*Significant differences from the PM-Search-full-3. Significance is defined as $P<.05$ in 2-sided paired *t* test.

General Relevance (Phase 1)

Our submissions scored higher than the topic-wise median submission, but the best submission (infNDCG: 0.5325, P@10: 0.5645, R-prec: 0.4358) outperformed our submissions (infNDCG: 0.4533, P@10: 0.4742, R-prec: 0.3593). Our PM-Search runs (*PM-Search-full-1* to *3*; ie, PM-Search) showed no significant improvements over the runs without active learning (*PM-Search-auto-1* and *2*). It is not surprising, since we focused on modeling evidence quality, and articles that are highly related to the queries but are of low evidence quality (eg, narrative reviews) will be ranked lower. As a result, our submissions performed only moderately in the phase 1 assessment that mainly judges the general relevance.

Evidence Quality (Phase 2)

Our PM-Search system *PM-Search-full-3* achieved the highest scores for standard gain NDCG@30 of 0.4780 and exponential gain NDCG@30 of 0.4519. As expected, the *PM-Search-full* settings outperform the *PM-Search-auto* settings that only use the features (0.4503 vs 0.4255 for averaged exponential NDCG@30). This shows that our expert annotation procedure as well as the expert-in-the-loop active learning strategy can improve the performance of evidence quality ranking. Remarkably, all our settings outperform the second-best system (0.4238 for standard NDCG@30 and 0.4193 for exponential NDCG@30) [29], including the *PM-Search-auto* settings that do not rely on expert annotations (exponential NDCG@30: 0.4255). The results show that the proposed PM-Search system is a robust evidence retriever that can be potentially applied to assist the practice of PM.

Ablations and Feature Importance

We also experimented with different settings and studied the importance of PM-Search components, including the baseline retriever, active learning, and the reranking features.

Baseline Retriever Settings

In Table 5, we show the performance of the baseline retriever without query expansion or keyword matching. The results show that query expansion is an important module to improve the recall of relevant articles. However, we find that boosting keywords such as "trial" and "patient" do not significantly change the performance. This is inconsistent with the study of Faessler et al [10], which shows that boosting a range of keywords helps improve the performance. One key difference between our system and Faessler et al [10] is that we only use 2 positive keywords, whereas they use various positive and negative keywords, so increasing the number and diversity of keywords could be a future work for improvements.

Table 5. Ablation results of different baseline retriever settings (in percentages).

Method	Evidence quality (phase 2)			General relevance (phase 1)		
	R@0.5k[a]	R@1k[b]	R@10k[c]	R@0.5k	R@1k	R@10k
Baseline retriever	68.99	75.96	81.00	65.51	72.30	77.71
Baseline retriever without query expansion	66.84*	72.61*	76.94*	61.85*	67.21*	72.90*
Baseline retriever without keyword matching	68.85	76.06	81.00	65.65	72.33	77.71

[a] R@0.5k: recall at the top 500 positions.

[b] R@1k: recall at the top 1000 positions.

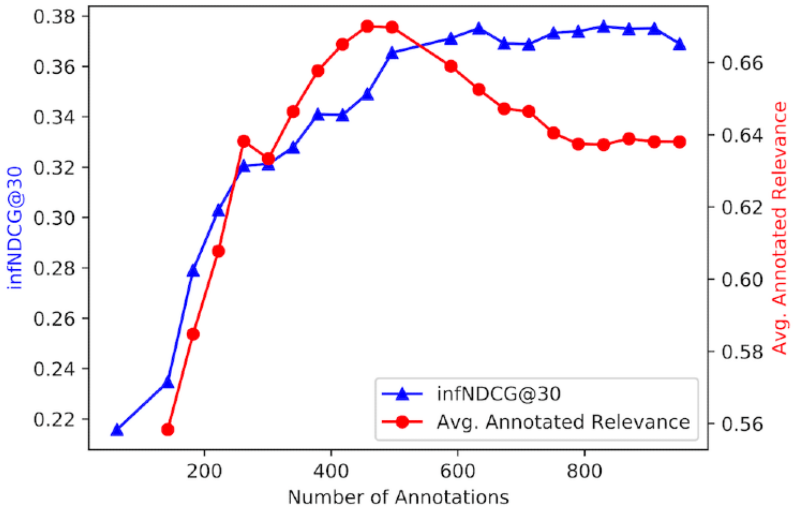
[c] R@10k: recall at the top 10,000 positions.

*Significant differences than the original retrieval. Significance is defined as $P<.05$ in 2-sided paired t test.

Active Learning

In Figure 4, we show the performance of the BioBERT predictions at each iteration in active learning, evaluated with infNDCG@30 by the evidence quality (phase 2) assessments. The performance increases with the iteration when the number of annotations is less than 500 and then converges after the number of annotations is greater than 500. Interestingly, we find that the average annotated relevance by our annotator also reaches its maximum at around 500 annotations, which indicates that this metric can be empirically used as the stop criterion.

Figure 4. InfNDCG@30 and average annotated relevance at each iteration in active learning. InfNDCG@30: inferred normalized discounted cumulative gain at rank 30.



Reranker Features

To analyze the importance of the used features, we show the ablation experiments in Table 4 and Pearson correlations between them and the official scores in both phases in Table 6.

General relevance (phase 1): BioBERT that is further pretrained by the annotations of previous TREC PM (pb) had the highest correlation (0.5771) with the phase 1 scores, and the baseline retriever with the pretrained BioBERT had the highest performance (infNDCG: 52.26%) in our ablation experiments. This is probably because the evaluations of previous tasks are also based on general relevance. The ElasticSearch scores (es) achieved the second highest correlation of 0.3892, and the fine-tuned BioBERT by active learning (FB) had a Pearson correlation of 0.3733. However, our expert annotations for the evidence quality only had a Pearson correlation of 0.2157 with the general relevance scores, which indicates that generally relevant papers might not have high evidence quality. In addition, the features of publication types (ty) and the citation counts (ct), which are designed for the evidence quality ranking and are positively correlated with the evidence quality, were negatively correlated with the general relevance scores.

Evidence quality (phase 2): The trends of ablation results and correlations between features and evidence quality scores were similar in both the standard and exponential scores. The most important features in the evidence quality evaluation included publication types and active learning. Interestingly, only using the publication type and the baseline retriever achieves comparable performance to the second-best system in TREC PM (0.4146 vs 0.4193 for exponential NDCG@30). BioBERT fine-tuned by the expert annotations (FB) had the highest performance in the ablation experiments (exponential NDCG@30: 0.4440) and its correlation to the official annotations was close to that of our expert annotations (0.3309 vs 0.2937 for exponential gains; 0.2847 vs 0.3073 for standard gains). Besides, the fine-tuned BioBERT outperformed the expert annotations by a large margin (0.3733 vs 0.2157) in the phase 1 assessment, indicating that it can rerank the documents by evidence quality while retaining the original general relevance ranks to some extent. The most correlated features of phase 1, that is, the pretrained BioBERT (pb) and the ElasticSearch score (es), had the lowest correlations with the phase 2 scores, which further confirms that the evidence quality assessment is distinct from the general relevance assessment.

In summary, the 2 assessment phases might have opposite considerations since features that are highly related to the score of one phase tended to be much less related to the score of the other phase, with the exception of the fine-tuned BioBERT. As a result, specific modeling of evidence quality beyond general relevance is required for a PM search engine.

Table 6. Feature correlations to the official scores.

Features	es[a]	pb[b]	ty[c]	ct[d]	LR[e]	FB[f]	Expert annotation
General relevance	0.3892	*0.5771*	–0.0621	–0.0435	0.1341	0.3733	0.2157
Evidence quality							
Standard gains	0.0752	0.0621	0.2564	0.0696	0.2728	*0.3309*	0.2937
Exponential gains	0.0474	0.0338	0.2772	0.0806	0.2816	0.2847	*0.3073*

[a] es: ElasticSearch score.

[b] pb: pretrained Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT).

[c] ty: publication type.

[d] ct: citation count.

[e] LR: linear regressor.

[f] FB: fine-tuned BioBERT.

Discussion

Topic-Level Generalizability Analysis

Each instance used to train the PM-Search reranker contained a topic-article pair and its relevance score. The main results show that PM-Search is generalizable at *instance-level*, where the model is trained and evaluated by different instances. However, *topic-level* generalizability of the PM-Search was not evaluated since our expert annotations and the official annotations, that is, the training and evaluation instances, used the same set of topics.

Here, we analyze how PM-Search generalizes to unseen topics using a leave-one-out evaluation strategy. Each time, we use the official annotations of only one topic to evaluate the models that are trained by our expert annotations without the evaluation topic. The results of each topic as the evaluation topic are calculated and the averaged performance is shown in Table 4. The leave-one-out results are close to the results when all expert annotations are used for training: 0.4415 versus 0.4440 for exponential NDCG@30 and 0.4658 versus 0.4710 for standard NDCG@30. This shows that the model is also generalizable to unseen topics.

Error Analysis

We show several typical cases in Table 7 to qualitatively analyze some errors in the evidence quality assessment. It should be noted that most errors cannot be attributed to a specific cause since the predictions of BioBERT are not explainable, so developing explainable models is a vital future direction to explore.

Table 7. Typical error cases in the evidence quality assessment. Topics are shown in Table 1.

Case	Topic	Article	Official, rank (normalized relevance)	PM[a]-Search, rank (normalized relevance)	Error type
1	1	PMID[b]: 23177515; Title: Efficacy and safety of regorafenib for advanced gastrointestinal stromal tumours after failure of imatinib and sunitinib (GRID): an international, multicentre, randomised, placebo-controlled, phase 3 trial	1 (1.00)	N/A[c]	Concept recognition
2	1	PMID: 24150533; Title: Risk of hypertension with regorafenib in cancer patients: a systematic review and meta-analysis	1 (1.00)	148 (0.47)	Different understanding
3	1	PMID: 25213161; Title: Randomized phase III trial of regorafenib in metastatic colorectal cancer: analysis of the CORRECT Japanese and non-Japanese subpopulations	1 (1.00)	297 (0.29)	Unclassified
4	11	PMID: 29147869; Title: Hematological adverse effects in breast cancer patients treated with cyclin-dependent kinase 4 and 6 inhibitors: a systematic review and meta-analysis	1 (1.00)	N/A	Full article visibility
5	11	PMID: 28540640; Title: A Population Pharmacokinetic and Pharmacodynamic Analysis of Abemaciclib in a Phase I Clinical Trial in Cancer Patients	1 (1.00)	53 (0.50)	Full article visibility
6	11	PMID: 29700711; Title: Cyclin-dependent kinase 4/6 inhibitors in hormone receptor-positive early breast cancer: preliminary results and ongoing studies	61 (0.25)	6 (0.71)	Different understanding

[a] PM: precision medicine.

[b] PMID: PubMed IDentifier.

[c] N/A: not applicable.

Full Article Visibility

The PM-Search system can only access the title and abstract of PubMed articles. However, vital article information (eg, detailed gene variant types, treatments) might only appear in the full article, especially for meta-analyses and systematic reviews where abstracts tend to use more general concepts. For example, PM-Search fails to retrieve the Case 5 article where the queried disease "breast cancer" is only mentioned in the full article, not in the abstract. For this, future models can use the full article information from PubMed Central to better retrieve and rank relevant papers.

Different Understanding

In some cases, we have a different understanding of how clinically significant the evidence is that an article provides. For example, the article "Risk of hypertension with regorafenib in cancer patients: a systematic review and meta-analysis" in Case 2 is focused on the hypertension side effect of the therapy, not the therapeutic effects, which we think is not significant. However, it was given the highest score in the official evaluation but ranked much lower in the PM-Search prediction. This issue should be solved by community efforts for the development of standards.

Concept Recognition

The baseline retriever of PM-Search uses query expansion to recognize relevant concepts in the article. However, this step is error prone since biomedical terms are highly variable and thus cannot be represented by a list of synonyms. For example, in Case 1, the "colorectal cancer" in the query appears as "gastrointestinal stromal tumours" in the article, which was missed in the query expansion step of PM-Search. As a result, this article was not returned by the PM-Search but ranked the highest in the official assessment. Improving similar concept recognition, such as using distributed representations of concepts, remains an important direction to explore.

Comparison With Prior Work

Many IR systems for precision medicine have been proposed in the TREC PM tracks [7-9,30], where the key issue to solve is that queries and their related documents might use different terms to describe the same concepts. Some studies [31-33] have attempted to use BERT-based models for ranking in previous TREC PM tracks, showing various levels of improvements. Thalia is a semantic search engine for biomedical abstracts that is updated on a daily basis [34]. It tackles the vocabulary mismatch problem by mapping the queries to predefined concepts by which the documents are indexed. The HPI-DHC team shows that query expansion associated with hand-crafted rules improves the retrieval performance [35]. Faessler et al [10,36] systematically analyze the individual contributions of relevant system features such as BM25 weights, query expansion, and boosting settings. PRIMROSE is a PM search engine that expands the queries with an internal knowledge graph [37]. Noh and Kavuluru [38] use a basic BERT with specific components for reranking. Koopman et al [39] present a search engine for clinicians to find tailored treatments for children with cancer. For the vocabulary mismatch issue, PM-Search uses a similar query expansion strategy to previous studies. However, PM-Search differs from all prior work in that it is specifically designed to rank the retrieval results by their evidence quality, which is an important feature for PM search engines.

Conclusions and Future Work

In this paper, we present PM-Search, a search engine for PM that achieved state-of-the-art performance in TREC PM 2020. PM-Search uses an ElasticSearch-based baseline retriever with query expansion and keyword matching and an evidence reranker that uses the BioBERT fine-tuned by an active learning strategy. Our analyses show that the evidence quality is a distinct aspect from the general relevance, and thus, specific modeling of it is necessary to assist the practices for evidence-based PM.

The deployment and evaluation of PM-Search in real clinical settings remains a clear future direction. It is also worth exploring the use of dense vectors for baseline retrieval and incorporating full-text information into the ranking process.

Acknowledgments

We thank the organizers of Text Retrieval Conference (TREC) Precision Medicine (PM) 2020 for their efforts in holding this task and manual evaluations of the submitted systems.

Conflicts of Interest

None declared.

References

1. Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials 2010 Aug 12;11(1):85 [FREE Full text] [doi: 10.1186/1745-6215-11-85] [Medline: 20704705]
2. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med 2015 Feb 26;372(9):793-795 [FREE Full text] [doi: 10.1056/NEJMp1500523] [Medline: 25635347]
3. Paez JG, Jänne PA, Lee JC, Tracy S, Greulich H, Gabriel S, et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science 2004 Jun 04;304(5676):1497-1500. [doi: 10.1126/science.1099314] [Medline: 15118125]
4. Romond EH, Perez EA, Bryant J, Suman VJ, Geyer CE, Davidson NE, et al. Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. N Engl J Med 2005 Oct 20;353(16):1673-1684. [doi: 10.1056/NEJMoa052122] [Medline: 16236738]
5. Sackett DL. Evidence-based medicine. Semin Perinatol 1997 Feb;21(1):3-5. [doi: 10.1016/s0146-0005(97)80013-4] [Medline: 9190027]
6. Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, Lazar AJ, et al. Overview of the TREC 2017 Precision Medicine track. Text Retr Conf 2017 Nov;26 [FREE Full text] [Medline: 32776021]
7. Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, Lazar AJ. Overview of the TREC 2018 Precision Medicine track. Text Retr Conf 2018 [FREE Full text]
8. Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, Lazar AJ, et al. Overview of the TREC 2019 Precision Medicine track. Text Retr Conf 2019 Nov;1250 [FREE Full text] [Medline: 34849512]
9. Roberts K, Demner-Fushman D, Voorhees EM, Bedrick S, Hersh WR. Overview of the TREC 2020 Precision Medicine track. Text Retr Conf 2020 Nov;1266 [FREE Full text] [Medline: 34849513]
10. Faessler E, Oleynik M, Hahn U. What makes a top-performing precision medicine search engine? tracing main system features in a systematic way. 2020 Jul 25 Presented at: SIGIR '20: the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval; July 25-30, 2020; Virtual event, China p. 459-468. [doi: 10.1145/3397271.3401048]
11. Nguyen V, Karimi S, Jin B. An experimentation platform for precision medicine. 2019 Jul 18 Presented at: SIGIR'19: the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval; July 21-25, 2019; Paris, France p. 1357-1360. [doi: 10.1145/3331184.3331396]
12. Stokes N, Li Y, Cavedon L, Zobel J. Exploring criteria for successful query expansion in the genomic domain. Inf Retrieval 2008 Oct 29;12(1):17-50. [doi: 10.1007/s10791-008-9073-9]
13. Craswell N, Mitra B, Yilmaz E, Campos D, Voorhees EM. Overview of the TREC 2019 deep learning track. arXiv 2020 Mar 18:1-22. [doi: 10.48550/arXiv.2003.07820]
14. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K. Deep contextualized word representations. 2018 Jun Presented at: The 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 1-6, 2018; New Orleans, LA p. 2227-2237. [doi: 10.18653/v1/n18-1202]

15. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. 2019 Jun Presented at: The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); June 2-7, 2019; Minneapolis, MN p. 4171-4186.
16. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman S. GLUE: a multi-task benchmark and analysis platform for natural language understanding. 2018 Nov Presented at: The 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP; November 1, 2018; Brussels, Belgium p. 353-355. [doi: 10.18653/v1/w18-5446]
17. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in neural information processing systems. 2017 Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); December 4-9, 2017; Long Beach, CA p. 5998-6008 URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
18. Nogueira R, Yang W, Cho K, Lin J. Multi-stage document ranking with bert. arXiv 2019 Oct 31:1-13. [doi: 10.48550/arXiv.1910.14424]
19. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
20. MedlinePlus API. National Library of Medicine. URL: https://ghr.nlm.nih.gov/condition/{d}?report=json [accessed 2022-11-29]
21. MedlinePlus API. National Library of Medicine. URL: https://ghr.nlm.nih.gov/gene/{g}?report=json [accessed 2022-11-29]
22. Robertson S, Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond. FNT in Information Retrieval 2009 Dec 17;3(4):333-389. [doi: 10.1561/1500000019]
23. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A. Transformers: state-of-the-art natural language processing. 2020 Oct Presented at: The 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; November 16-20, 2020; Online Virtual Conference p. 38-45. [doi: 10.18653/v1/2020.emnlp-demos.6]
24. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv 2017 Jan 30:1-15. [doi: 10.48550/arXiv.1412.6980]
25. Shen X, Zhai C. Active feedback in ad hoc information retrieval. 2005 Aug 15 Presented at: SIGIR '05: the 28th annual international ACM SIGIR conference on Research and development in information retrieval; August 15-19, 2005; Salvador, Brazil p. 59-66. [doi: 10.1145/1076034.1076047]
26. Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: bringing order to the web. Stanford InfoLab 1998 Jan 29:1-17 [FREE Full text]
27. Almeida T, Matos S. BIT.UA@ TREC Precision Medicine track. 2020 Presented at: The Twenty-Ninth Text REtrieval Conference, TREC 2020; November 16-20, 2020; Virtual Event (Gaithersburg, MD) URL: https://trec.nist.gov/pubs/trec29/papers/BIT.UA.PM.pdf
28. Karimi M. CSIROmed at TREC Precision Medicine. 2020 Nov Presented at: The Twenty-Ninth Text REtrieval Conference (TREC 2020); November 16-20, 2020; Virtual Event (Gaithersburg, MD) URL: https://trec.nist.gov/pubs/trec29/papers/CSIROmed.PM.pdf
29. Pradeep R, Ma X, Zhang X, Cui H, Xu R, Nogueira R. H2oloo at TREC: when all you got is a hammer... deep learning, health misinformation, and precision medicine. 2020 Presented at: The Twenty-Ninth Text REtrieval Conference (TREC 2020); November 16-20, 2020; Virtual Event (Gaithersburg, MD) URL: https://trec.nist.gov/pubs/trec29/papers/h2oloo.DL.HM.PM.pdf
30. Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, Lazar A, et al. Overview of the TREC 2017 Precision Medicine track. Text Retr Conf 2017 Nov;26 [FREE Full text] [Medline: 32776021]
31. Jo SH, Lee KS. CBNU at TREC 2019 Precision Medicine track. 2019 Presented at: The Twenty-Eighth Text REtrieval Conference, TREC 2019; November 13-15, 2019; Gaithersburg, MD URL: https://trec.nist.gov/pubs/trec28/papers/cbnu.PM.pdf
32. Liu X, Li L, Yang Z, Dong S. SCUT-CCNL at TREC 2019 Precision Medicine track. 2019 Presented at: The Twenty-Eighth Text REtrieval Conference, TREC 2019; November 13-15, 2019; Gaithersburg, MD URL: https://trec.nist.gov/pubs/trec28/papers/CCNL.PM.pdf
33. Zheng Q, Li Y, Hu J, Yang Y, He L, Xue Y. ECNU-ICA team at TREC 2019 Precision Medicine track. 2019 Presented at: The Twenty-Eighth Text REtrieval Conference, TREC 2019; November 13-15, 2019; Gaithersburg, MD URL: https://trec.nist.gov/pubs/trec28/papers/ECNU_ICA.PM.pdf
34. Soto A, Przybyła P, Ananiadou S. Thalia: semantic search engine for biomedical abstracts. Bioinformatics 2019 May 15;35(10):1799-1801 [FREE Full text] [doi: 10.1093/bioinformatics/bty871] [Medline: 30329013]
35. Oleynik M, Faessler E, Sasso A, Kappattanavar A, Bergner B, Da CH. HPI-DHC at TREC 2018 Precision Medicine track. 2018 Presented at: The Twenty-Seventh Text REtrieval Conference (TREC 2018); November 14-16, 2018; Gaithersburg, MD URL: https://trec.nist.gov/pubs/trec27/papers/hpi-dhc-PM.pdf
36. Faessler E, Hahn U, Oleynik M. 2019 Presented at: The Twenty-Eighth Text REtrieval Conference, TREC 2019; November 13-15, 2019; Gaithersburg, MD URL: https://trec.nist.gov/pubs/trec28/papers/julie-mug.PM.pdf
37. Shenoi SJ, Ly V, Soni S, Roberts K. Developing a Search Engine for Precision Medicine. AMIA Jt Summits Transl Sci Proc 2020 May 30;2020:579-588 [FREE Full text] [Medline: 32477680]

38. Noh J, Kavuluru R. Literature retrieval for precision medicine with neural matching and faceted summarization. Proc Conf Empir Methods Nat Lang Process 2020 Nov;2020:3389-3399 [FREE Full text] [doi: 10.18653/v1/2020.findings-emnlp.304] [Medline: 34541588]
39. Koopman B, Wright T, Omer N, McCabe V, Zuccon G. Precision medicine search for paediatric oncology. 2021 Jul 11 Presented at: SIGIR '21: the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval; July 11-15, 2021; Virtual event, Canada p. 2536-2540. [doi: 10.1145/3404835.3462792]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
BioBERT: Bidirectional Encoder Representations from Transformers for Biomedical Text Mining
BNDCG@30: normalized discounted cumulative gain NDCG at rank 30
ct: citation count
es: ElasticSearch score
FB: fine-tuned BioBERT
infNDCG: inferred normalized discounted cumulative gain
IR: information retrieval
LR: linear regressor
NDCG: normalized discounted cumulative gain
NDCG@30: NDCG at rank 30
P@10: precision at rank 10
pb: pretrained BioBERT
PM: precision medicine
R-prec: R-precision
TREC: Text Retrieval Conference
ty: publication type

Edited by M Focsa; submitted 04.07.22; peer-reviewed by R Haynes, C Lokker; comments to author 05.10.22; revised version received 13.11.22; accepted 16.11.22; published 15.12.22.

Please cite as:
Jin Q, Tan C, Chen M, Yan M, Zhang N, Huang S, Liu X
State-of-the-Art Evidence Retriever for Precision Medicine: Algorithm Development and Validation
JMIR Med Inform 2022;10(12):e40743
URL: https://medinform.jmir.org/2022/12/e40743
doi:10.2196/40743
PMID:36409468

Original Paper

Pan-Canadian Electronic Medical Record Diagnostic and Unstructured Text Data for Capturing PTSD: Retrospective Observational Study

Leanne Kosowan[1], MSc; Alexander Singer[1], MB BAO BCh, CCFP; Farhana Zulkernine[2], PhD; Hasan Zafari[2], PhD; Marcello Nesca[3], MSc; Dhasni Muthumuni[4], MD

[1]Department of Family Medicine, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, MB, Canada
[2]School of Computing, Queen's University, Kingston, ON, Canada
[3]Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, MB, Canada
[4]Department of Psychiatry, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, MB, Canada

Corresponding Author:
Alexander Singer, MB BAO BCh, CCFP
Department of Family Medicine
Rady Faculty of Health Sciences
University of Manitoba
D009-780 Bannatyne Ave.
Winnipeg, MB, R3E0W2
Canada
Phone: 1 204 789 3314
Email: alexander.singer@umanitoba.ca

Abstract

Background: The availability of electronic medical record (EMR) free-text data for research varies. However, access to short diagnostic text fields is more widely available.

Objective: This study assesses agreement between free-text and short diagnostic text data from primary care EMR for identification of posttraumatic stress disorder (PTSD).

Methods: This retrospective cross-sectional study used EMR data from a pan-Canadian repository representing 1574 primary care providers at 265 clinics using 11 EMR vendors. Medical record review using free text and short diagnostic text fields of the EMR produced reference standards for PTSD. Agreement was assessed with sensitivity, specificity, positive predictive value, negative predictive value, and accuracy.

Results: Our reference set contained 327 patients with free text and short diagnostic text. Among these patients, agreement between free text and short diagnostic text had an accuracy of 93.6% (CI 90.4%-96.0%). In a single Canadian province, case definitions 1 and 4 had a sensitivity of 82.6% (CI 74.4%-89.0%) and specificity of 99.5% (CI 97.4%-100%). However, when the reference set was expanded to a pan-Canada reference (n=12,104 patients), case definition 4 had the strongest agreement (sensitivity: 91.1%, CI 90.1%-91.9%; specificity: 99.1%, CI 98.9%-99.3%).

Conclusions: Inclusion of free-text encounter notes during medical record review did not lead to improved capture of PTSD cases, nor did it lead to significant changes in case definition agreement. Within this pan-Canadian database, jurisdictional differences in diagnostic codes and EMR structure suggested the need to supplement diagnostic codes with natural language processing to capture PTSD. When unavailable, short diagnostic text can supplement free-text data for reference set creation and case validation. Application of the PTSD case definition can inform PTSD prevalence and characteristics.

(JMIR Med Inform 2022;10(12):e41312) doi:10.2196/41312

KEYWORDS

electronic health records; EHR; natural language processing; NLP; medical informatics; primary health care; stress disorders, posttraumatic; posttraumatic stress disorder; PTSD

Introduction

Primary care providers are typically the first point of contact for individuals within the health care system. Primary care services support patients throughout their health care experiences managing both acute and chronic conditions. Primary care electronic medical records (EMR) are a rich source of longitudinal patient data collected by health care providers throughout an individual's health care experience. EMR data can identify clinical phenotypes, describe care pathways, and inform quality improvement initiatives [1,2]. EMR-derived data typically include information related to patient characteristics, diagnoses, prescribed medications, and biometrics. They may also include information on social history, allergies, and risk factors for diseases [3-7]. Given the breadth of information available within EMRs, their use for disease surveillance continues to grow.

Identification of complex medical conditions may require multiple data points. Structured data fields such as standardized diagnosis or medication codes, as well as unstructured free-text data within the EMR can be assessed to describe complex conditions. Unstructured free text in the EMR can describe the observations, assessment, and plan for patient care providing depth to what is available in structured data fields [8,9]. More specifically, unstructured and short-text fields describe the patient context, including sociodemographic, risk behaviors and allergies, patient experience and interactions with the provider, and rational for the health care decisions that were made, which can inform disease surveillance and research [8]. Text analytics and, more specifically, natural language processing (NLP) of text data in the EMR can identify symptoms and variable interactions across multiple tables within data holdings [9-15]. Mining text data from health records typically includes refining procedures and knowledge extraction, aggregation, abstraction, and summarization of EMR information to transform text data into actionable insights such as inform phenotyping, disease prognosis and management, and disease surveillance [9,16,17]. Free-text information is not always available for research due to the technical limitations of EMR data systems or analysis, as well as privacy and data protection restrictions [18]. Due to this limitation, previous studies have relied on small data sets or a small number of institutions, preventing evidence of transferability of the models [17]. Primary care EMR short diagnostic text fields, more widely available than free-text data, have been suggested as a method for supplementing diagnostic definitions when free-text is unavailable [15,19,20]. Supplementation of free-text data with short-text fields, matched with refined processes for annotation and classification can support the use of EMR data in research [17].

Posttraumatic stress disorder (PTSD) is a complex mental health disorder characterized by a constellation of distressing symptoms that occur after witnessing or experiencing a traumatic event [21,22]. PTSD involves intrusive thoughts, persistent avoidance, negative alterations in cognition and mood, and alterations in arousal and reactivity (eg, irritability, reduced concentration, and exaggerated startle response) due to trauma recollection, which occur for greater than 1 month and result in significant impairment for the individual [20,22-24]. PTSD is associated with an array of multimodal risk indicators suggesting no single factor can account for the large variance in PTSD symptoms [19,20]. When encountered in primary care, PTSD is associated with considerable functional impairment and health care utilization [24]. This complex set of symptoms, combined with an individual's possible reluctance to seek help, infrequent patient-clinician interaction, and overlapping symptoms with other mental health conditions, makes PTSD difficult to accurately diagnose in primary care [20,22]. Identifying PTSD requires both depth and breadth to detail the patients' experience and capture associated factors [19,20].

This study had two objectives, which are as follows: (1) to compare the quality of capture when using free-text data compared to short diagnostic text fields from primary care EMRs for the creation of a reference set for a complex condition such as PTSD, and (2) test possible PTSD case definitions using single-province and pan-Canadian EMR reference standards. This study assesses the performance of 4 PTSD case definitions against reference standards to assess improved agreement when structured data fields are supplemented with NLP of EMR short diagnostic phrases.

Methods

Overview

This retrospective cross-sectional study used EMR data extracted and processed by the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). At the time of this study, there were 1574 consenting primary care providers (ie, family physicians, nurse practitioners, and community pediatricians) from 257 clinics representing 1,493,516 patients in 7 Canadian provinces (British Columbia, Alberta, Manitoba, Ontario, Quebec, Nova Scotia, and Newfoundland and Labrador) [3,7].

Data Sources

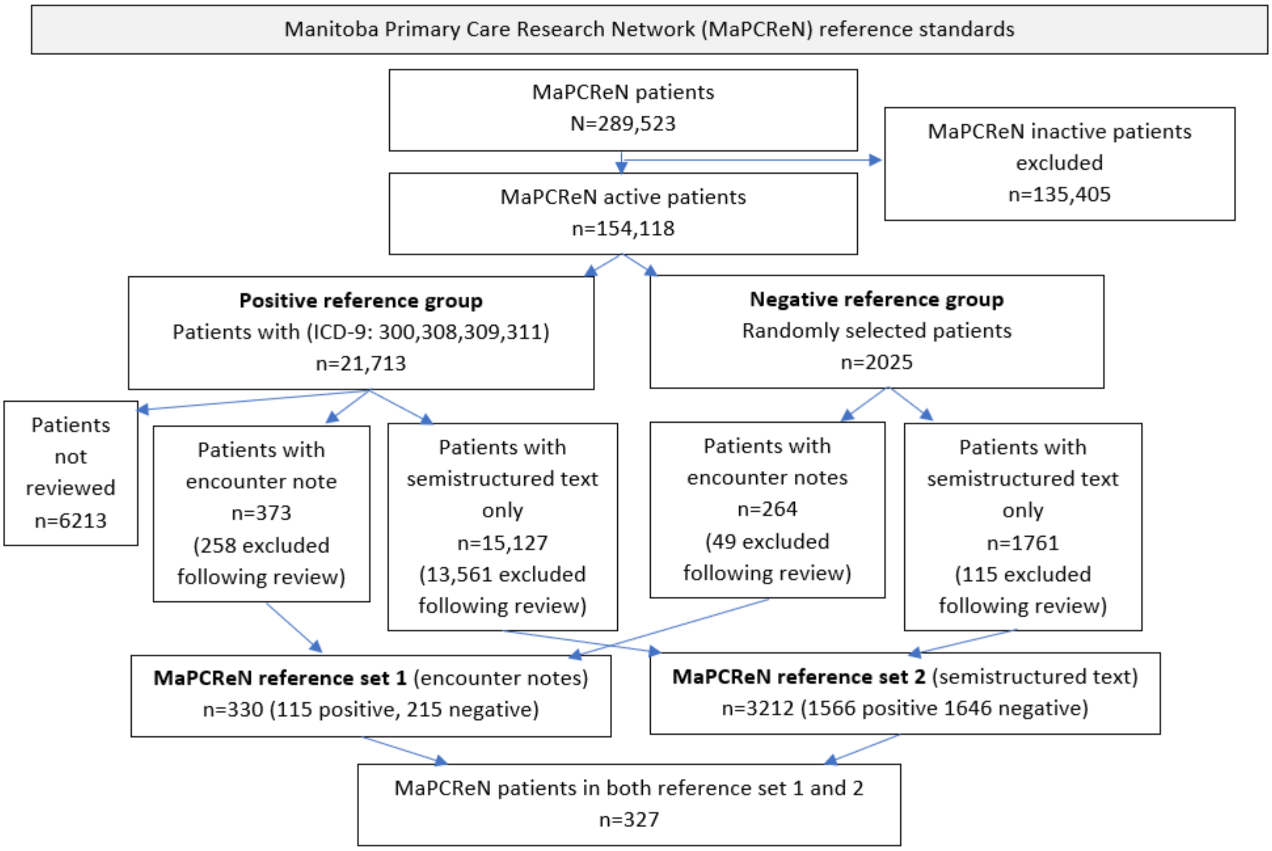
The CPCSSN repository is a pan-Canadian data set that is updated semiannually from regional practice-based research networks. The data in the repository comprised deidentified EMR data from consenting primary care providers that use 11 different EMR systems across Canada. Extracted EMR data are cleaned and standardized to map prescribed medications to Anatomical Therapeutic Chemical classification codes, laboratory tests to Logical Observation Identifiers Names and Codes, and medical diagnoses to International Classification of Disease, ninth edition, clinical modification (ICD-9-CM) codes. The CPCSSN repository also contains unstructured data in the form of short diagnostic text fields related to diagnoses, medication instructions, allergies, and social and behavioral risk factors. Additionally, some regional networks, such as the Manitoba Primary Care Research Network (MaPCReN), also extract free-text encounter notes that go through a deidentification algorithm to anonymize the data. Encounter notes are narrative entries created by primary care providers, typically structured in the problem-oriented medical record format [8]. MaPCReN represents 266 consenting primary care providers in 48 clinics in Manitoba, Canada. This study accessed a CPCSSN data set comprised of structured and short diagnostic text fields, and a MaPCReN data set containing structured, short diagnostic text fields, and free-text encounter notes.

Manitoba Primary Care Patients

The MaPCReN database includes 289,523 patients, of which 154,118 (52.23%) were considered active because they had seen a primary care provider participating in MaPCReN in the prior 2 years (between January 1, 2017, and December 31, 2019) [25]. In addition to structured and short diagnostic text data available for all patients, 19.6% (56,795/289,523) of the patients have free-text encounter notes available in the MaPCReN repository (2,125,961 encounter notes). Two medical students conducted a complete review of the medical records of a subset of patients from the MaPCReN repository. The reviewers were instructed to use the criteria from the Diagnostic and Statistics Manual of Mental Disorders, Fifth Edition [26] or specific documentation to indicate whether a patient was diagnosed with PTSD. A data extraction form was developed to capture patients living with PTSD and related signs or symptoms (Multimedia Appendix 1).

To create the subset for medical record review, we identified 21,713 patients with one more of the following ICD-9-CM codes in the health condition table of the EMR starting 300 (anxiety), 308 (acute reaction to stress), 309 (adjustment reaction), or 311 (depression). A total of 373 patients had a complete record reviewed by 2 students. Medical record review without free text was also completed by 2 medical students for 15,127 (69.67%) of these 21,713 patients to create positive reference sets. To identify patients without PTSD (negative reference set), patients were randomly selected for review by 2 medical students. In the negative reference set, 264/2025 (13.0%) patients had full medical records review (including free text), and 1761/2025 (87.0%) patients were reviewed without free-text encounter notes. Patients were labeled as “PTSD,” “possible PTSD,” or “no PTSD” in the data extraction form (Multimedia Appendix 1). Any discrepancies were reviewed by a family physician clinician researcher (AS). The final reference set included patients who were considered “PTSD” or “no PTSD” and excluded patients with “possible PTSD.” This process created the following two MaPCReN reference standards: (1) a total of 330 patients (n=115, 34.8% positive and n=215, 65.2% negative) had full medical record review including free-text data, and (2) a total of 3212 patients (n=1566, 48.75% positive and n=1646, 51.25% negative) had medical record review without free-text data. There were 327 patients who were included in both MaPCReN reference sets (Figure 1) [20].

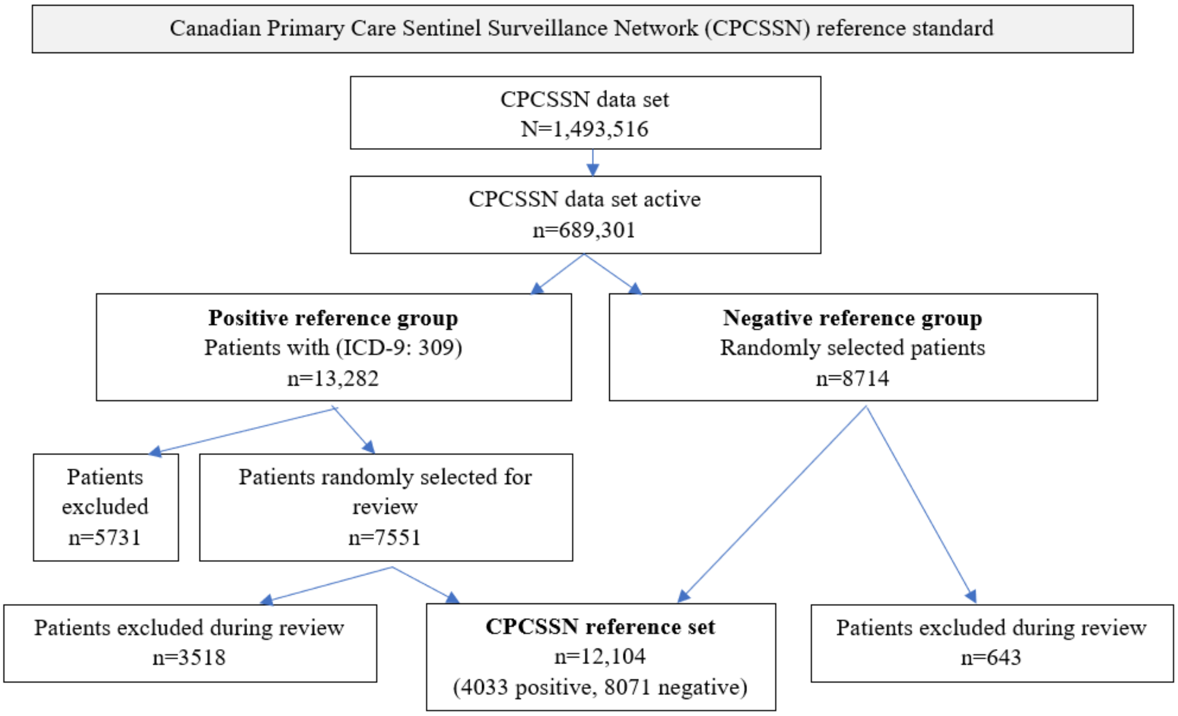
Figure 1. Flow diagram for creation of posttraumatic stress disorder reference standard in the Manitoba Primary Care Research Network.


Manitoba Primary Care Research Network (MaPCReN) reference standards
MaPCReN patients N=289,523
MaPCReN inactive patients excluded n=135,405
MaPCReN active patients n=154,118
Positive reference group Patients with (ICD-9: 300,308,309,311) n=21,713
Negative reference group Randomly selected patients n=2025
Patients not reviewed n=6213
Patients with encounter note n=373 (258 excluded following review)
Patients with semistructured text only n=15,127 (13,561 excluded following review)
Patients with encounter notes n=264 (49 excluded following review)
Patients with semistructured text only n=1761 (115 excluded following review)
MaPCReN reference set 1 (encounter notes) n=330 (115 positive, 215 negative)
MaPCReN reference set 2 (semistructured text) n=3212 (1566 positive 1646 negative)
MaPCReN patients in both reference set 1 and 2 n=327

Pan-Canadian Primary Care Patients

From the CPCSSN repository, a subset of patient records was extracted for medical record review to create a pan-Canadian reference set for PTSD. The CPCSSN repository contains EMR data for 1,493,516 patients, of which 689,301 (46.15%) were considered active because they had an appointment within the previous 2 years [25]. Within CPCSSN, there is no free-text encounter note data available. Medical record review was performed by 12 medical students using short diagnostic text fields. In total, there were 6 cohorts of ~2700 randomly selected records, each reviewed by 2 medical students for a total of 16,265 records reviewed. We included patients from each of the 7 participating provinces. There were 13,282 patients with an ICD-9-CM code (309, adjustment reaction), of which 7551 (56.85%) were randomly selected for medical record review. Moreover, there were 8714 patients randomly selected for creation of the negative reference set. We used the same data extract table and process as conducted for the MaPCReN reference set. Discrepancies were reviewed by a family physician (AS). There were 3518/7551 (46.6%) who were excluded due to poor interrater agreement or being classified as “possible PTSD.” Our final reference set had 12,104 patients (n=4033, 33.32% positive and n=8071, 66.68% negative; Figure 2).

Figure 2. Flow diagram for creation of posttraumatic stress disorder reference standard in the Canadian Primary Care Sentinel Surveillance Network.


Canadian Primary Care Sentinel Surveillance Network (CPCSSN) reference standard
CPCSSN data set N=1,493,516
CPCSSN data set active n=689,301
Positive reference group Patients with (ICD-9: 309) n=13,282
Negative reference group Randomly selected patients n=8714
Patients excluded n=5731
Patients randomly selected for review n=7551
Patients excluded during review n=3518
CPCSSN reference set n=12,104 (4033 positive, 8071 negative)
Patients excluded during review n=643

Case Definitions

Four case definitions for PTSD were developed by consensus discussion and evidence review by a research team including clinicians and researchers. Case definitions included ICD-9-CM and Anatomical Therapeutic Chemical codes from the health condition, billing, encounter diagnosis, and medication tables of CPCSSN (Table 1). The ICD-9-CM code for PTSD is 309.81; however, some providers use a less specific ICD-9-CM code 309 (adjustment reaction) because of billing rules in some justifications (ie, Ontario) which require that only the first 3 digits of the ICD-9-CM code be entered. Additionally, during medical record review, medical students found that patients with a diagnostic text entry for “PTSD” also had the following ICD-9-CM codes associated with that encounter: 300 (anxiety), 308 (acute reaction to stress), 309 (adjustment reaction), or 311 (depressive disorder). Medical student reviewers were instructed to create a list of spelling mistakes, abbreviations, and phrases that were recorded by primary care providers to identify PTSD in the short diagnostic text field (Multimedia Appendix 2). These codes and list were incorporated into data preprocessing stages prior to applying the case definitions (Table 1).

Table 1. Posttraumatic stress disorder (PTSD) test case definitions.

Case definition 1	Case definition 2	Case definition 3	Case definition 4
≥1 health condition, billing, or encounter diagnosis for ICD-9-CM[a] 309.81	≥1 health condition for ICD-9-CM 309.81 OR ≥2 billing, encounter diagnosis for ICD-9-CM 309.81 separated by at least 1 week	≥1 health condition for ICD-9-CM 309.81 OR ≥1 billing, encounter diagnosis for ICD-9-CM 309.81 AND PTSD medication (ATC[b] code starting with N05 or N06) OR ≥2 billing, encounter diagnosis for ICD-9-CM 309.81 separated by at least 1 week	≥1 health condition, billing, or encounter diagnosis for ICD-9-CM 309.81 OR ≥1 health condition, billing, or encounter diagnosis for ICD-9-CM starting with 290-316 AND PTSD recorded as the diagnosis name by the provider (Multimedia Appendix 2)

[a] ICD-9-CM: International Classification of Disease, ninth edition, clinical modification.

[b] ATC: anatomical therapeutic chemical.

Preprocessing Steps

Primary care EMR data are collected for clinical purposes and therefore often include domain-specific language and acronyms as well as spelling and typographical errors. To prepare the data for validation (ie, capture in case definition 4), we removed stop words, removed special characters, and adjusted capitalization in the short diagnostic text fields of the EMR. Short diagnostic text fields document diagnosis name and reasons for the encounter. During medical record review, medical student reviewers recorded PTSD acronyms and spelling errors that were later converted into “PTSD” prior to applying the case definition (Multimedia Appendix 2).

Statistical Analyses

We compared the agreement of EMR free-text encounter notes and EMR short diagnostic text fields using a 2x2 contingency table and the following metrics: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and overall accuracy. Further, we assessed agreement between the PTSD case definitions and each of the 3 reference sets (MaPCReN free text, MaPCReN short diagnostic text, and CPCSSN) with sensitivity, specificity, PPV, NPV, and overall accuracy. The equations for these metrics are presented below:

Using the PTSD case definitions, the prevalence and 95% confidence limits were computed using an exact binomial test to estimate prevalence of PTSD in a pan-Canadian data set. Statistical analyses were conducted using SAS V9.4 (SAS Institute).

Ethics Approval

Ethical approval for this study was obtained from the Health Research Ethics Board at the University of Manitoba, approval number HS21053(2017:257).

Results

Manitoba Primary Care Patients

There were 154,118 patients in MaPCReN who attended an appointment with a participating provider between January 1, 2017, and December 31, 2019. There were 330 patients in MaPCReN reference set 1 (free-text data), and 3212 patients in MaPCReN reference set 2 (short diagnostic text). There were 327 patients who were included in both reference sets. There was a strong agreement between free-text and short diagnostic text reference sets with an overall accuracy of 93.6% (CI 90.4%-96.0%). There were 20 patients who had ongoing symptoms of PTSD documented in free-text EMR data (not an explicit PTSD diagnosis) that were not identified through review of short diagnostic text fields. Despite this, there was strong agreement between the 2 reference sets with a sensitivity of 82.5% (CI 74.2%-88.9%) and specificity of 99.5% (CI 97.4%-100%; Table 2).

Case definitions 1 and 4 performed similarly in both MaPCReN reference sets (Table 3). Reference set 1 had a sensitivity of 82.6% (CI 74.4%-89.0%), specificity of 99.5% (CI 97.4%-100%), PPV of 99.0% (CI 93.1%-99.9%), NPV of 91.5% (87.8%-94.1%), and accuracy of 93.6% (CI 90.4%-96.0%) for both case definitions. Similarly, reference set 2 had a sensitivity of 100% (CI 99.8%-100%), specificity of 98.4% (CI 97.7%-99.0%), PPV of 98.4 (CI 97.6%-98.9%), NPV of 100%, and accuracy of 99.2% (CI 98.8%-99.5%) for both case definitions. Within the MaPCReN repository, supplementation with NLP (case definition 4) did not capture any additional patients when compared to case definition 1, which focused only on diagnostic codes for PTSD (ICD-9-CM 309.81). Requiring a second billing code for PTSD (case definition 2) or a medication that may be used to treat PTSD (case definition 3) produced lower sensitivity (57.4%, CI 47.8%-66.6% and 79.1%, CI 70.6%-86.2%; Table 3).

Table 2. Agreement between Manitoba Primary Care Research Network (MaPCReN) reference set 1 (with encounter notes) and MaPCReN reference set 2 (with short diagnostic text fields only; N=327).

Performance metric[a]	Value (95% CI)
Accuracy	93.6 (90.4-96.0)
Sensitivity	82.5 (74.2-88.9)
Specificity	99.5 (97.4-100)
Positive predictive value	99.0 (93.0-99.9)
Negative predictive value	91.4 (87.7-94.0)

[a] Cell occurrence <5 required suppression of numbers in 2x2 contingency table.

Table 3. Agreement between patients captured using the posttraumatic stress disorder case definitions and the Manitoba Primary Care Research Network (MaPCReN) reference sets.

Case definitions	[a]TP (n)	[b]TN (n)	[c]FN (n)	[d]FP (n)	[e]SE (%, CI)	[f]SP (%, CI)	[g]PPV (%, CI)	[h]NPV (%, CI)	Accuracy (%, CI)
MaPCReN reference set 1 (with encounter notes; N=330)									
Case definition 1	95	214	*Suppressed*	<5	82.6 (74.4-89.0)	99.5 (97.4-100)	99.0 (93.1-99.9)	91.5 (87.8-94.1)	93.6 (90.4-96.0)
Case definition 2	66	214	*Suppressed*	<5	57.4 (47.8-66.6)	99.5 (97.4-100)	98.5 (90.3-99.8)	81.4 (77.9-84.4)	84.9 (80.5-88.5)
Case definition 3	91	214	*Suppressed*	<5	79.1 (70.6-86.2)	99.5 (97.4-100)	98.9 (92.8-99.9)	89.9 (86.2-92.7)	92.4 (89.0-95.0)
Case definition 4	95	214	*Suppressed*	<5	82.6 (74.4-89.0)	99.5 (97.4-100)	99.0 (93.1-99.9)	91.5 (87.8-94.1)	93.6 (90.4-96.0)
MaPCReN reference set 2 (no encounter notes; N=3212)									
Case definition 1	1566	1620	0	26	100 (99.8-100)	98.4 (97.7-99.0)	98.4 (97.6-98.9)	100	99.2 (98.8-99.5)
Case definition 2	1135	1640	431	6	72.5 (70.2-74.7)	99.6 (99.2-99.9)	99..5 (98.8-99.8)	79.2 (77.8-80.5)	86.4 (85.2-87.6)
Case definition 3	1469	1620	97	26	93.8 (92.5-95.0)	98.4 (97.7-99.0)	98.3 (97.5-98.8)	94.4 (93.2-95.3)	96.2 (95.5-96.8)
Case definition 4	1566	1620	0	26	100 (99.8-100)	98.4 (97.7-99.0)	98.4 (97.6-98.9)	100	99.2 (98.8-99.5)

[a]TP: true positive.

[b]TN: true negative.

[c]FN: false negative.

[d]FP: false positive.

[e]SE: sensitivity.

[f]SP: specificity.

[g]PPV: positive predictive value.

[h]NPV: negative predictive value.

Pan-Canadian Primary Care Patients

In the CPCSSN data set, case definition 4 had the strongest agreement with our reference set with a sensitivity of 91.1% (CI 90.1%-91.9%), specificity of 99.1% (CI 98.9%-99.3%), PPV of 98.1% (CI 97.6%-98.5%), NPV of 95.7% (CI 95.3%-96.1%), and accuracy of 96.4% (CI 96.1%-96.8%). In comparison, case definition 1 had a sensitivity of 72.3% (CI 70.9%-73.7%), specificity of 99.1% (CI 98.9%-99.3%), PPV of 97.6% (CI 97.0%-98.1%), NPV of 87.8% (CI 87.2%-88.3%), and accuracy of 90.2% (CI 89.7%-90.7%). The inclusion of multiple billing codes (case definition 2) or medications that can be used to treat PTSD (case definition 3) did not improve the agreement of the case definitions (Table 4).

When we apply each of the definitions to the CPCSSN data set of active patients, PTSD prevalence estimates suggest a range of 0.8% (CI 0.77%-0.81%; n=5565) with case definition 2 to 1.3% (CI 1.25%-1.31%; n=8913) with case definition 4. Case definition 1, which required at least one specific ICD-9-CM code 309.81, had a prevalence of 1.1% (CI 1.08%-1.13%; n=7718).

Table 4. Agreement between the posttraumatic stress disorder case definitions in the Canadian Primary Care Sentinel Surveillance Network and the reference data set (N=12,104).

Case definitions	TP[a] (n)	TN[b] (n)	FN[c] (n)	FP[d] (n)	SE[e] (%, CI)	SP[f] (%, CI)	PPV[g] (%, CI)	NPV[h] (%, CI)	Accuracy (%, CI)
1	2917	8000	1116	71	72.3 (70.9-73.7)	99.1 (98.9-99.3)	97.6 (97.0-98.1)	87.8 (87.2-88.3)	90.2 (89.7-90.7)
2	2502	8045	1531	26	62.0 (60.5-63.5)	99.7 (99.5-99.8)	99.0 (98.5-99.3)	84.0 (83.5-84.5)	87.1 (86.5-87.7)
3	2917	8004	1116	67	72.3 (70.9-73.7)	99.2 (99.0-99.4)	97.8 (97.2-98.2)	87.8 (87.2-88.3)	90.2 (89.7-90.8)
4	3672	8000	361	71	91.1 (90.1-91.9)	99.1 (98.9-99.3)	98.1 (97.6-98.5)	95.7 (95.3-96.1)	96.4 (96.1-96.8)

[a] TP: true positive.

[b] TN: true negative.

[c] FN: false negative.

[d] FP: false positive.

[e] SE: sensitivity.

[f] SP: specificity.

[g] PPV: positive predictive value.

[h] NPV: negative predictive value.

Discussion

Principal Results

We found strong agreement between reference standards created through review of EMR free-text encounter notes compared to EMR short diagnostic text fields. Similar to other studies, we also found that when available, free-text encounter notes can capture additional information about a patient for identification of disease, symptoms, and management strategies [7,12,14,15]. Although free-text encounter notes provided additional information regarding risk factors and symptoms, when compared to short diagnostic text fields, their inclusion did not dramatically impact the validation of algorithms intended to identify diagnosed cases. Primary care settings in our sample include regionally or privately operated clinics, different EMR systems, and privacy and confidentiality regulations that can make free-text data difficult to obtain [27]. We found that when free-text encounter notes are unavailable, short diagnostic text data offer a viable option for identification of a confirmed diagnosis among primary care patients, even when this condition is complex such as PTSD.

Comparison With Prior Work

The estimated PTSD prevalence ranged from 0.8% to 1.3%. Case definition 1, which focused on specific ICD-9-CM code for PTSD (309.81) found a prevalence of 1.1% but may not be viable if 5-digit billing codes (ie, ICD-9-CM) are not available. Within the Manitoba data set, diagnostic code alone and diagnostic codes supplemented with NLP both had high agreement with reference sets. Inclusion of free-text encounter notes during medical record review did not significantly change agreement metrics. Contrary to similar studies, we did not find that the inclusion of NLP improved the agreement of our case definition in Manitoba [7,12,14,15]. However, when we applied the case definitions to the pan-Canadian CPCSSN reference set, provincial differences in diagnostic codes and EMR structure were noticed. Seungwon et al [27] conducted a scoping review of 274 articles representing 299 algorithms for Charlson conditions reporting that case validation studies frequently focused on a single-center, limiting generalizability of created algorithms. Similarly, we found that our algorithm tested in MaPCReN, which includes only 3 distinct EMR venders, performed better than when tested in a pan-Canadian CPCSSN data set representing 11 different EMR venders across Canada.

Consistent with other literature regarding complex phenotypes, we found that reliance on diagnostic codes can vary in accuracy depending on the jurisdiction [14,27]. System-level and jurisdictional differences in diagnostic coding requirements reduced the sensitivity of case definition 1 in the CPCSSN reference set. Depending on the condition, a 3-digit ICD-9-CM code may still indicate disease presence. For example, ICD-9-CM 250 indicates diabetes with ICD-9-CM subcodes indicating the type and severity of the diabetes [28]. However, the 3-digit ICD-9-CM code for PTSD is 309, indicating an adjustment reaction which is not specific to PTSD. When using free-text data to improve PTSD capture, tools such as well-developed and defined NLP or lasso regression can aid in the identification of patients [7,12,14,15]. Case definition 4 supplemented specific diagnostic codes with NLP of short diagnostic text fields in the EMR to identify patients with PTSD. Similar to other works, we found that combining structured EMR data and unstructured free text significantly improved diagnostic capture in our pan-Canadian data set yielding higher performance [7,15,20,27]. However, we did not ascertain additional benefit from using free-text encounter notes when compared to short diagnostic text fields that are more widely available. Doan et al [12] found that NLP showed comparable performance in disease identification to clinician manual chart review. Although literature suggests the need to capture multiple risk factors for the identification of PTSD [19], in this study, we focused NLP on explicit PTSD diagnostic text documented in short diagnostic text fields of the EMR. We demonstrated that explicit PTSD diagnostic text can improve PTSD capture in a pan-Canadian data set. NLP can serve as a model for decision support closing documentation gaps and overcoming barriers present when only structured data fields are available [12,15].

Following free-text encounter note review, 6.1% (20/327) of patients in our purposefully selected reference standard were identified as having “possible PTSD.” These patients did not have an explicit PTSD diagnosis in the text or structured data fields of the EMR. Characterizing patients with “possible PTSD” may identify patients who warrant further clinical investigation to inform diagnosis. Identification of patients with “possible PTSD” can support patient care by informing diagnostic investigations, as well as promoting documentation of mental health symptoms, treatments, and improvements in symptoms [15]. This may be a role for clinical decision support systems that can provide passive alerts to primary care providers indicating the need for further PTSD assessment [7,15].

Depending on study objectives and data set, researchers may choose to use different combinations of coded and free-text data, the former being more readily available and commonly used in many jurisdictions [14,27]. However, previous studies have demonstrated that using diagnostic codes from one part of the EMR alone may be problematic due to data quality concerns [18,29]. Furthermore, changes in terminology and coding standards can make it difficult to compare and share algorithms between EMR systems and jurisdictions. Understanding the health system structure and setting of the study is crucial in algorithm development [27]. Interpretability is an important consideration within the clinical domain, which may suggest the use of an NLP rule-based system, particularly when a data set has limited free-text information. Despite this, the supplementation of structured EMR data with NLP-derived data is important to overcome documentation gaps [9,15,20]. Our pan-Canadian data set only included short diagnostic fields and did not include free-text encounter notes. The availability of free-text encounter notes may suggest the use of a pretrained model for both text representation and classification. Pretrained model such as the Bidirectional Encoder Representations from Transformers can transform free-text data into a standardized form [9]. Specialist models such as MentalBERT have developed domain-specific pretrained language models in the area of mental health that can further benefit machine learning models aimed at capturing mental health conditions [30]. Matching data sets to appropriate methods can balance interpretability of the model and improve prediction leading to results that can inform clinical decision-making and health system planning [9,15,19,20].

Limitations

This study relied on primary care provider documentation in the EMR. NLP assessment of clinical notes entered by a primary care provider requires processing of clinical narratives that were entered by providers with limited time and may therefore include domain-specific abbreviations and spelling or editorial errors [7]. Due to variation in primary care provider documentation and coding, our study may have underestimated the presence of PTSD in its patient population. Additionally, clinicians primarily use their EMR for clinical purposes and therefore are less concerned with the secondary use of specific ICD-9-CM codes. This may contribute to issues with data capture or completeness. The use of NLP must be developed within context to meet organizational challenges of structured data fields [14]. Tools developed through this study can support identification in a Canadian EMR data repository but have not been validated in other jurisdictions. CPCSSN represents care received from a primary care provider and therefore does not represent care received from a specialist, such as a psychiatrist or psychologist. Future studies linking this data set to other data holdings representing care providing by specialist providers may improve our case definition accuracy by including more dedicated assessments and information related to PTSD care.

Conclusions

Inclusion of free-text encounter notes during medical record review did not lead to dramatically improved capture of PTSD cases, nor did it lead to significant improvements in case definition agreement. However, incorporating NLP of short diagnostic text fields into a case definition for a complex condition, such as PTSD, improved the capture of our case definition when compared to case definitions that used structured data fields alone. Depending on the jurisdiction and EMR systems in use, specific diagnostic codes can still provide a good estimate of patients with PTSD in a population.

Further research is required to refine NLP algorithms to be able to detect PTSD from free-text encounter notes lacking a formal coded diagnosis entry. In this large primary care data set, PTSD affected between 0.8% and 1.3% of the population, demonstrating that primary care EMR data are a rich source of data for this complex condition.

Acknowledgments

We would like to thank all of the following medical record reviewers: Alyzia Horsefall, Amy Lam, Anna Liu, Courtney Bell, Daniel Shenoda, Dhasni Muthurmuni, Gabriel Furman, Hannah Stirton, John Bahng Seokjae, Joseph Asamineq, Katrina Leong, Mary Pambid, Peter Trokajilo, Shivani Mathur, Suman Lakhi, and Sydnee Tuckett. We also acknowledge the assistance of William Peeler in data acquisition.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data extraction form.
[PDF File (Adobe PDF File), 65 KB - medinform_v10i12e41312_app1.pdf]

Multimedia Appendix 2
Posttraumatic stress disorder terms.
[PDF File (Adobe PDF File), 60 KB - medinform_v10i12e41312_app2.pdf]

References

1. Cavlan O, Dash P, Drouin J, Fountaine T, Riahi F. Using care pathways to improve health systems. Health International. 2011. URL: https://www.mckinsey.com/client_service/healthcare_systems_and_services/people/~/media/3EAC5D0AD0D440BD9500799BAD632DE0.ashx [accessed 2022-11-08]
2. Rubin G, Berendsen A, Crawford SM, Dommett R, Earle C, Emery J, et al. The expanding role of primary care in cancer control. Lancet Oncol 2015 Sep;16(12):1231-1272. [doi: 10.1016/S1470-2045(15)00205-3] [Medline: 26431866]
3. Queenan JA, Williamson T, Khan S, Drummond N, Garies S, Morkem R, et al. Representativeness of patients and providers in the Canadian Primary Care Sentinel Surveillance Network: a cross-sectional study. CMAJ Open 2016;4(1):E28-E32 [FREE Full text] [doi: 10.9778/cmajo.20140128] [Medline: 27331051]
4. Singer AG, Kosowan L, Nankissoor N, Phung R, Protudjer JLP, Abrams EM. Use of electronic medical records to describe the prevalence of allergic diseases in Canada. Allergy Asthma Clin Immunol 2021 Aug 18;17(1):85-90 [FREE Full text] [doi: 10.1186/s13223-021-00580-z] [Medline: 34407859]
5. Singer A, Kosowan L, Loewen S, Spitoff S, Greiver M, Lynch J. Who is asked about alcohol consumption? A retrospective cohort study using a national repository of Electronic Medical Records. Prev Med Rep 2021 Jun;22:101346-101352 [FREE Full text] [doi: 10.1016/j.pmedr.2021.101346] [Medline: 33767948]
6. Greiver M, Aliarzadeh B, Meaney C, Moineddin R, Southgate CA, Barber DT, et al. Are We Asking Patients if They Smoke?: Missing Information on Tobacco Use in Canadian Electronic Medical Records. Am J Prev Med 2015 Aug;49(2):264-268. [doi: 10.1016/j.amepre.2015.01.005] [Medline: 25997907]
7. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform 2009 Oct;42(5):760-772 [FREE Full text] [doi: 10.1016/j.jbi.2009.08.007] [Medline: 19683066]
8. Wright A, Sittig DF, McGowan J, Ash JS, Weed LL. Bringing science to medicine: an interview with Larry Weed, inventor of the problem-oriented medical record. J Am Med Inform Assoc 2014;21(6):964-968 [FREE Full text] [doi: 10.1136/amiajnl-2014-002776] [Medline: 24872343]
9. Elbattah M, Arnaud É, Gignon M, Dequen G. The Role of Text Analytics in Healthcare: A Review of Recent Developments and Applications. 2021 Presented at: Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies; February 11-13, 2021; Vienna, Austria p. 825-832. [doi: 10.5220/0010414508250832]
10. 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]
11. Coleman N, Halas G, Peeler W, Casaclang 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;16:11-19 [FREE Full text] [doi: 10.1186/s12875-015-0223-z] [Medline: 25649201]
12. Doan S, Maehara CK, Chaparro JD, Lu S, Liu R, Graham A, Pediatric Emergency Medicine Kawasaki Disease Research Group. Building a Natural Language Processing Tool to Identify Patients With High Clinical Suspicion for Kawasaki Disease from Emergency Department Notes. Acad Emerg Med 2016 May;23(5):628-636 [FREE Full text] [doi: 10.1111/acem.12925] [Medline: 26826020]
13. Gigengack MR, van Meijel EPM, Alisic E, Lindauer RJL. Comparing three diagnostic algorithms of posttraumatic stress in young children exposed to accidental trauma: an exploratory study. Child Adolesc Psychiatry Ment Health 2015;9:14-22 [FREE Full text] [doi: 10.1186/s13034-015-0046-7] [Medline: 25984233]
14. Harrington KM, Quaden R, Stein MB, Honerlaw JP, Cissell S, Pietrzak RH, VA Million Veteran Program and Cooperative Studies Program. Validation of an Electronic Medical Record-Based Algorithm for Identifying Posttraumatic Stress Disorder in U.S. Veterans. J Trauma Stress 2019 Apr 22;32(2):226-237 [FREE Full text] [doi: 10.1002/jts.22399] [Medline: 31009556]
15. Shiner B, Levis M, Dufort VM, Patterson OV, Watts BV, DuVall SL, et al. Improvements to PTSD quality metrics with natural language processing. J Eval Clin Pract 2022 Aug 24;28(4):520-530. [doi: 10.1111/jep.13587] [Medline: 34028937]
16. Hao T, Huang Z, Liang L, Weng H, Tang B. Health Natural Language Processing: Methodology Development and Applications. JMIR Med Inform 2021 Oct 21;9(10):e23898 [FREE Full text] [doi: 10.2196/23898] [Medline: 34673533]
17. Spasic I, Nenadic G. Clinical Text Data in Machine Learning: Systematic Review. JMIR Med Inform 2020 Mar 31;8(3):e17984 [FREE Full text] [doi: 10.2196/17984] [Medline: 32229465]
18. Singer A, Yakubovich S, Kroeker AL, Dufault B, Duarte R, Katz A. Data quality of electronic medical records in Manitoba: do problem lists accurately reflect chronic disease billing diagnoses? J Am Med Inform Assoc 2016 Nov;23(6):1107-1112. [doi: 10.1093/jamia/ocw013] [Medline: 27107454]
19. Karstoft K, Galatzer-Levy IR, Statnikov A, Li Z, Shalev AY, members of Jerusalem Trauma Outreach Prevention Study (J-TOPS) group. Bridging a translational gap: using machine learning to improve the prediction of PTSD. BMC Psychiatry 2015 Mar 16;15(1):30-37 [FREE Full text] [doi: 10.1186/s12888-015-0399-8] [Medline: 25886446]

20. Zafari H, Kosowan L, Zulkernine F, Signer A. Diagnosing post-traumatic stress disorder using electronic medical record data. Health Informatics J 2021 Nov 24;27(4):14604582211053259 [FREE Full text] [doi: 10.1177/14604582211053259] [Medline: 34818936]
21. Koenen KC, Ratanatharathorn A, Ng L, McLaughlin KA, Bromet EJ, Stein DJ, et al. Posttraumatic stress disorder in the World Mental Health Surveys. Psychol. Med 2017 Apr 07;47(13):2260-2274. [doi: 10.1017/s0033291717000708]
22. Van Ameringen M, Mancini C, Patterson B, Boyle MH. Post-traumatic stress disorder in Canada. CNS Neurosci Ther 2008;14(3):171-181 [FREE Full text] [doi: 10.1111/j.1755-5949.2008.00049.x] [Medline: 18801110]
23. Sareen J, Cox BJ, Stein MB, Afifi TO, Fleet C, Asmundson GJG. Physical and mental comorbidity, disability, and suicidal behavior associated with posttraumatic stress disorder in a large community sample. Psychosom Med 2007 Apr;69(3):242-248. [doi: 10.1097/PSY.0b013e31803146d8] [Medline: 17401056]
24. Stein MB, McQuaid JR, Pedrelli P, Lenox R, McCahill ME. Posttraumatic stress disorder in the primary care medical setting. Gen Hosp Psychiatry 2000;22(4):261-269. [doi: 10.1016/s0163-8343(00)00080-3] [Medline: 10936633]
25. Menec V, Black C, Roos N, Bogdanovic B, Reid R. Defining practice populations for primary care: methods and issues. Manitoba Centre for Health Policy and Evaluation. 2000 Feb. URL: http://mchp-appserv.cpe.umanitoba.ca/reference/roster.pdf [accessed 2022-11-08]
26. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fifth edition (DSM-5). Washington, DC, USA: American Psychiatric Association Publishing; 2013.
27. Lee S, Doktorchik C, Martin EA, D'Souza AG, Eastwood C, Shaheen AA, et al. Electronic Medical Record-Based Case Phenotyping for the Charlson Conditions: Scoping Review. JMIR Med Inform 2021 Feb 01;9(2):e23934 [FREE Full text] [doi: 10.2196/23934] [Medline: 33522976]
28. International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/icd/icd9cm.htm [accessed 2022-02-03]
29. Singer A, Kroeker AL, Yakubovich S, Duarte R, Dufault B, Katz A. Data quality in electronic medical records in Manitoba: Do problem lists reflect chronic disease as defined by prescriptions? Can Fam Physician 2017 May;63(5):382-389 [FREE Full text] [Medline: 28500199]
30. Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E. MentalBERT: Publicly available pretrained language models for mental healthcare. 2022 Presented at: Proceedings of the Language Resources and Evaluation Conference; June 21-23, 2022; Marseille, Francce p. 7184-7190. [doi: 10.48550/arXiv.2110.15621]

Abbreviations

CPCSSN: Canadian Primary Care Sentinel Surveillance Network
EMR: electronic medical record
ICD-9-CM: International Classification of Disease, ninth edition, clinical modification
MaPCReN: Manitoba Primary Care Research Network
NLP: natural language processing
NPV: negative predictive value
PPV: positive predictive value
PTSD: posttraumatic stress disorder

Edited by C Lovis, A Benis; submitted 21.07.22; peer-reviewed by M Elbattah, J Candeliere; comments to author 18.09.22; revised version received 09.11.22; accepted 13.11.22; published 13.12.22.

Please cite as:
Kosowan L, Singer A, Zulkernine F, Zafari H, Nesca M, Muthumuni D
Pan-Canadian Electronic Medical Record Diagnostic and Unstructured Text Data for Capturing PTSD: Retrospective Observational Study
JMIR Med Inform 2022;10(12):e41312
URL: https://medinform.jmir.org/2022/12/e41312
doi:10.2196/41312
PMID:36512389

Original Paper

Natural Language Processing and Graph Theory: Making Sense of Imaging Records in a Novel Representation Frame

Laurent Binsfeld Gonçalves[1], MA; Ivan Nesic[1], MSc; Marko Obradovic[1], MSc; Bram Stieltjes[1], MD, PhD; Thomas Weikert[1], MD; Jens Bremerich[1], MD, PhD

[1] Clinic of Radiology & Nuclear Medicine, University Hospital Basel, University of Basel, Basel, Switzerland

Corresponding Author:
Laurent Binsfeld Gonçalves, MA
Clinic of Radiology & Nuclear Medicine
University Hospital Basel
University of Basel
Petersgraben, 4
Basel, 4031
Switzerland
Phone: 352 621517916
Email: laurent.binsfeld@gmail.com

Abstract

Background: A concise visualization framework of related reports would increase readability and improve patient management. To this end, temporal referrals to prior comparative exams are an essential connection to previous exams in written reports. Due to unstructured narrative texts' variable structure and content, their extraction is hampered by poor computer readability. Natural language processing (NLP) permits the extraction of structured information from unstructured texts automatically and can serve as an essential input for such a novel visualization framework.

Objective: This study proposes and evaluates an NLP-based algorithm capable of extracting the temporal referrals in written radiology reports, applies it to all the radiology reports generated for 10 years, introduces a graphical representation of imaging reports, and investigates its benefits for clinical and research purposes.

Methods: In this single-center, university hospital, retrospective study, we developed a convolutional neural network capable of extracting the date of referrals from imaging reports. The model's performance was assessed by calculating precision, recall, and F1-score using an independent test set of 149 reports. Next, the algorithm was applied to our department's radiology reports generated from 2011 to 2021. Finally, the reports and their metadata were represented in a modulable graph.

Results: For extracting the date of referrals, the named-entity recognition (NER) model had a high precision of 0.93, a recall of 0.95, and an F1-score of 0.94. A total of 1,684,635 reports were included in the analysis. Temporal reference was mentioned in 53.3% (656,852/1,684,635), explicitly stated as not available in 21.0% (258,386/1,684,635), and omitted in 25.7% (317,059/1,684,635) of the reports. Imaging records can be visualized in a directed and modulable graph, in which the referring links represent the connecting arrows.

Conclusions: Automatically extracting the date of referrals from unstructured radiology reports using deep learning NLP algorithms is feasible. Graphs refined the selection of distinct pathology pathways, facilitated the revelation of missing comparisons, and enabled the query of specific referring exam sequences. Further work is needed to evaluate its benefits in clinics, research, and resource planning.

(JMIR Med Inform 2022;10(12):e40534) doi:10.2196/40534

KEYWORDS

radiology; deep learning; NLP; radiology reports; imaging record; temporal referrals; date extraction; graph theory; health care information system; resource planning.

Introduction

Radiology departments generate tremendous amounts of reports every day. Narrative radiology reports are the primary communication medium between radiologists and referring physicians, thus playing a central role in patient care and containing a large variety of health care information [1,2]. From 1996 to 2010, image study volume for computed tomography (CT) and magnetic resonance imaging (MRI) increased by 280% to 380% [3]. Radiology embraced digital workflows and electronic information transfer to referring colleagues early on, which virtually eradicated analog data in this field [4]. This early commitment provides enormous quantities of digitalized reporting data containing interpretative image descriptions. However, the extraction of this information is hampered because unstructured reports are poorly computer-readable [5]. Semantic reports contain valuable information at a granular level (eg, multiple temporal referrals) that can be evoked for the overall report or specific findings in multiple document locations. This multilocular information cannot easily be determined on a whole document level [6].

Natural language processing (NLP) is one solution to the problem of extracting specific information from the plethora of free-text radiology reports. NLP is defined as the analysis of linguistic data, most commonly in the form of textual data, using computational methods [7-13]. NLP has evolved from rule-based to machine learning algorithms [14-20], deep learning being a subset of the latter that applies multilayer neural networks [21,22]. Its capability to automatically extract structured information has been described in many medical research settings [23-29]. Especially in radiology, there are numerous instances where it has demonstrated excellent text mining performances, including the detection of incidental findings and recommendations [30-32], actionable findings [33], specific findings [34-41], quality assessment of reports [42,43], and the generation of curated data sets [44-49].

The quantitative accumulation of radiology reports per patient over the years has led to a highly interconnected network of exams. Modern picture archiving and communication systems (PACS) represent the different exams as a list sorted by their acquisition date. Most systems can highlight the previous exams of roughly the same region in the study description to the user. This type of comparative visualization does not consider multiregional studies or often-encountered findings at the margins of the acquired field of view. It does not foreground the dates to which the radiologist compared his findings in the report. This last part especially is a significant shortcoming for clinicians reviewing patient history. They have to read every report carefully to see to which point in time the radiologist compared tumor progress, for example, or if the images from an external institute were available to the radiologist at the exact time when reading the follow-up exam.

One crucial connection in this context is dated referrals to prior exams. The good practice guidelines for radiological reporting from the European Society of Radiology [50] and the 2020 revised American College of Radiology practice parameter for communication of diagnostic imaging findings emphasize the need for comparison with previous investigations, including the date of previous reports and mentioning the absence of previous imaging. By using comparison studies, radiologists make more observations, gain confidence in their interpretation, and provide more diagnoses [51-55]. One study found that the diagnostic accuracy, sensitivity, and specificity in mammography increased as the false-positive rate decreased [56]. Various recent studies relied on NLP techniques to extract the temporality of measurements in imaging reports (ie, attributing an observation on the current or prior exam) [39-62]. However, to the best of our knowledge, no methods that extract every referring date from semantic radiological texts have been researched. Moreover, no studies in the literature have focused on the overall temporal indexing of the report assessed, in most instances, by the radiologist at the beginning of the report.

One solution to displaying connections between a multitude of different reports is graph representation. Graph theory defines graphs as a set of properties stored in nodes connected by edges, which represent a relationship between the connected nodes [63,64]. A review paper from 2020 found that graphs, as defined by graph theory, are hardly used to represent patient data in a clinical context; in the literature review, only 11 papers matched the description [65].

This study aimed to develop a novel and concise visualization framework of related reports.

To this end, we applied a self-designed NLP algorithm capable of extracting the referencing dates from unstructured radiology reports on all the reports generated for 10 years at a university hospital. This information was an essential input for a relational graph in which the nodes represent the radiology reports with their associated metadata and the dated referrals are their connecting edges. Finally, we investigated the potential benefits of such a graph representation and storage for clinical and research purposes.

Methods

Ethics Approval

Institutional review board approval and the requirement for informed consent were waived (institutional review board: Ethikkommission Nordwest- und Zentralschweiz) since no patient identifiers were used. Collected data consisted of plain text from radiology reports and randomized metadata, neither of which could be tracked back to radiologists, individual patients, nor referring clinicians.

Data Set Acquisition and Description

We extracted all radiology reports from January 2011 to December 2021 as well as a selection of their associated Digital Imaging and Communications in Medicine (DICOM) metadata (ie, randomized patient ID, modality type, body region, study date) from the hospital database. All reports were written in German and derived from all the imaging modalities (ie, ultrasound, radiography, mammography, x-ray angiography, CT, MRI, nuclear medicine exams, and positron emission tomography [PET]-CT). The reports were a mix of unstructured free-text reports and standardized templates, either containing subheadings for distinct organs with prewritten normal findings (eg, CT chest-abdomen) or checklists for standardized reporting features (eg, Liver Imaging Reporting and Data System for liver MRI). The broad structure of the reports was usually divided into 5 sections: medical history, medical question, examination protocol, radiological finding, and impression.

Every radiology exam had a predefined body region and modality type in its DICOM metadata. There were 14 body regions and 9 modalities (see Multimedia Appendix 1).

Construction of a Temporal Reference Extraction Algorithm

Data Selection for Training

We randomly selected 5187 reports from the previously extracted radiology reports.

Data Annotation

An internally developed data annotation tool, “xtag,” was used. A second-year medical resident (LBG) manually labeled 5187 reports with 5 classes indicating the temporal reference (Table 1). The annotation classes “date,” “today,” “yesterday,” and “no previous” were applied on the text sequence level (ie, annotating sequences of numbers or words). The annotation class “missing” was applied at the document level and was exclusive, meaning that no other annotation could be applied. On the other hand, “date,” “today,” “yesterday,” and “no previous” could be applied multiple times per report. To assess the necessity of a second reading, a fifth-year medical resident in radiology (TW) annotated 100 randomly selected reports. This process yielded 100% agreement among readers. Considering the simplicity of the task and based on this result, we refrained from a second reading of the whole data set.

Table 1. Annotated classes and their defined meaning.

Class	Meaning
Date	Precise numerical date referring to a comparative exam; any numerical or partially numerical format was accepted.
Today	Non-numerical temporal reference to a comparative study done on the same day as the actual report (ie, any literal expression meaning today)
Yesterday	Non-numerical temporal reference to a comparative study done on the day before the actual report (ie, any literal expression meaning yesterday)
No previous	Explicit statement that no comparable previous exams are available
Missing	No mention of a comparative study

Data Format

The training pipelines required the annotations to conform to the IOB2 format [66,67]. The predictions were also produced in the same format (further technical information can be found in Multimedia Appendix 2 [5,68-72]).

Algorithm Training and Testing

We excluded 2392 reports from the annotated data set, as they did not contain temporal links. We split the data into a training/validation data set of 2646 reports (94.6%) and an independent test data set of 149 reports (5.4%). We estimated that 5% for an independent and second test data set is a valid representation, as we verified the algorithm's robustness using the 5-fold cross-validation [73]. We also considered the low output variability of the problem to be solved. We used the Spacy sentencizer to text into sentences before training. We then used the ktrain library to produce a bidirectional long short-term memory (LSTM) [74] model starting with pretrained fastText word embeddings [75] (for details, see Multimedia Appendix 2). We applied various rule-based date extraction algorithms on the predicted date sequences to extract as many dates as possible. The non-numerical classes today and yesterday were converted into a numerical format using the date of the referring report as a reference. The dates missing the year specification were assigned the same year as the referencing report. The prediction was ignored if the day or month was missing. A grid search algorithm tested different combinations of learning rates and batch sizes to find the near optimal parameters for our training algorithm. A 5-fold cross-validation [76] of the training data set with 20% of the reports as validation in each round was performed to evaluate the model's performance on large independent data sets. The data set split into folds was done at the report level. The model was tested on the independent test data set in a final evaluation step**.** The following performance evaluation metrics were used to assess the trained model's quality: precision, recall, and F_1-score [77].

Extraction of the Referenced Modality and Body Region

The referenced modality was extracted using a simple rule-based approach. After extracting the temporal references from the report, the algorithm searched for a mention of the modality in the sentence with the date reference. The previous report's body part was derived from its metadata and was assumed to be the same as the referencing report's body region.

Graduation of the Predicted Link's Confidence

We graduated the prediction's confidence as follows: (1) date, modality, and body part; (2) date and modality; (3) date and body part; (4) date. This confidence graduation was established as a link property, in which 1 was the most confident and 4 was the least confident. The link was discarded if it was impossible to generate it based on these 4 principles. This approach permitted narrowing and increasing the accuracy of the referenced reports if more than one exam was acquired on the referenced date.

Algorithm Application and Data Extraction on the Complete Data Set

The preparatory steps for extraction of the temporal information by the trained model were the same as for the training part. The result of the model's application to all the reports from 2011 to 2021 was a per-token table with labels for each token in the IOB2 format. Predictions that did not comply with the IOB2 format were removed.

Populating the Graph Database

The graph database system used was Neo4j (Version 4.4.). All the reports and a selection of their associated metadata from 2011 to 2021 were imported via the py2neo library. The metadata consisted of the exam's acquisition date, name, modality, and body region, as well as its randomized patient ID. The reports and their metadata were assigned to the vertices, and unidirectional edges from referencing reports to referenced reports were created. We assigned 3 properties to the edges: The first consisted of the inferred class called "reference class," the second showed the extracted string, and the third displayed the prediction's confidence.

Interactive Exploration of the Graph

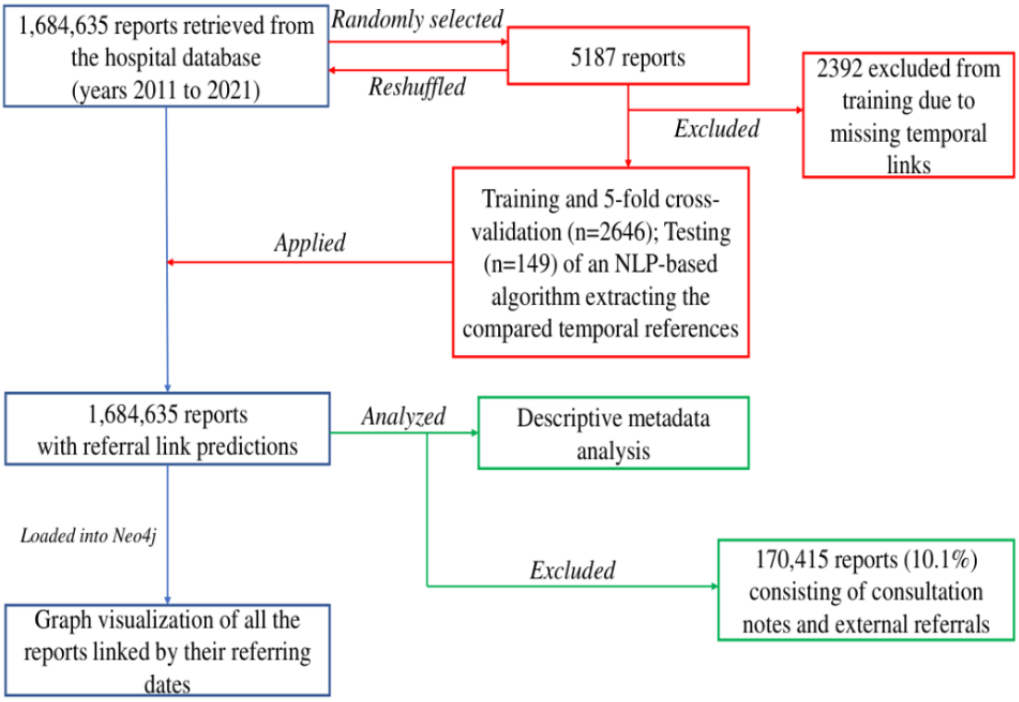
The assessment of the potential benefits of patient data visualization in a graph was explored interactively. The aim was to offer, at one glance, a well-ordered overview of the patient's imaging history with the related reports; enable comparison to previous exams; and represent the desired pathology pathway in a concise way (eg, oncological or postoperative follow-up imaging). In addition, it reveals to the clinician and the radiologist at what point in time the radiologist made his or her comparison. The user should be able to restrict his or her search to individually adaptable filters in the report's metadata (eg, body region, modality type, report date, or keywords in the reports' text). Another important feature would be to provide precisely filtered examinations in a concise order, in which every exam has its precisely defined position in a sequence. A final goal was to assess missed comparisons to previous exams, which was hoped to be achieved visually by spotting the missing link in the graph and by self-designable search algorithms.

Results

Data Set

In total, 1,684,635 reports from 264,655 distinct patients were extracted. We excluded 170,415 (10.1%) reports from the metadata analysis because they consisted of consultation notes and external referrals (detailed count in Multimedia Appendix 3). Figure 1 shows the detailed methodical flowchart.

Figure 1. Study flowchart of 1,684,635 patient reports retrieved from the hospital database (2011 to 2021). NLP: natural language processing.


1,684,635 reports retrieved from the hospital database (years 2011 to 2021)
Randomly selected
Reshuffled
5187 reports
Excluded
2392 excluded from training due to missing temporal links
Applied
Training and 5-fold cross-validation (n=2646); Testing (n=149) of an NLP-based algorithm extracting the compared temporal references
1,684,635 reports with referral link predictions
Analyzed
Descriptive metadata analysis
Loaded into Neo4j
Excluded
170,415 reports (10.1%) consisting of consultation notes and external referrals
Graph visualization of all the reports linked by their referring dates

Annotation Distribution

A total of 7860 annotations were applied to 5187 reports from 2011 to 2019. Class distribution of the training data set was as follows: 44% date reference, 27% no previous comparative exam, 23% missing temporal link and 6% referral class "today." We removed the semantic referencing class "yesterday" from our data set as there were not enough training samples (34/5187, 0.7%).

Temporal Information Extraction Algorithm

Hyperparameter Optimization

The algorithm's output yielded an optimal learning rate of 1e-2 and a batch size of 1024. The random state was fixed for reproducibility. The maximal number of training epochs was limited to a never reached limit of 30.

Training and Testing

The stagnation in the validation performance of 3 epochs was targeted for the early stopping. During training, the model is stored after each epoch. After the completion of the training process, the best-performing epoch weights were used for the final model. The same procedure was used for all the steps in which training was involved. After 5-fold cross-validation (results in Multimedia Appendix 4), the algorithm's performance was tested on the previously unused test data set (Table 2).

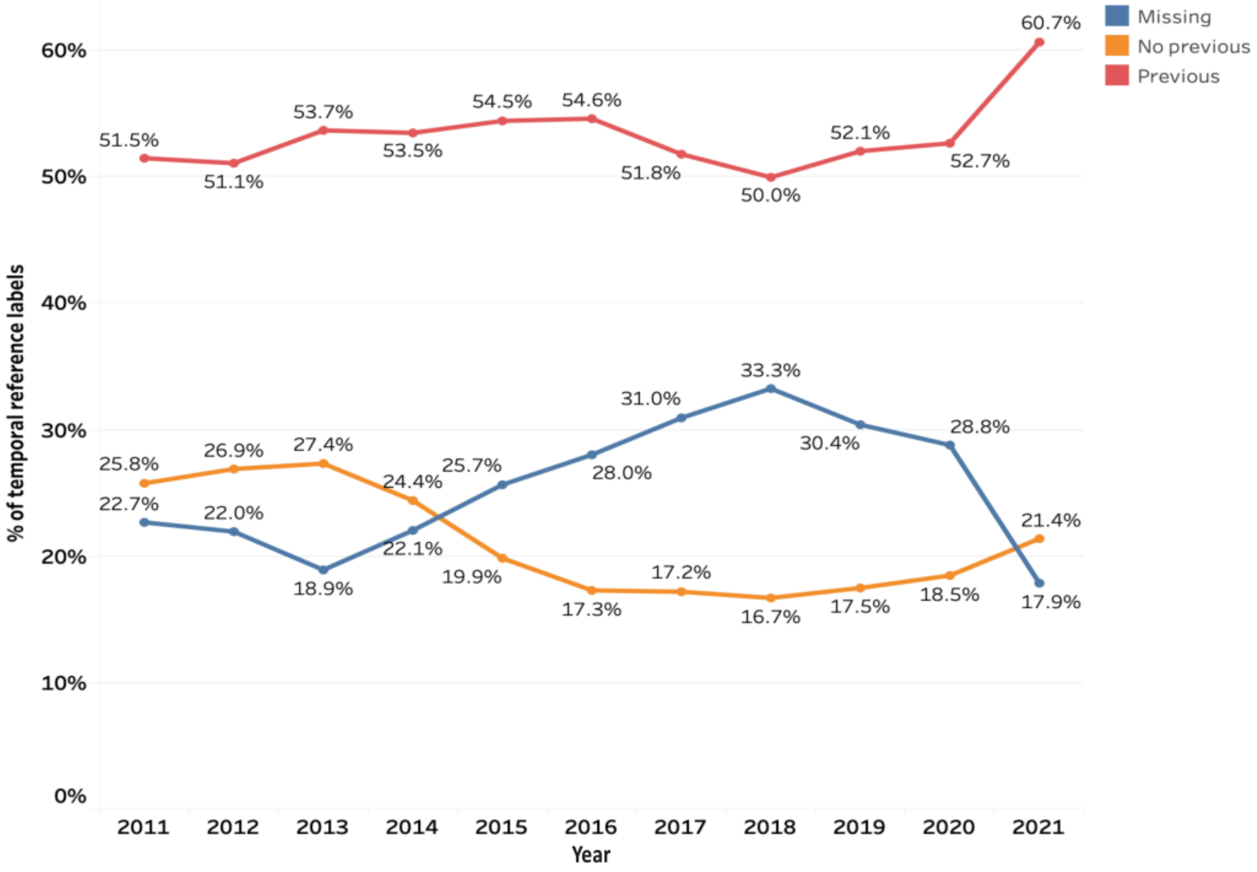
Table 2. Test results on 149 previously unused reports.

Variable	Precision (95% CI)	Recall (95% CI)	F_1-score (95% CI)
Date	0.93 (0.89-0.93)	0.9 (0.86-0.93)	0.93 (0.91-0.94)
No previous	0.94 (0.95-0.97)	0.98 (0.96-0.98)	0.96 (0.93-0.98)
Today	0.76 (0.73-0.88)	0.85 (0.79-0.90)	0.83 (0.79-0.93)
Micro average	0.93 (0.91-0.94)	0.92 (0.90-0.95)	0.94 (0.89-0.93)
Macro average	0.86 (0.84-0.95)	0.91 (0.87-0.95)	0.91 (0.80-0.94)
Weighted average	0.93 (0.91-0.94)	0.93 (0.90-0.94)	0.94 (0.91-0.95)

Temporal Referencing Analysis

A temporal reference to comparable exams was mentioned in 53.3% (656,852/1,232,297), explicitly stated as not available in 21.0% (258,386/1,232,297), and omitted in 25.7% (317,059/1,232,297) of the reports. Variability over the years was asserted (Figure 2). The modalities with the least amount of missing references were mammography (41,197/545,636, 7.6%), PET/CT (1850/18,500, 10.3%), and CT (278,286/2,399,017, 11.6%). On the other hand, angiography (33,924/40,872, 83.2%) and ultrasound (94,080/254,270, 37.2%) had the most missing references (Table S4 in Multimedia Appendix 5). The body regions with the lowest amount of missing references were trunk (3072/39,639, 7.8%), breast (5727/70,617, 8.1%), and thorax (25,646/276,060, 9.3%). On the other hand, the heart (19,030/26,090, 72.9%) and neck (14,716/23,230, 63.4%) regions had the most missing links (Table S5 in Multimedia Appendix 5). Modalities primarily referred to the same modality except for angiography referring to plain radiographs in 39.8% (1790/4503), PET/CT referring to MRI in 45.1% (456/1013), and nuclear medicine exams referring to CT in 33.9% (3500/10294; Table S6 in Multimedia Appendix 5). Every body region predominantly referred to the same body region. The most extreme example was “breast,” which was referenced in 99.0% (59,619/60,221) of the cases by the other breast studies.

Figure 2. Temporal reference of the reports (n=1,514,220) over the years.



Analysis of the Median Time Period of the Referencing Reports

The median period between referencing reports from 2011 to 2021 was determined in days, per modality (Table 3) and body region (Table 4). The most extended periods were found in mammography (372 days) and the corresponding body region breast (370 days). The shortest periods were observed in plain radiograph reports (19 days) and the thorax region (10 days).

Table 3. Median time period between referencing reports (n=757,249) per modality.

Modality	Time period (days), median (Q1-Q3)	IQR	*P* value
Computed radiography	19 (2-118)	116	<.001
X-ray angiography	35 (7-137)	130	.048
Computed tomography (CT)	42 (3-231)	228	<.001
Magnetic resonance	65 (3-344)	341	.002
Nuclear medicine	114 (8-440)	432	.06
PET[a]/CT	129.5 (30-366)	336	<.001
Ultrasound	344 (24-386)	362	.002
Mammography	372 (352-722)	368	.03

[a] PET: positron emission tomography.

Table 4. Median time period between referencing reports (n=757,249) per body region.

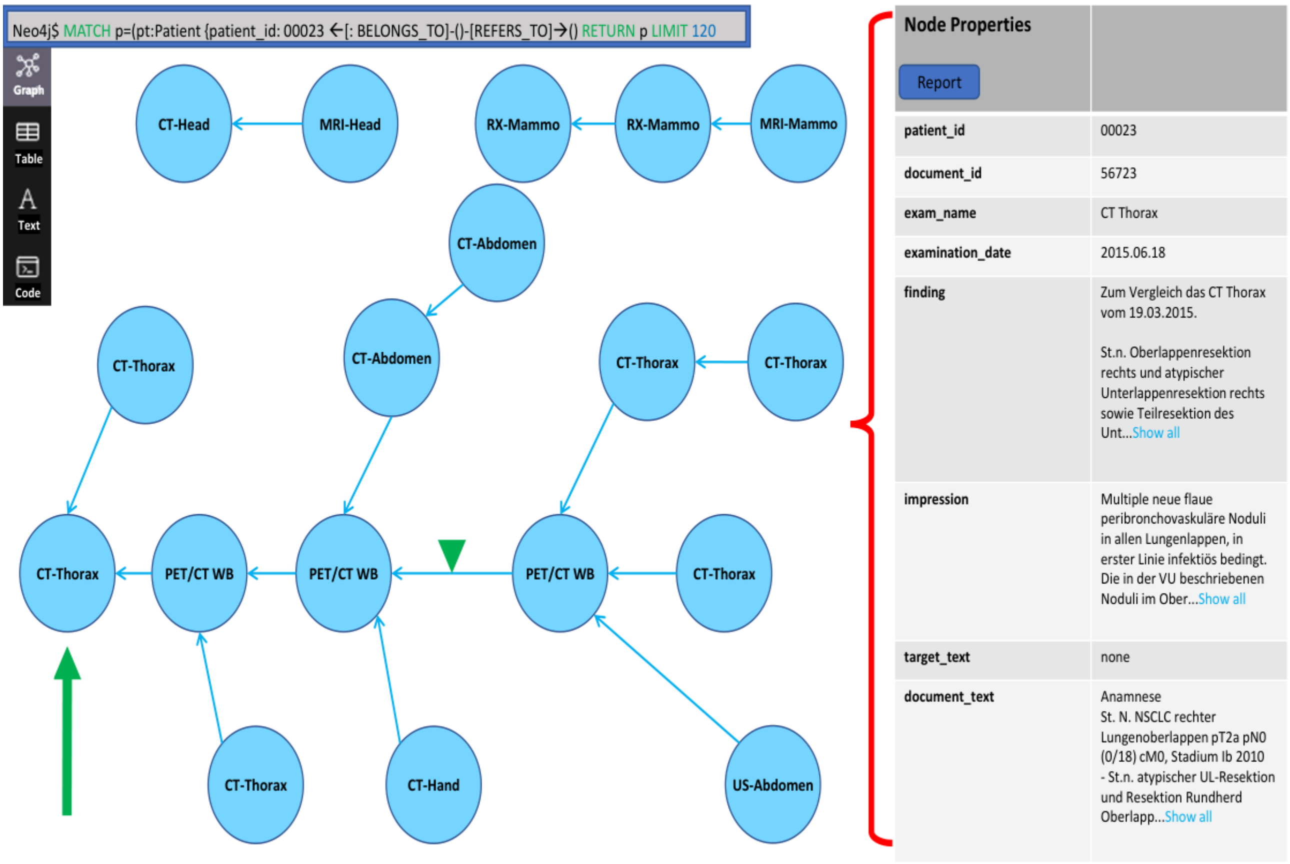
Body region	Time period (days), median (Q1-Q3)	IQR	*P* value
Thorax	10 (2-156)	154	.01
Upper extremity	11 (1-48)	47	.01
Abdomen	35 (4-237)	233	.006
Spine	35 (3-207)	204	<.001
Pelvis	39 (3-146)	143	.001
Lower extremity	42 (6-136)	130	.01
Head	65 (2-364)	362	<.001
Trunk	89 (34-196)	162	.03
Heart	125 (8-378)	370	=.40
Whole body	128 (8-427.3)	419.3	<.001
Neck	182 (29-395)	366	.045
Breast	370 (348-550)	202	.009

Exploration of Imaging Records in a Graph

General Overview

All the imaging reports and metadata from 2011 to 2021 were successfully loaded into a directed graph. The blue nodes represented the different patient reports labeled with their examination name (eg, CT-chest or MRI-head), and the connecting links were their automatically extracted referral dates. The interface was individually adaptable (eg, the user could freely position the nodes as desired, and the colors of the individual components and displayed metadata were customizable). The total number of distinct patient reports could be selected at the beginning of any query. This view permitted a rapid visual assessment of the earliest comparison exam (Figure 3).

Figure 3. Single patient case example (lung cancer with cancer-related studies) including the full user interface in which the blue box in the top left is the query interface; the blue nodes contain the examination name, represent all the imaging studies stored in the picture archiving and communication system (PACS), and are ordered from oldest on the left to newest on the right; the connecting blue arrows represent their referral links; and a node’s metadata (examination name, acquisition date, text of the selected and referenced reports, and finding and impression sections) appears on the right side when the node is clicked on. CT: computed tomography; MRI: magnetic resonance imaging; PET: positron emission tomography; RX: x-ray; US: ultrasonography; WB: whole body.



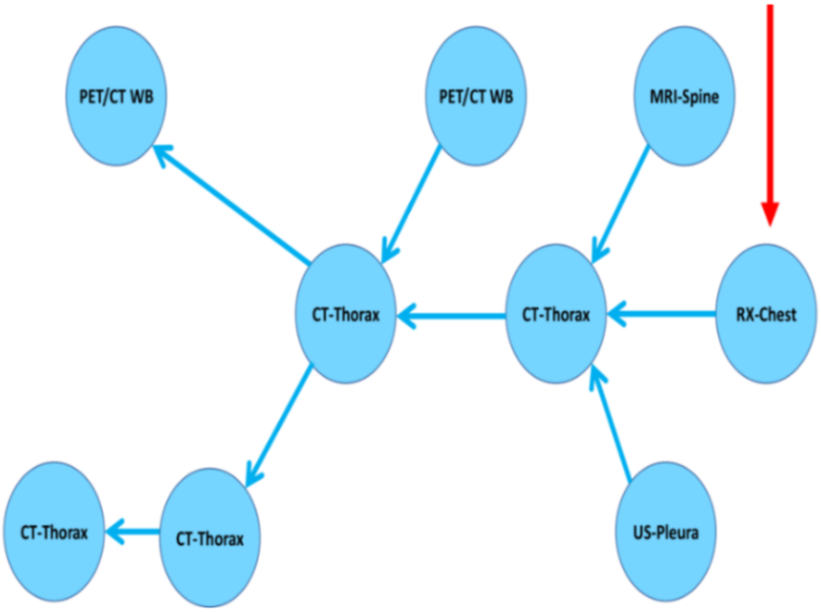
```
Neo4j$ MATCH p=(pt:Patient {patient_id: 00023 ←[: BELONGS_TO]-()-[REFERS_TO]→() RETURN p LIMIT 120
```

Node Properties	
patient_id	00023
document_id	56723
exam_name	CT Thorax
examination_date	2015.06.18
finding	Zum Vergleich das CT Thorax vom 19.03.2015. St.n. Oberlappenresektion rechts und atypischer Unterlappenresektion rechts sowie Teilresektion des Unt...Show all
impression	Multiple neue flaue peribronchovaskuläre Noduli in allen Lungenlappen, in erster Linie infektiös bedingt. Die in der VU beschriebenen Noduli im Ober...Show all
target_text	none
document_text	Anamnese St. N. NSCLC rechter Lungenoberlappen pT2a pN0 (0/18) cM0, Stadium Ib 2010 - St.n. atypischer UL-Resektion und Resektion Rundherd Oberlapp...Show all

Multiparametric Filtered Representations

Narrowing the reports down to the most relevant and thus facilitating visualization are of utmost relevance with the high number of exams per patient. By clicking on the node of interest, the user could opt to display solely the linked reports (visualized in Figure 4). Another possible method of restricting the view and looking for specific findings was a search filter related to the associated metadata and specific words in the report's text. One possible concept would be to look for specific exams with no previous reference and a defined pathologic condition as a keyword in the report's text, which would speed up the selection of the first exam associated with this condition.

Figure 4. Filtered representation of related studies as a partial screenshot within the user interface, in which all the related prior exams appear at one glance by clicking on the last node referring to a lung cancer (red arrow). Although the user selected the most recent study, clicking on every other node in this network would have resulted in the same view. CT: computed tomography; MRI: magnetic resonance imaging; PET: positron emission tomography; US: ultrasonography; WB: whole body.



Specific Exam Sequence Selection

Selecting highly customizable sequences of referring exams with specific metadata attributes (eg, chest x-ray followed by chest CT) was possible. This can be refined, for example, with a period restriction or restricted time interval between the related exams (Figures 5 and 6).

Figure 5. Specific exam sequence selection as a partial screenshot within the user interface, in which we used the query field to randomly select 300 reports (blue nodes) of head computed tomography (CT) referenced by a head magnetic resonance image (MRI) that was acquired no longer than 3 days later and contained the keyword “infarct” in the impression field.

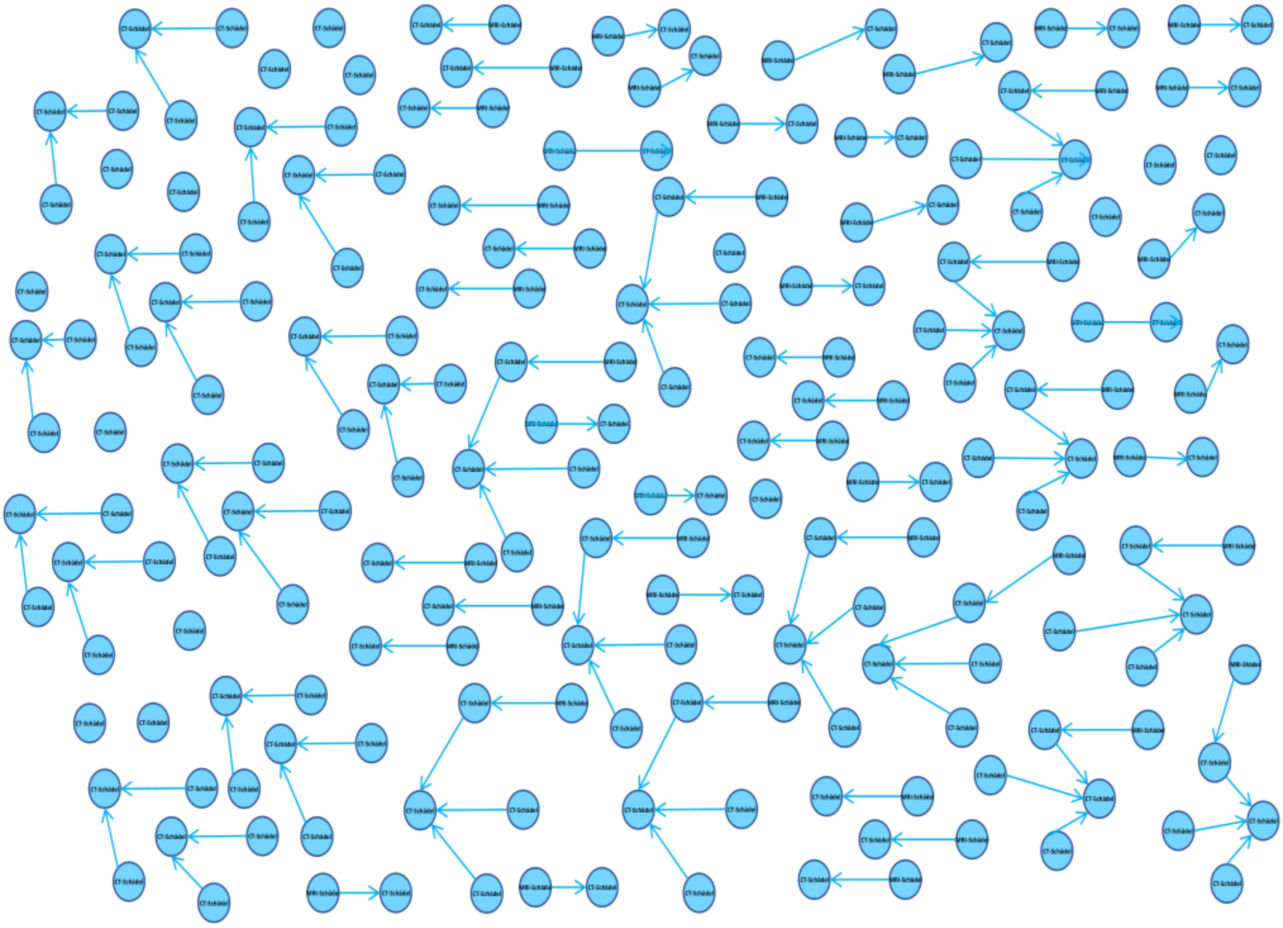
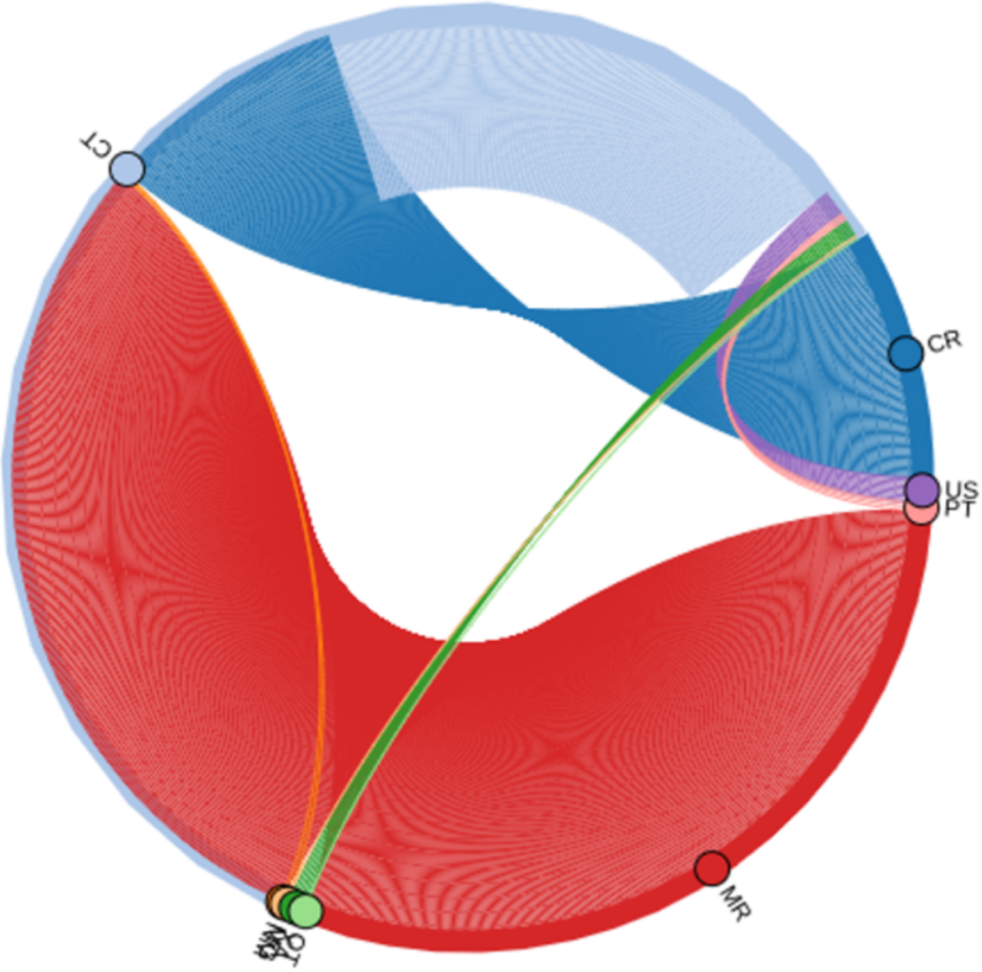


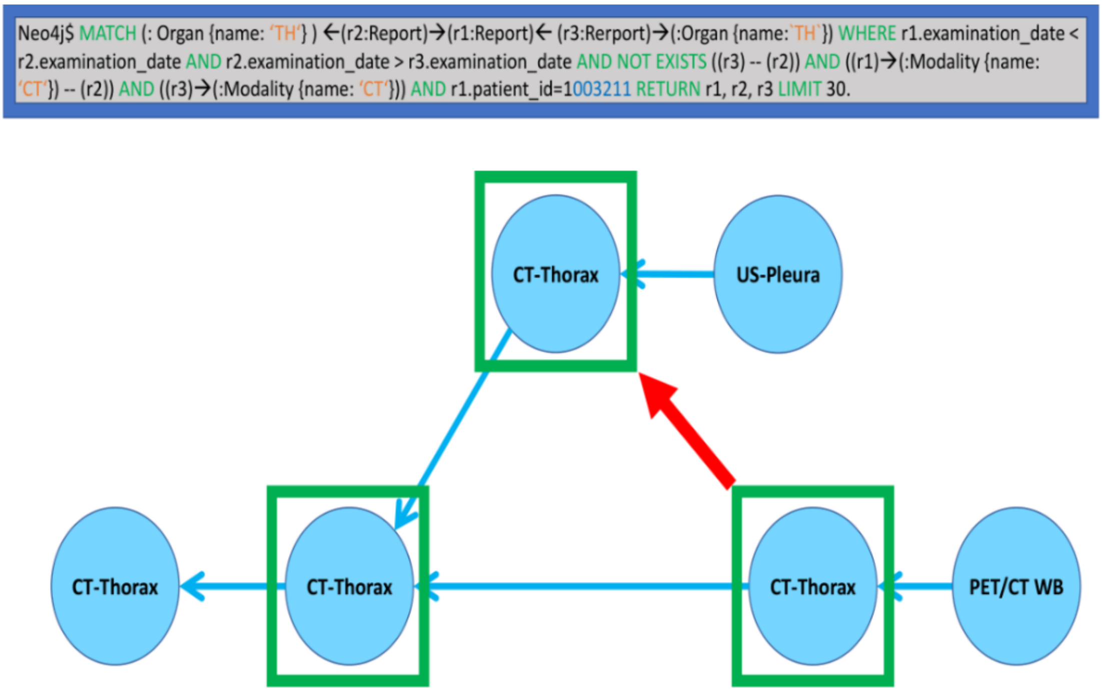
Figure 6. Chord diagram representing the connections between modalities in the head region referencing a head computed tomography (CT) image (light blue rim) during the 7 days after its acquisition. The size of the arc is proportional to the number of referenced reports. Most referring reports are head magnetic resonance images (MRIs), followed by other head CT images. CR: computed radiography; MR: magnetic resonance; NM: nuclear medicine; OT: other; PT: positron emission tomography; US: ultrasound.



Visual and Filter-Aided Detection of Missing Comparative Connections

Selective queries with sequential filters and graph visualization permitted a rapid assessment of situations in which referral links were missing (Figure 7). This feature was helpful when preceding comparative exams had been overlooked due to the poor list-like appearance of exam history in PACS or radiology information systems as well as when previous external images were imported into the PACS after the acquisition and reading of the following exam.

Figure 7. Single patient case example illustrating a missing temporal reference (red arrow) between subsequent reports (blue nodes, ordered by the earliest acquisition on the left to the latest on the right) of computed tomography (CT) studies of the thorax (green boxes). It is easy to detect a suspected missed link between the earlier CT-Thorax report at the top was not referenced by the later CT-Thorax report at the bottom right as well as via the search queries exposing temporal inconsistencies in the referrals (blue box at the top). PET: positron emission tomography; US: ultrasonography; WB: whole body.



```
Neo4j$ MATCH (: Organ {name: 'TH'} ) ←(r2:Report)→(r1:Report)← (r3:Rerport)→(:Organ {name:`TH`}) WHERE r1.examination_date <
r2.examination_date AND r2.examination_date > r3.examination_date AND NOT EXISTS ((r3) -- (r2)) AND ((r1)→(:Modality {name:
'CT'}) -- (r2)) AND ((r3)→(:Modality {name: 'CT'})) AND r1.patient_id=1003211 RETURN r1, r2, r3 LIMIT 30.
```

Discussion

Principal Findings

As shown in this paper, representing imaging records in a directed graph is feasible. Connecting them via their referring dates improved visualization of related imaging pathways and detected missed exam comparisons. We also showed that automated extraction of referring dates from written radiology reports using a deep learning–based NLP algorithm, the groundwork needed to create the representation, is feasible and achievable with high significance (F_1-score of 0.94).

Considering the extraction of concepts of temporality using NLP, our method can be compared with a publication from 2019 by Bozkurt et al [60]. Their main focus was extracting measurements and their core descriptors, among other things, their temporal context, for which they used rule-based NLP with predefined regular expressions. They solely focused on 2 temporal aspects (ie, current or prior), and their pipeline had a high F_1-score of 0.85. Our approach uses a date-extracting LSTM. It focuses on all the referring dates in a written report, including the ones without a precise measurement, for example, the lesions that cannot be measured due to an amorphous configuration or the overall comparison date of the report. In addition, our algorithm has the crucial and unique advantage of detecting the explicit absence and missingness of a comparative exam from written text. Furthermore, we extracted every date of comparison from the report, thus permitting a comminuted and precise linkage for constructing a general graph.

Our approach, however, has the main disadvantage of not attributing the comparison date to specific findings or measurements, which will slow down the focused review of specific entities in complex patient histories. Another disadvantage of our more granular extraction method is the high complexity of the task, which consecutively increases its dependency on correctly spelled referencing dates. Following this logic, omitted or wrongly chosen dates would have a greater impact on the integrity of the machine learning model and the graph in addition to the effect of varying writing habits or report templates between different institutions or radiologists. Although the reporting guidelines favor precisely dated comparisons, the radiologist does not always explicitly write the exact date of the compared finding in the text. As this omission mostly happens in comparison to the most recent report, which would be mentioned at the beginning of the report as the last referenced report, our method covers the majority of these cases. These aspects may render the overall applicability of our model more complex and susceptible to smaller errors than the temporality extraction algorithms developed so far.

In 2006, Lakhani et al [78] explored, in their large-scale database analysis of 1.8 million reports, how often radiologists compared with prior studies using a SQL approach. They found that 42.5% of reports completely omitted any reference to previous studies, 38.7% mentioned a comparison, and no relevant comparison was explicitly pointed out in only 18.8%. Although not entirely comparable, as they focused on a purely semantic approach of referring information extraction, it provides a good approximation, because if the reports contained phrases hinting toward a comparison, the date of the compared exam was most probably mentioned. In our study, reports referenced the date of a comparable exam (53%), explicitly stated that there was no previous exam (21%) more often, and were less prone to miss the referring link (26%). The best year for indicating temporal references was 2021, with only 17.9% of reports missing a reference, down from 30.4% in 2019 and 28.8% in 2020. This tendency toward more temporal referrals could result from the increased emphasis on comparison exam consultation and report structuring in current reporting guidelines and digitalization, with many previous studies easily accessible. However, these percentages of prior exam consultation based on the written references in radiology reports are most probably underestimates. Haygood et al [79] concluded in their study from 2018 that the assumption that an older radiologic image or medical document was not consulted during radiologic interpretation merely because it is not cited in the report was not valid. This causes medicolegal issues for the reader. Radiologists were found negligent by juries for failure to compare a new chest radiograph with all previous chest radiographs [80]. Without written proof, this gets more difficult to defend. Another relevant aspect is the different extraction approaches. Our sentence-based named-entity recognition system analyzed data on a granular level, thus not missing single dates meant for comparison in parenthesis or other dates without a clear semantic indication of referral like the SQL approaches required.

Studies analyzing errors in the radiology reporting process emphasized the importance of comparing findings [52-56]. The good practice guidelines from the European Society of Radiology [50] and the 2020 revised American College of Radiology practice parameter for communication of diagnostic imaging findings support this affirmation. Kim and Mansfield [55] found that 5% of all errors in radiology resulted from failure to consult prior radiographic studies that could have led to the correct diagnosis. However, a critical review of the previous radiologists' findings or impressions should prevail when comparing previous exams. One must be careful not to follow an incorrect path; this error, called “satisfaction of report,” accounted for 6% of all errors reported in radiology in the study by Kim and Mansfield [55]. The widespread availability of previous exams in modern PACS renders an excuse for failing to compare findings with prior exams obsolete. The automatic selection of comparative exams offered by modern PACS is inherently biased because it primarily considers the locoregional aspect, thus losing focus on multiregionality. For example, a CT of the cervical spine or shoulder may be overlooked as a potential comparison source when evaluating apical lung masses, or abdominal radiographs when interpreting hips. The same logic applies to clinicians or radiologists reviewing the imaging history of a given finding, especially in oncology, which has many multiregional studies and findings.

These complex considerations call for a well-arranged and organized visualization system. Poor usability and hampered visualization of patient data reduce the motivation of thoroughly reviewing them, which remains a challenge in health care and is associated with increased error rates due to missing pertinent details, user fatigue, and frustration [78,79]. A study from 2022 analyzing the impact of intensive care unit clinical information systems showed that poor interface design and visual representations are major sources of dissatisfaction among users [80]. Our explorations indicate that grouping related exams together in a graph could help improve this fundamental and increasingly pressing user-friendliness issue.

We hope that, by reinforcing the radiologist's organizing role and improving the case overview by replacing the list appearance of imaging history, he or she will tend to omit the referring links less often, thus minimizing comparison error. Another critical aid is the improved detection of omitted connections in situations in which, for example, previously acquired external scans were loaded into the PACS after reading the following exam. This would be of great value for the subsequent physicians reviewing the imaging history. Temporal referrals in a report prove to the reader that the radiologist has not forgotten to compare a specific finding. This is a valuable asset, considering that a finding's relevance is often determined by its temporal course. For example, lung nodules, brain atrophy changes, or vascular aneurysms showing no dynamic changes over a long period are less alarming, especially in infants and older adults, for whom noninvasive imaging follow-ups are favored over invasive medical investigations. Optimizing visualization with a graph representation could save time as well as decrease unnecessary exams and radiation exposure for patients.

In specialized medicine, clinicians are more focused on specific regions or findings. Manually filtering out the irrelevant exams adds work and a source of potential error (eg, an orthopedic surgeon is more inclined to investigate images implying the healing process of a fracture or a neurologist the exams related to cerebral or spinal findings). Our graph enables the user with an exam of interest to select all the related studies and to omit, if desired, all the unrelated reports, thus substantially and instantly reducing the number of studies to be reviewed.

Our system can assist quality control and review of guideline adherence by rapidly filtering out selectable sequences of exams (eg, CT performed after an x-ray) refined by the possibility of restricting the search for an interstudy period. This highly customizable review based on the reports' metadata could also help research projects. For example, when evaluating the features of a brain lesion over time, one could filter out all the reports in the database in which the finding is described in the report text; these reports will then be shown, if desired independently of the patient, with their respective related reports. This approach rapidly and intuitively speeds up an otherwise fastidious query, offering the researcher a follow-up and quick method for taking the measurement steps on the associated images. The quantitative and qualitative predictions as well as the period of the related following radiology exams could be of great value for clinical management purposes, permitting optimal prediction of the necessary human and material resources.

Limitations

Our study has several limitations. The main limitation was that the analysis was based on a single tertiary care university hospital and depended strongly on our reporting customs. Second, reports were labeled by only 1 reader (a second-year resident). Given the low grade of complexity in labeling the referencing dates and the 100% agreement in a subset of 100 reports, we refrained from a second reading of the whole data set. The more challenging task of determining the comparative study was done during the reporting process by at least one board-certified radiologist. Third, there were insufficient samples to train the non-numerical referencing dates expressing “yesterday.” This should be addressed in future work. One solution could be to use active learning algorithms prioritizing the model's most uncertain predictions. Fourth, there was a lack of external validation. Also, to our knowledge, there is no comparable study in the literature. Nevertheless, the methodology should be reproducible in other radiology department setups to allow for future comparison. To this end, we have also made the codebase that allows for internal testing available (Multimedia Appendix 2). Fifth, the focus here was on the feasibility of an entire pipeline, including extraction and representation. Thus, we did not thoroughly evaluate its clinical usefulness but, instead, illustrated the potential usefulness in several use cases.

Future Prospects

The high performance of our NLP-based model at processing immense amounts of free-text data underlines its potential for future research projects. The process of filtering out comparative studies could be accelerated substantially, which could greatly benefit the development of image detection–based and NLP-based algorithms. The concept of a related graph database could optimize the engineering and designing of other medical software tools in radiology by improving visualization and user-friendliness, accelerating data selection in research projects, and enhancing quality control and clinical review processes. An important amelioration could be the connection of the dates to the specific findings or measurements to which they are referring. Furthermore, it could enable resource planners to separately predict the necessary human and material resources. A significant asset of these databases is the easy-to-implement expansions (eg, integration of pathology reports or associated images). By giving users the possibility of correcting and adding links, it would be conceivable to create a continuously self-improving algorithm.

Conclusion

We established a proof of concept of an NLP-based algorithm capable of accurately extracting the dates of referrals on a granular level from unstructured radiology reports. We successfully generated customizable graphs of referring radiology reports, in which multiple filters may freely be applied, providing a well-arranged visual overview. This type of visualization permitted new possibilities for querying specific exam sequences, facilitated the detection of missed comparisons by the radiologist, and offers health care professionals a wide range of review opportunities. The radiologist's awareness and motivation for the comparative aspect of his or her findings could be increased, and his or her worth for clinicians could be augmented by not solely providing information but also actively helping to organize it. Further work is needed to expand its features and evaluate its definite benefits in day-to-day clinical practice.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Defining body regions and modality types.
[DOCX File , 14 KB - medinform_v10i12e40534_app1.docx]

Multimedia Appendix 2
Data format and machine learning.
[DOCX File , 14 KB - medinform_v10i12e40534_app2.docx]

Multimedia Appendix 3
Distinct count in the whole data set.
[DOCX File , 35 KB - medinform_v10i12e40534_app3.docx]

Multimedia Appendix 4
Five-fold cross-validation results.
[DOCX File , 13 KB - medinform_v10i12e40534_app4.docx]

Multimedia Appendix 5
Metadata analysis of the temporal references.
[DOCX File , 19 KB - medinform_v10i12e40534_app5.docx]

References

1. Grieve FM, Plumb AA, Khan SH. Radiology reporting: a general practitioner's perspective. Br J Radiol 2010 Jan;83(985):17-22 [FREE Full text] [doi: 10.1259/bjr/16360063] [Medline: 19470574]

2. Lee B, Whitehead MT. Radiology reports: What YOU think you're saying and what THEY think you're saying. Curr Probl Diagn Radiol 2017 May;46(3):186-195. [doi: 10.1067/j.cpradiol.2016.11.005] [Medline: 28069356]
3. Smith-Bindman R, Miglioretti DL, Johnson E, Lee C, Feigelson HS, Flynn M, et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996-2010. JAMA 2012 Jun 13;307(22):2400-2409 [FREE Full text] [doi: 10.1001/jama.2012.5960] [Medline: 22692172]
4. Avrin DE, Urbania TH. Demise of film. Acad Radiol 2014 Mar;21(3):303-304. [doi: 10.1016/j.acra.2013.12.008] [Medline: 24507419]
5. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform 2020 Mar 31;8(3):e17984 [FREE Full text] [doi: 10.2196/17984] [Medline: 32229465]
6. Short RG, Bralich J, Bogaty D, Befera NT. Comprehensive word-Level classification of screening mammography reports using a neural network sequence labeling approach. J Digit Imaging 2019 Oct 18;32(5):685-692 [FREE Full text] [doi: 10.1007/s10278-018-0141-4] [Medline: 30338478]
7. Verspoor K, Cohen KB. Natural Language Processing. In: Dubitzky W, Wolkenhauer O, Cho KH, Yokota H, editors. Encyclopedia of Systems Biology. New York, NY: Springer; 2013:1495-1498.
8. Goldberg Y. A primer on neural network models for natural language processing. Journal of Artificial Intelligence Research 2016 Nov 20;57:345-420 [FREE Full text] [doi: 10.1613/jair.4992]
9. Hao T, Huang Z, Liang L, Weng H, Tang B. Health natural language processing: methodology development and applications. JMIR Med Inform 2021 Oct 21;9(10):e23898 [FREE Full text] [doi: 10.2196/23898] [Medline: 34673533]
10. Dreisbach C, Koleck TA, Bourne PE, Bakken S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. Int J Med Inform 2019 May;125:37-46 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.02.008] [Medline: 30914179]
11. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform 2019 Apr 27;7(2):e12239 [FREE Full text] [doi: 10.2196/12239] [Medline: 31066697]
12. Bae JH, Han HW, Yang SY, Song G, Sa S, Chung GE, et al. Natural language processing for assessing quality indicators in free-text colonoscopy and pathology reports: development and usability study. JMIR Med Inform 2022 Apr 15;10(4):e35257 [FREE Full text] [doi: 10.2196/35257] [Medline: 35436226]
13. Chen J, Gong Z, Liu W. A nonparametric model for online topic discovery with word embeddings. Information Sciences 2019 Dec;504:32-47. [doi: 10.1016/j.ins.2019.07.048]
14. European Society of Radiology (ESR). What the radiologist should know about artificial intelligence - an ESR white paper. Insights Imaging 2019 Apr 04;10(1):44 [FREE Full text] [doi: 10.1186/s13244-019-0738-2] [Medline: 30949865]
15. Chen P, Zafar H, Galperin-Aizenberg M, Cook T. Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports. J Digit Imaging 2018 Apr 27;31(2):178-184 [FREE Full text] [doi: 10.1007/s10278-017-0027-x] [Medline: 29079959]
16. Luo JW, Chong JJ. Review of natural language processing in radiology. Neuroimaging Clin N Am 2020 Nov;30(4):447-458. [doi: 10.1016/j.nic.2020.08.001] [Medline: 33038995]
17. Dahl FA, Rama T, Hurlen P, Brekke PH, Husby H, Gundersen T, et al. Neural classification of Norwegian radiology reports: using NLP to detect findings in CT-scans of children. BMC Med Inform Decis Mak 2021 Mar 04;21(1):84 [FREE Full text] [doi: 10.1186/s12911-021-01451-8] [Medline: 33663479]
18. Liu H, Zhang Z, Xu Y, Wang N, Huang Y, Yang Z, et al. Use of BERT (Bidirectional Encoder Representations from Transformers)-based deep learning method for extracting evidences in Chinese radiology reports: development of a computer-aided liver cancer diagnosis framework. J Med Internet Res 2021 Jan 12;23(1):e19689 [FREE Full text] [doi: 10.2196/19689] [Medline: 33433395]
19. Bressem K, Adams L, Gaudin R, Tröltzsch D, Hamm B, Makowski MR, et al. Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. Bioinformatics 2021 Jan 29;36(21):5255-5261. [doi: 10.1093/bioinformatics/btaa668] [Medline: 32702106]
20. Barash Y, Guralnik G, Tau N, Soffer S, Levy T, Shimon O, et al. Comparison of deep learning models for natural language processing-based classification of non-English head CT reports. Neuroradiology 2020 Oct 25;62(10):1247-1256. [doi: 10.1007/s00234-020-02420-0] [Medline: 32335686]
21. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, et al. Deep learning: a primer for radiologists. Radiographics 2017;37(7):2113-2131. [doi: 10.1148/rg.2017170077] [Medline: 29131760]
22. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019 Jan 7;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
23. 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 Jul 26;6(1):30 [FREE Full text] [doi: 10.1186/1472-6947-6-30] [Medline: 16872495]
24. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ 2015 Apr 24;350(apr24 11):h1885 [FREE Full text] [doi: 10.1136/bmj.h1885] [Medline: 25911572]

25. 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]
26. Jung K, LePendu P, Iyer S, Bauer-Mehren A, Percha B, Shah N. Functional evaluation of out-of-the-box text-mining tools for data-mining tasks. J Am Med Inform Assoc 2015 Jan;22(1):121-131 [FREE Full text] [doi: 10.1136/amiajnl-2014-002902] [Medline: 25336595]
27. Carrell D, Halgrim S, Tran D, 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]
28. Huang SH, LePendu P, Iyer SV, Tai-Seale M, Carrell D, Shah NH. Toward personalizing treatment for depression: predicting diagnosis and severity. J Am Med Inform Assoc 2014 Nov 01;21(6):1069-1075 [FREE Full text] [doi: 10.1136/amiajnl-2014-002733] [Medline: 24988898]
29. Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, et al. Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inform Assoc 2016 Nov 12;23(6):1166-1173 [FREE Full text] [doi: 10.1093/jamia/ocw028] [Medline: 27174893]
30. Trivedi G, Dadashzadeh ER, Handzel RM, Chapman WW, Visweswaran S, Hochheiser H. Interactive NLP in clinical care: identifying incidental findings in radiology reports. Appl Clin Inform 2019 Aug 04;10(4):655-669 [FREE Full text] [doi: 10.1055/s-0039-1695791] [Medline: 31486057]
31. Dutta S, Long WJ, Brown DF, Reisner AT. Automated detection using natural language processing of radiologists recommendations for additional imaging of incidental findings. Ann Emerg Med 2013 Aug;62(2):162-169. [doi: 10.1016/j.annemergmed.2013.02.001] [Medline: 23548405]
32. Trivedi G, Hong C, Dadashzadeh ER, Handzel RM, Hochheiser H, Visweswaran S. Identifying incidental findings from radiology reports of trauma patients: An evaluation of automated feature representation methods. Int J Med Inform 2019 Sep;129:81-87 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.05.021] [Medline: 31445293]
33. Visser JJ, de Vries M, Kors JA. Automatic detection of actionable findings and communication mentions in radiology reports using natural language processing. Eur Radiol 2022 Jun 06;32(6):3996-4002. [doi: 10.1007/s00330-021-08467-8] [Medline: 34989840]
34. Li AY, Elliot N. Natural language processing to identify ureteric stones in radiology reports. J Med Imaging Radiat Oncol 2019 Jun 05;63(3):307-310. [doi: 10.1111/1754-9485.12861] [Medline: 30720244]
35. Dublin S, Baldwin E, Walker RL, Christensen LM, Haug PJ, Jackson ML, et al. Natural language processing to identify pneumonia from radiology reports. Pharmacoepidemiol Drug Saf 2013 Aug 01;22(8):834-841 [FREE Full text] [doi: 10.1002/pds.3418] [Medline: 23554109]
36. Grundmeier R, Masino A, Casper T, Dean J, Bell J, Enriquez R, et al. Identification of long bone fractures in radiology reports using natural language processing to support healthcare quality improvement. Appl Clin Inform 2017 Dec 18;07(04):1051-1068. [doi: 10.4338/aci-2016-08-ra-0129]
37. Kolanu N, Brown AS, Beech A, Center JR, White CP. Natural language processing of radiology reports for the identification of patients with fracture. Arch Osteoporos 2021 Jan 06;16(1):6. [doi: 10.1007/s11657-020-00859-5] [Medline: 33403479]
38. Senders JT, Cho LD, Calvachi P, McNulty JJ, Ashby JL, Schulte IS, et al. Automating clinical chart review: an open-source natural language processing pipeline developed on free-text radiology reports from patients with glioblastoma. JCO Clin Cancer Inform 2020 Jan;4:25-34 [FREE Full text] [doi: 10.1200/CCI.19.00060] [Medline: 31977252]
39. Sevenster M, Buurman J, Liu P, Peters J, Chang P. Natural language processing techniques for extracting and categorizing finding measurements in narrative radiology reports. Appl Clin Inform 2017 Dec 19;06(03):600-610. [doi: 10.4338/aci-2014-11-ra-0110]
40. Cheng LTE, Zheng J, Savova GK, Erickson BJ. Discerning tumor status from unstructured MRI reports--completeness of information in existing reports and utility of automated natural language processing. J Digit Imaging 2010 Apr 30;23(2):119-132 [FREE Full text] [doi: 10.1007/s10278-009-9215-7] [Medline: 19484309]
41. Senders JT, Karhade AV, Cote DJ, Mehrtash A, Lamba N, DiRisio A, et al. Natural language processing for automated quantification of brain metastases reported in free-text radiology reports. JCO Clinical Cancer Informatics 2019 Dec(3):1-9. [doi: 10.1200/cci.18.00138]
42. Lacson R, Prevedello LM, Andriole KP, Gill R, Lenoci-Edwards J, Roy C, Fleischner Society. Factors associated with radiologists' adherence to Fleischner Society guidelines for management of pulmonary nodules. J Am Coll Radiol 2012 Jul;9(7):468-473. [doi: 10.1016/j.jacr.2012.03.009] [Medline: 22748786]
43. Duszak R, Nossal M, Schofield L, Picus D. Physician documentation deficiencies in abdominal ultrasound reports: frequency, characteristics, and financial impact. J Am Coll Radiol 2012 Jun;9(6):403-408. [doi: 10.1016/j.jacr.2012.01.006] [Medline: 22632666]
44. Petkov VI, Penberthy LT, Dahman BA, Poklepovic A, Gillam CW, McDermott JH. Automated determination of metastases in unstructured radiology reports for eligibility screening in oncology clinical trials. Exp Biol Med (Maywood) 2013 Dec 09;238(12):1370-1378 [FREE Full text] [doi: 10.1177/1535370213508172] [Medline: 24108448]

45. Zhou Y, Amundson PK, Yu F, Kessler MM, Benzinger TLS, Wippold FJ. Automated classification of radiology reports to facilitate retrospective study in radiology. J Digit Imaging 2014 Dec 30;27(6):730-736 [FREE Full text] [doi: 10.1007/s10278-014-9708-x] [Medline: 24874407]
46. Hripcsak G, Austin JHM, Alderson PO, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. Radiology 2002 Jul;224(1):157-163. [doi: 10.1148/radiol.2241011118] [Medline: 12091676]
47. Chapman BE, Lee S, Kang HP, Chapman WW. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. J Biomed Inform 2011 Oct;44(5):728-737 [FREE Full text] [doi: 10.1016/j.jbi.2011.03.011] [Medline: 21459155]
48. Chen MC, Ball RL, Yang L, Moradzadeh N, Chapman BE, Larson DB, et al. Deep learning to classify radiology free-text reports. Radiology 2018 Mar;286(3):845-852. [doi: 10.1148/radiol.2017171115] [Medline: 29135365]
49. Yu S, Kumamaru KK, George E, Dunne RM, Bedayat A, Neykov M, et al. Classification of CT pulmonary angiography reports by presence, chronicity, and location of pulmonary embolism with natural language processing. J Biomed Inform 2014 Dec;52:386-393 [FREE Full text] [doi: 10.1016/j.jbi.2014.08.001] [Medline: 25117751]
50. European Society of Radiology (ESR). Good practice for radiological reporting. Guidelines from the European Society of Radiology (ESR). Insights Imaging 2011 Apr 6;2(2):93-96 [FREE Full text] [doi: 10.1007/s13244-011-0066-7] [Medline: 22347937]
51. Aideyan UO, Berbaum K, Smith WL. Influence of prior radiologic information on the interpretation of radiographic examinations. Academic Radiology 1995 Mar;2(3):205-208. [doi: 10.1016/s1076-6332(05)80165-5]
52. White K, Berbaum K, Smith WL. The role of previous radiographs and reports in the interpretation of current radiographs. Invest Radiol 1994 Mar;29(3):263-265. [doi: 10.1097/00004424-199403000-00002] [Medline: 8175298]
53. Hunter T, Boyle R. The value of reading the previous radiology report. AJR Am J Roentgenol 1988 Mar 01;150(3):697-698. [doi: 10.2214/ajr.150.3.697] [Medline: 3257629]
54. Bruno MA, Walker EA, Abujudeh HH. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics 2015 Oct;35(6):1668-1676. [doi: 10.1148/rg.2015150023] [Medline: 26466178]
55. Kim YW, Mansfield LT. Fool me twice: delayed diagnoses in radiology with emphasis on perpetuated errors. AJR Am J Roentgenol 2014 Mar;202(3):465-470. [doi: 10.2214/AJR.13.11493] [Medline: 24555582]
56. Yankaskas BC, May RC, Matuszewski J, Bowling JM, Jarman MP, Schroeder BF. Effect of observing change from comparison mammograms on performance of screening mammography in a large community-based population. Radiology 2011 Dec;261(3):762-770 [FREE Full text] [doi: 10.1148/radiol.11110653] [Medline: 22031709]
57. Sevenster M, Buurman J, Liu P, Peters J, Chang P. Natural language processing techniques for extracting and categorizing finding measurements in narrative radiology reports. Appl Clin Inform 2017 Dec 19;06(03):600-610. [doi: 10.4338/aci-2014-11-ra-0110]
58. Sevenster M, Bozeman J, Cowhy A, Trost W. A natural language processing pipeline for pairing measurements uniquely across free-text CT reports. J Biomed Inform 2015 Feb;53:36-48 [FREE Full text] [doi: 10.1016/j.jbi.2014.08.015] [Medline: 25200472]
59. Yim W, Kwan SW, Yetisgen M. Classifying tumor event attributes in radiology reports. Journal of the Association for Information Science and Technology 2017 Sep 14;68(11):2662-2674. [doi: 10.1002/asi.23937]
60. Bozkurt S, Alkim E, Banerjee I, Rubin DL. Automated detection of measurements and their descriptors in radiology reports using a hybrid natural language processing algorithm. J Digit Imaging 2019 Aug;32(4):544-553 [FREE Full text] [doi: 10.1007/s10278-019-00237-9] [Medline: 31222557]
61. Banerjee I, Bozkurt S, Caswell-Jin JL, Kurian AW, Rubin DL. Natural language processing approaches to detect the timeline of metastatic recurrence of breast cancer. JCO Clinical Cancer Informatics 2019 Dec(3):1-12. [doi: 10.1200/cci.19.00034]
62. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004;11(5):392-402 [FREE Full text] [doi: 10.1197/jamia.M1552] [Medline: 15187068]
63. Cox S, Ahalt SC, Balhoff J, Bizon C, Fecho K, Kebede Y, et al. Visualization environment for federated knowledge graphs: development of an interactive biomedical query language and web application interface. JMIR Med Inform 2020 Nov 23;8(11):e17964 [FREE Full text] [doi: 10.2196/17964] [Medline: 33226347]
64. Sun H, Xiao J, Zhu W, He Y, Zhang S, Xu X, et al. Medical knowledge graph to enhance fraud, waste, and abuse detection on claim data: model development and performance evaluation. JMIR Med Inform 2020 Jul 23;8(7):e17653 [FREE Full text] [doi: 10.2196/17653] [Medline: 32706714]
65. Schrodt J, Dudchenko A, Knaup-Gregori P, Ganzinger M. Graph-representation of patient data: a systematic literature review. J Med Syst 2020 Mar 12;44(4):86 [FREE Full text] [doi: 10.1007/s10916-020-1538-4] [Medline: 32166501]
66. Ujiie S, Yada S, Wakamiya S, Aramaki E. Identification of adverse drug event-related Japanese articles: natural language processing analysis. JMIR Med Inform 2020 Nov 27;8(11):e22661 [FREE Full text] [doi: 10.2196/22661] [Medline: 33245290]

67. Dai H, Lee Y, Nekkanti C, Jonnagaddala J. Family history information extraction with neural attention and an enhanced relation-side scheme: algorithm development and validation. JMIR Med Inform 2020 Dec 01;8(12):e21750 [FREE Full text] [doi: 10.2196/21750] [Medline: 33258777]
68. Charniak E. Statistical Techniques for Natural Language Parsing. AI Magazine 1997;18(4):33 [FREE Full text] [doi: 10.1609/aimag.v18i4.1320]
69. Martinez AR. Part-of-speech tagging. WIREs Comp Stat 2011 Sep 30;4(1):107-113. [doi: 10.1002/wics.195]
70. Stubbs A, Kotfila C, Uzuner Ö. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. J Biomed Inform 2015 Dec;58 Suppl(Suppl):S11-S19 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.007] [Medline: 26225918]
71. Maiya AS. ktrain: a low-code library for augmented machine learning. Journal of Machine Learning Research 2022;23:1-6 [FREE Full text]
72. usb-ai / apophenator. GitHub. URL: https://github.com/usb-ai/apophenator [accessed 2022-12-05]
73. Lin C, Hsu C, Lou Y, Yeh S, Lee C, Su S, et al. Artificial intelligence learning semantics via external resources for classifying diagnosis codes in discharge notes. J Med Internet Res 2017 Nov 06;19(11):e380 [FREE Full text] [doi: 10.2196/jmir.8344] [Medline: 29109070]
74. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw 2005;18(5-6):602-610. [doi: 10.1016/j.neunet.2005.06.042] [Medline: 16112549]
75. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 2017 Dec;5:135-146 [FREE Full text] [doi: 10.1162/tacl_a_00051]
76. Hastie T, Tibshirani R, Friedman J. Model Assessment and Selection. In: The Elements of Statistical Learning. New York, NY: Springer; 2009:219-259.
77. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 2015 Aug 12;15(1):29 [FREE Full text] [doi: 10.1186/s12880-015-0068-x] [Medline: 26263899]
78. Lakhani P, Menschik ED, Goldszal AF, Murray JP, Weiner MG, Langlotz CP. Development and validation of queries using structured query language (SQL) to determine the utilization of comparison imaging in radiology reports stored on PACS. J Digit Imaging 2006 Mar 1;19(1):52-68 [FREE Full text] [doi: 10.1007/s10278-005-7667-y] [Medline: 16132483]
79. Haygood TM, Mullins B, Sun J, Amini B, Bhosale P, Kang HC, et al. Consultation and citation rates for prior imaging studies and documents in radiology. J. Med. Imag 2018 Jul 1;5(03):1. [doi: 10.1117/1.jmi.5.3.031409]
80. Berlin L. Must new radiographs be compared with all previous radiographs, or only with the most recently obtained radiographs? AJR Am J Roentgenol 2000 Mar;174(3):611-615. [doi: 10.2214/ajr.174.3.1740611] [Medline: 10701597]

Abbreviations

CT: computed tomography
DICOM: Digital Imaging and Communications in Medicine
LSTM: long short-term memory
MRI: magnetic resonance imaging
NLP: natural language processing
PACS: picture archiving and communication system
PET: positron emission tomography

Edited by A Benis; submitted 29.06.22; peer-reviewed by I Banerjee, R Sauvayre; comments to author 21.07.22; revised version received 13.09.22; accepted 30.11.22; published 21.12.22.

Please cite as:
Binsfeld Gonçalves L, Nesic I, Obradovic M, Stieltjes B, Weikert T, Bremerich J
Natural Language Processing and Graph Theory: Making Sense of Imaging Records in a Novel Representation Frame
JMIR Med Inform 2022;10(12):e40534
URL: https://medinform.jmir.org/2022/12/e40534
doi:10.2196/40534
PMID:36542426

Original Paper

Telehealth System Based on the Ontology Design of a Diabetes Management Pathway Model in China: Development and Usability Study

ZhiYuan Fan[1,2], MSCE; LiYuan Cui[3], MSci; Ying Ye[1], MSci; ShouCheng Li[1], BA; Ning Deng[1,2], PhD

[1]College of Biomedical Engineering and Instrument Science, Ministry of Education Key Laboratory of Biomedical Engineering, Zhejiang University, Hangzhou, China

[2]Binjiang Institute of Zhejiang University, Hangzhou, China

[3]School of Medical Imaging, Hangzhou Medical College, HangZhou, China

Corresponding Author:
Ning Deng, PhD
College of Biomedical Engineering and Instrument Science
Ministry of Education Key Laboratory of Biomedical Engineering
Zhejiang University
Zhouyiqing Bldg 512
38 Zheda Rd
Hangzhou, 310000
China
Phone: 86 571 2295 2693
Email: zju.dengning@gmail.com

Abstract

Background: Diabetes needs to be under control through management and intervention. Management of diabetes through mobile health is a practical approach; however, most diabetes mobile health management systems do not meet expectations, which may be because of the lack of standardized management processes in the systems and the lack of intervention implementation recommendations in the management knowledge base.

Objective: In this study, we aimed to construct a diabetes management care pathway suitable for the actual situation in China to express the diabetes management care pathway using ontology and develop a diabetes closed-loop system based on the construction results of the diabetes management pathway and apply it practically.

Methods: This study proposes a diabetes management care pathway model in which the management process of diabetes is divided into 9 management tasks, and the Diabetes Care Pathway Ontology (DCPO) is constructed to represent the knowledge contained in this pathway model. A telehealth system, which can support the comprehensive management of patients with diabetes while providing active intervention by physicians, was designed and developed based on the DCPO. A retrospective study was performed based on the data records extracted from the system to analyze the usability and treatment effects of the DCPO.

Results: The diabetes management pathway ontology constructed in this study contains 119 newly added classes, 28 object properties, 58 data properties, 81 individuals, 426 axioms, and 192 Semantic Web Rule Language rules. The developed mobile medical system was applied to 272 patients with diabetes. Within 3 months, the average fasting blood glucose of the patients decreased by 1.34 mmol/L (P=.003), and the average 2-hour postprandial blood glucose decreased by 2.63 mmol/L (P=.003); the average systolic and diastolic blood pressures decreased by 11.84 mmHg (P=.02) and 8.8 mmHg (P=.02), respectively. In patients who received physician interventions owing to abnormal attention or low-compliance warnings, the average fasting blood glucose decreased by 2.45 mmol/L (P=.003), and the average 2-hour postprandial blood glucose decreased by 2.89 mmol/L (P=.003) in all patients with diabetes; the average systolic and diastolic blood pressure decreased by 20.06 mmHg (P=.02) and 17.37 mmHg (P=.02), respectively, in patients with both hypertension and diabetes during the 3-month management period.

Conclusions: This study helps guide the timing and content of interactive interventions between physicians and patients and regulates physicians' medical service behavior. Different management plans are formulated for physicians and patients according to different characteristics to comprehensively manage various cardiovascular risk factors. The application of the DCPO in the diabetes management system can provide effective and adequate management support for patients with diabetes and those with both diabetes and hypertension.

(JMIR Med Inform 2022;10(12):e42664) doi:10.2196/42664

KEYWORDS

diabetes; chronic disease management; Chronic Disease Management Pathway; ontology; Semantic Web Rule Language rules; SWRL rules

Introduction

Background

Diabetes mellitus is one of the most common chronic diseases in China [1]. People with diabetes have an increased risk of developing serious health problems [2]. In patients with diabetes, consistent high blood glucose (BG) levels can lead to serious diseases affecting the heart and blood vessels [3]. The combination of lifestyle modifications and self-care therapies as part of diabetes management can significantly increase the treatment rate of diabetes, reduce the incidence of cardiovascular disease, and improve the quality of life of patients [4]. The chronic care model is the most widely used chronic disease management model, which emphasizes that physicians and patients participate in the management process together for collaborative management [5,6]. The purpose of the chronic care model is to remind physicians to provide patients with timely and efficient management, which means the physician will immediately intervene and give feedback on the patient's behavior when the patients complete the management tasks [7]. However, the traditional diabetes management method cannot meet the long-term management needs of patients owing to time and space constraints [8]. In addition, most patients with diabetes have multiple cardiovascular risk factors, such as hypertension, hyperglycemia, and obesity, which are the main causes of death in patients with diabetes [9]. Therefore, in addition to the comprehensive management of patients with diabetes based on the BG level, interventions to control multiple cardiovascular risk factors are also required. With the development of mobile internet technology, diabetes management tends to be digitalized, which relieves the time and space limitations of traditional management methods and realizes the dynamic monitoring and maintenance of patients throughout the entire process managed by medical service providers [10].

Previous Work

Compared with traditional diabetes management methods, the application of mobile medical technology can change the role of patients from passively accepting management services to having the core role in management work, positively improving their self-management awareness. Wyne et al [11] used mobile health technology to manage patients with type 2 diabetes mellitus (T2DM) and effectively prevented serious complications, demonstrating that mobile health technology improved the management of patients with diabetes. Quinn et al [12] conducted a cluster-randomized trial using the BlueStar diabetes care system (WellDoc Inc). The trial results showed that the glycosylated hemoglobin level was significantly reduced and the depression levels and other physiological indicators (blood pressure [BP], lipids, etc) were also improved in the patients engaged in the care and intervention of this system over a 1-year treatment period.

It is necessary to transform medical knowledge and clinical data into computer-recognizable knowledge models using the Semantic Web. As a formal representation of knowledge that can accurately describe the relationship between concepts, ontology has gradually become a key technology for realizing the Semantic Web. The use of ontology to express domain knowledge can facilitate knowledge sharing and dissemination. It is also key to realizing a complete knowledge base and an intelligent clinical decision support system. El-Sappagh et al [13] constructed the Diabetes Diagnosis Ontology based on diabetes clinical practice guidelines and principles of standard medical terminology, established Semantic Web Rule Language (SWRL) diagnostic rules, and used an inference engine to perform diagnostic inference on the T2DM diagnostic knowledge base. Krishnan et al [14] constructed the Diabetes Mellitus Treatment Ontology as a basis for sharing semantic, domain-specific, standard, machine-readable, and interoperable knowledge related to T2DM treatment. Fast Healthcare Interoperability Resources and Semantic Sensor Network–based type 1 diabetes mellitus Ontology is designed for managing patients with type 1 diabetes, which has the Health Level 7-Fast Healthcare Interoperability Resources standard and Semantic Sensor Network sensor ontology integrated, and provides patients with a complete and personalized treatment plan based on the complete patient information [15]. Sherimon et al [16] proposed an ontology-based clinical decision support system for patients with diabetes, which predicts the risk of patients according to various risk factors, including smoking, alcohol consumption, and cardiovascular family history. Chen et al [17] proposed an ontology-based model for the diagnosis and treatment of patients with diabetes who are in hospital, and it can help reduce medication errors.

Key Issues

Previous studies have used diabetes clinical practice guidelines as the theoretical basis for the use of ontologies and considered patient profile data in knowledge-based systems for decision support in hospitals and in-home diabetes monitoring and management, including patient self-monitoring, data recording, and physicians' simple management advice. However, because of the lack of clear regulations on physician-patient interaction and intervention feedback mechanisms in medical guidelines, these studies ignore the importance of physicians' active intervention on patients, do not define the interactive intervention mechanism between physicians and patients, and lack active intervention time and initiative [15-17]. The definition of intervention content and digital management recommendations failed to combine standardized processes and evidence-based medical knowledge, resulting in the inability of patients with diabetes to receive effective intervention guidance promptly in actual management. No research has considered the appropriate timing and details of physician intervention in patients with diabetes when constructing diabetes management ontology.

Furthermore, the existing ontology construction process focuses on hyperglycemia control in patients with diabetes, ignoring the importance of controlling and managing other cardiovascular risk factors and treating these factors as patient-related indicators without targeted attention and guidance in the care process.

Objective

A previous study proposed the concept and construction method for the Chronic Disease Management Pathway (CDMP) [18,19]. CDMP has been proven effective in the treatment of hypertension and chronic obstructive pulmonary disease [20,21]. It is becoming one of the practical approaches for chronic disease management owing to its continuous, closed-loop, and standardized features and provides reliable implementation guidelines for chronic disease management. The introduction and application of the CDMP offer a new solution for the digital management of patients with diabetes.

In this study, we extracted the key issues in diabetes management through evidence-based medical guidelines and expert recommendations, combined the issues with the CDMP approach, and constructed the Type 2 Diabetes Mellitus Management Pathway (T2DMMP) model. A closed-loop Type 2 Diabetes Mellitus Management System (T2DMMS) was developed based on the T2DMMP model, which was expressed by ontology modeling to ensure that computer operations could execute it. Our management pathway was shown to be usable and effective in clinical diabetes management as an implementable intervention mechanism for physician-patient interactions.

Methods

Type 2 Diabetes Mellitus Management Pathway Model

Overview

The diagnosis and management of diabetes mellitus is a complex process. To help with the diabetes management process, in this study, the T2DMMP was constructed based on numerous diabetes prevention guidelines and other cardiovascular risk factors, focusing on clarifying the responsibilities of management roles and providing a standardized and complete management pathway for comprehensive diabetes management. The T2DMMP is a process that divides management tasks into physician intervention plans and patient self-management plans by defining 2 different management roles for physicians and patients.

Summary Extraction of Task Sets

The following literature on current diabetes prevention and management guidelines with high recognition was collected through literature research, and the task sets were extracted from them: Guidelines for the prevention and control of type 2 diabetes in China (2017) Edition [22], National guidelines for the prevention and control of diabetes in primary care (2018) [23], the comprehensive type 2 diabetes management algorithm [24], Management of Hyperglycemia in Type 2 Diabetes [25], and Diabetes Canada Clinical Practice Guidelines Expert Committee [26]. The analysis of the core content of diabetes management was performed based on the guidelines for the prevention and control of type 2 diabetes in China (2017 Edition) and National guidelines for the prevention and control of diabetes in primary care (2018).

Type 2 Diabetes Mellitus Management Pathway

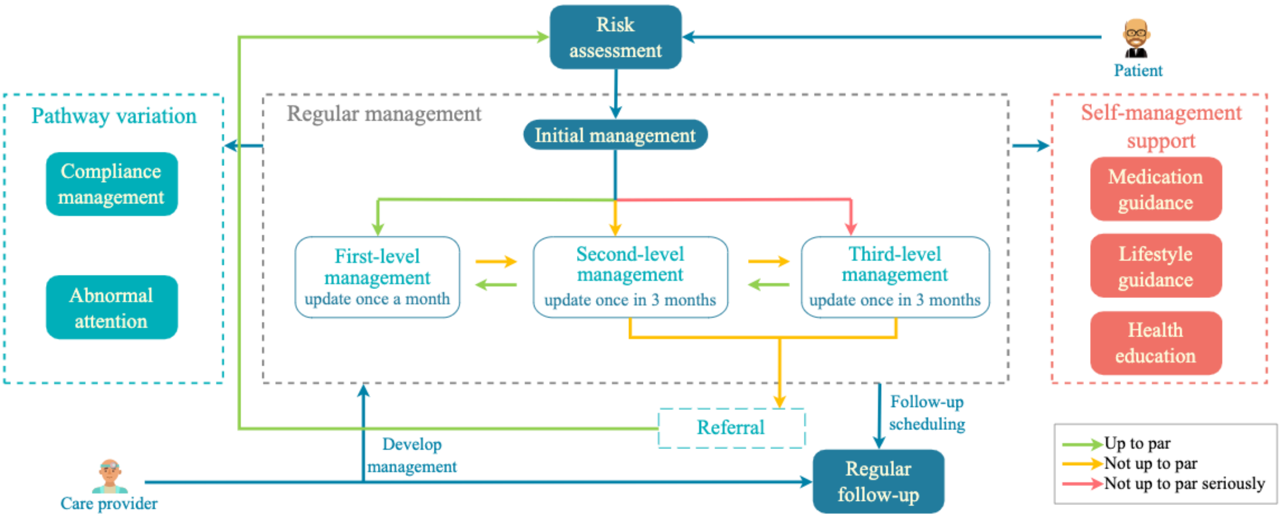
With regard to the CDMP model, 9 common tasks were defined in the T2DMMP, which could be grouped into 3 parts as follows: the task set for regular management, task set for pathway variation, and task set for self-management support. A pathway map of T2DMMP is shown in Figure 1. When a patient is diagnosed with diabetes, they will enter the diabetes management care pathway. A comprehensive cardiovascular risk assessment will be conducted subsequently, and different management levels will be formulated according to the assessment results as follows: patients with stable blood sugar are at routine first-level management; patients with unsatisfactory blood sugar are at routine second-level management; and patients with fasting BG (FBG) >11.1 mmol/L are in intensive third-level management. Different management plans are provided to patients at different management levels. The patient's condition will be periodically evaluated according to the patient's self-monitoring data, and the patient will be dynamically adjusted to the appropriate management level in the management path.

The task set for pathway variation consists of 2 tasks as follows: abnormal attention and compliance management. Once the patient's self-monitoring data (such as BG or BP) is abnormal, the caregiver needs to intervene immediately and appropriately. Once the patient's compliance is low, the care provider needs to conduct additional follow-up to motivate the patient.

The task set for self-management support consists of the following 3 tasks: medication guidance, lifestyle guidance, and health education. Medication guidance is designed to provide medication treatment plans, whereas lifestyle guidance provides non–drug treatment plans such as diet and physical activity. Health education aims to increase patients' awareness of the disease, thereby improving their self-management skills.

Two parties will play a role in management: the care provider and the patient. The care provider team comprises general practitioners, case managers in community health services, and specialists in secondary or tertiary hospitals. According to this pathway, care providers should work collaboratively for day-to-day management, such as regular follow-up and interventions for abnormal conditions. Patients need self-monitoring according to their self-management plans and receive timely intervention from their care providers.

Figure 1. Diagram of the type 2 diabetes mellitus management pathway.

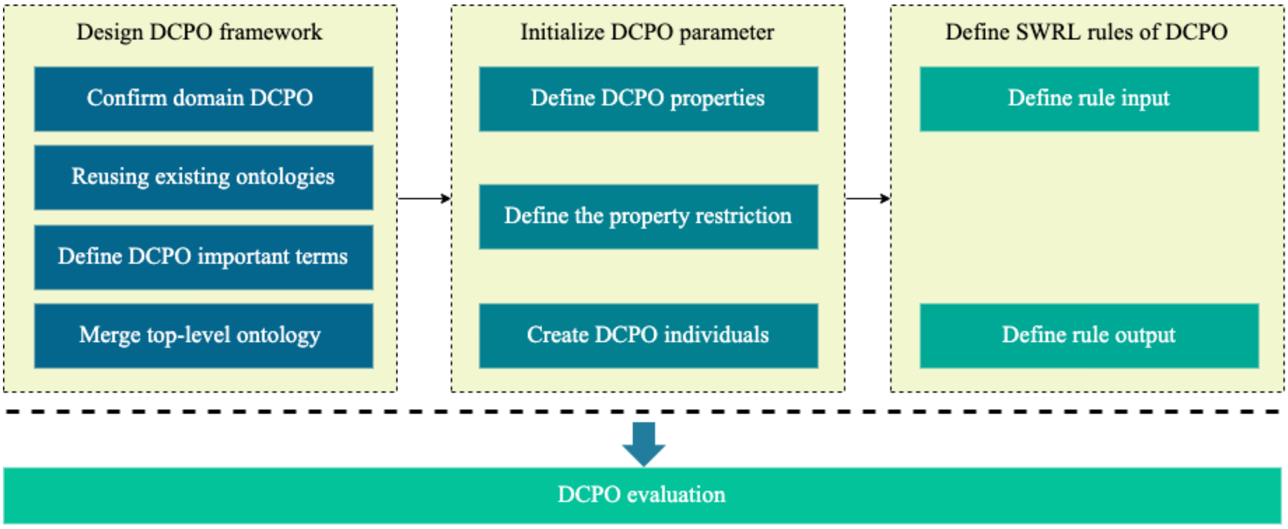


Construction of Diabetes Care Pathway Ontology

Overview

Ontology represents domain knowledge in a machine-readable and formal format [27]. It can be incorporated into a clinical decision support system as a knowledge base [28,29]. The Diabetes Care Pathway Ontology (DCPO) was constructed to describe the concept of T2DMMP. As shown in Figure 2, the DCPO is divided into 3 phases. Stages 1 and 2 follow the ontology engineering approach widely used to represent the structural information of the model, whereas stage 3 incorporates the medical knowledge of the model through external semantic rules.

Figure 2. The Diabetes Care Pathway Ontology (DCPO) construction methodology. SWRL: Semantic Web Rule Language.



Stage 1: DCPO Framework Design

At this stage, a set of questions that the ontology should answer the domain and scope of the DCPO are listed—the so-called capability questions. In this step, the DCPO will be identified as a representative of the T2DMMP model. The expected output of the DCPO is the monitoring and management of patients with diabetes, which includes individualized treatment plans and specific management tasks for physicians and patients.

On the basis of the scalability, standardization, and reusability of the ontologies, already existing diabetes ontologies were considered reusable. Keywords such as “diabetes management,” “diabetes ontology,” “diabetes medication ontology,” “diabetes diet ontology,” and “diabetes treatment ontology” were used to search for content on the BioPortal and PubMed websites. Diabetes Diagnosis Ontology and Diabetes Mellitus Treatment Ontology were also identified for use. To achieve the diabetes management path, the *TIME.owl* ontology of the W3C (World Wide Web Consortium) standard was used as the temporality module. It defines the timing of the management tasks by defining the time intervals and moments. Broad coverage of diabetes treatment medication drugs was introduced, and the drug-drug interaction ontology was reused to describe the types of drugs in the medication guidance module. The lifestyle module included a diet plan and an exercise plan. The diet plan was a set of dietary recommendations for the patient; therefore, the *OntoFood* ontology was reused. On the basis of the T2DMMP model, the DCPO model is defined as a class and class hierarchy and is divided into the following 2 primary levels of abstraction: the first level for the core concepts in the path and the second level for the detailed elements of the first-level classes. The first-level core concepts include patient profile, management roles, management plans, management tasks, and chronological expressions. The patient's self-management plan and the physician's diagnosis and treatment plan defined under the management plan are the contents of the second level. This definition can accurately describe the real world and is widely used in ontological design. The DCPO was merged with 2 top-level ontologies—Basic Formal Ontology and Ontology for General Medical Science—in a method of merging top-level ontologies that are often used by researchers [30-32]; it has been shown to facilitate the reuse of terms from existing ontologies constructed under top-level ontologies [33].

Stage 2: DCPO Model Initialization

In the second stage, class attributes are defined to describe the internal structure of concepts because the class hierarchy is insufficient to distinguish the relationships between concepts. These attributes include object attributes that describe the relationship between 2 individuals and data attributes that describe the relationship between an individual and a data value [34,35]. The precise semantics of restricting classes by adding attributes is accomplished, and these restrictions are expressed as a set of axioms. These axioms include property axioms that describe aspects of properties such as domain and range and individual axioms that describe anonymous individual classes. The instances of each class are created in the hierarchy. The core part of an instance is a class called the patient profile. Related feature instances are created and bound to the patient profile instance for further rule-based reasoning.

Stage 3: Rule Definition

In the third phase, external semantic rules are used to implement the complex deductive reasoning required for path-driven decision support. In this study, the SWRL rules are used to incorporate the medical knowledge of the pathway model. The input and output of the rule and the state representation of the reasoning are defined. The inputs and outputs of the rules vary significantly for different pathway tasks. On the basis of the basic DCPO and predefined SWRL rule sets, various pathway tasks will be generated based on raw patient data and then converted into executable management plans, including physician intervention plans and patient self-management plans. This is a crucial part of the knowledge-based clinical decision support system engines.

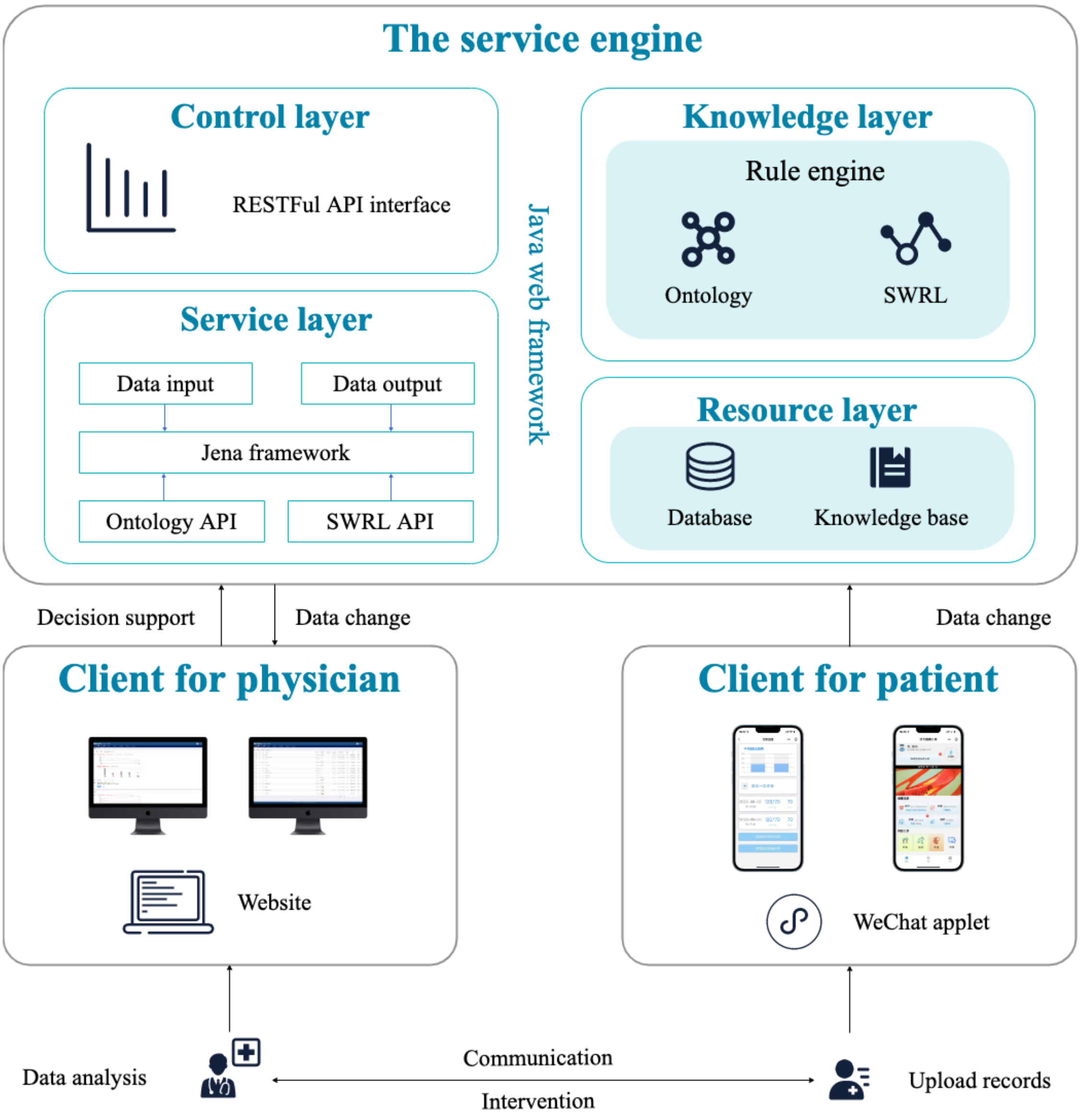
System Deployment

A DCPO-based closed-loop diabetes management system was designed and applied to evaluate the practical applications of the research results. The T2DMMS is involved in a diabetes management service scenario to achieve comprehensive management and intervention guidance for patients with type 2 diabetes.

System Framework Design

The architecture of the T2DMMS is shown in Figure 3, which includes 3 parts as follows: an intelligent service engine, a physician-oriented client, and a patient-oriented client. The service engine runs on the cloud server and is the core module of T2DMMS. Its primary function is to integrate the ontology knowledge base with the SWRL rule base and patient data; it also plays a role in realizing logical reasoning and providing the web service interface to interact with physician-oriented clients and realize decision support based on the T2DMMP. Physician-oriented clients are mainly responsible for displaying patients' health data. Physicians can complete a comprehensive assessment of patients by viewing the data and complete management tasks by providing intervention guidance and improving patient compliance efficiently. For patient-oriented clients, patients can view their self-management plans and complete their daily management tasks by uploading their health data records. In addition, patients can communicate with their physicians through their clients.

Figure 3. The Type 2 Diabetes Mellitus Management System architecture. API: application program interface; REST: representational state transfer; SWRL: Semantic Web Rule Language.



Development Tools

The intelligent service engine is based on the SpringBoot+MyBatis+MySQL design architecture. The presentation form on the physician-oriented client is a Webpage and WeChat Mini Program (Tencent). The patient-oriented client is mainly the WeChat Mini Program.

Retrospective Evaluation

As the initial dates of patients' inclusion in the closed-loop diabetes management system for management were different, the date of each patient's inclusion in the management was used as the start of the study, whereby the management data for the following 3 months were recorded. If the management period was longer than 3 months, only the data for the first 3 months were collected for analysis. All patient data in this study were analyzed statistically using SPSS software (version 24.0; IBM Corp). The continuous assessment indexes were analyzed for significant differences in the management process using Student's *t* test (2-tailed), and the assessment indexes for the attainment rate were analyzed using the chi-square test to determine whether the data had significant differences.

The study involved 272 patients with diabetes who were included in the T2DMMS for typical management from January 2020 to August 2020 in the Ning-xia Medical University General Hospital Group. The patient inclusion criteria were as follows: (1) patients who signed informed consent form for the trial, (2) patient whose management time >3 months, and (3) patients whose BP and BG recorded >3 times. The patient exclusion criteria were as follows: (1) patients whose BG was not recorded after inclusion in management, (2) patients who were lost to follow-up, (3) patients with complex and severe comorbidities, and (4) patients with mental cognitive or physical dysfunction.

Ethics Approval

All the patients who entered the telehealth system signed an informed consent form. The nursing staff also signed informed consent forms. All the procedures were conducted according to the ethical guidelines for biomedical research involving humans at Ning-xia Medical University. Ethics approval was granted by the Ethics Committee for the Conduct of Human Research at the General Hospital of Ning-xia Medical University (NXMU-GH-2017-273).

Results

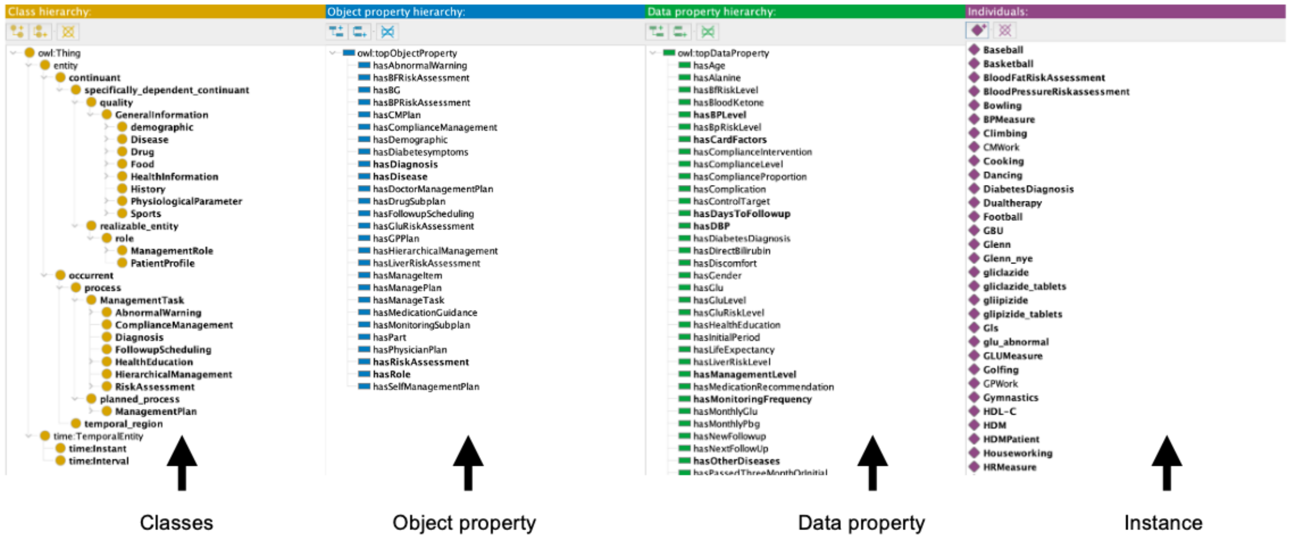
Ontology Construction and Evaluation

The DCPO was developed in the Web Ontology Language file format by the ontology editor Protégé 5.5.0. A snapshot of the DCPO is shown in Figure 4. The newest version of the DCPO contains 119 newly added classes, 28 object properties, 58 data properties, 81 individuals, and 426 axioms. In addition, 192 SWRL rules were newly added to implement the diabetes management care pathway. The SWRL rules are divided into 10 modules, with 22.9% (44/192) of rules based on clinical expert experience, 43.2% (83/192) based on medical guidelines, and 33.9% (65/192) based on both clinical expert experience and medical guidelines. Table 1 presents the specific distributions. The complete SWRL is presented in Multimedia Appendix 1.

The class diagram of the main core of the DCPO is shown in Figure 5. The DCPO mainly consists of the following 3 levels: level 0 includes several common top-level ontologies, which are regarded as standard to implement and improve the interoperability of other ontologies; level 1 consists of 5 terms that described the core concepts in the diabetes management pathway model; and level 2 is the detailed elements for each level-1 term.

First, the patient profile class is used to generate the instance of the patient condition through the object properties to connect to other main class instances. The management task represented the main content of the diabetes management pathway model, and it was synergistic and sequential, covering the whole process of diabetes diagnosis and treatment. In addition, the DCPO can generate management tasks with different contents for patients with different characteristics using instances and a set of SWRL rules. The content of the generated management tasks is converted into management plans.

Figure 4. The snapshot of Diabetes Care Pathway Ontology (DCPO) from the Protégé 5.5.0.



Class hierarchy:

```
owl:Thing
  entity
    continuant
      specifically_dependent_continuant
        quality
          GeneralInformation
            demographic
            Disease
            Drug
            Food
            HealthInformation
            History
            PhysiologicalParameter
            Sports
        realizable_entity
          role
            ManagementRole
            PatientProfile
    occurrent
      process
        ManagementTask
          AbnormalWarning
          ComplianceManagement
          Diagnosis
          FollowupScheduling
          HealthEducation
          HierarchicalManagement
          RiskAssessment
        planned_process
          ManagementPlan
      temporal_region
  time:TemporalEntity
    time:Instant
    time:Interval
```

Object property hierarchy:

```
owl:topObjectProperty
  hasAbnormalWarning
  hasBFRiskAssessment
  hasBG
  hasBPRiskAssessment
  hasCMPlan
  hasComplianceManagement
  hasDemographic
  hasDiabetesymptoms
  hasDiagnosis
  hasDisease
  hasDoctorManagementPlan
  hasDrugSubplan
  hasFollowupScheduling
  hasGluRiskAssessment
  hasGPPlan
  hasHierarchicalManagement
  hasLiverRiskAssessment
  hasManageItem
  hasManagePlan
  hasManageTask
  hasMedicationGuidance
  hasMonitoringSubplan
  hasPart
  hasPhysicianPlan
  hasRiskAssessment
  hasRole
  hasSelfManagementPlan
```

Data property hierarchy:

```
owl:topDataProperty
  hasAge
  hasAlanine
  hasBfRiskLevel
  hasBloodKetone
  hasBPLevel
  hasBpRiskLevel
  hasCardFactors
  hasComplianceIntervention
  hasComplianceLevel
  hasComplianceProportion
  hasComplication
  hasControlTarget
  hasDaysToFollowup
  hasDBP
  hasDiabetesDiagnosis
  hasDirectBilirubin
  hasDiscomfort
  hasGender
  hasGlu
  hasGluLevel
  hasGluRiskLevel
  hasHealthEducation
  hasInitialPeriod
  hasLifeExpectancy
  hasLiverRiskLevel
  hasManagementLevel
  hasMedicationRecommendation
  hasMonitoringFrequency
  hasMonthlyGlu
  hasMonthlyPbg
  hasNewFollowup
  hasNextFollowUp
  hasOtherDiseases
  hasPassedThreeMonthOrInitial
```

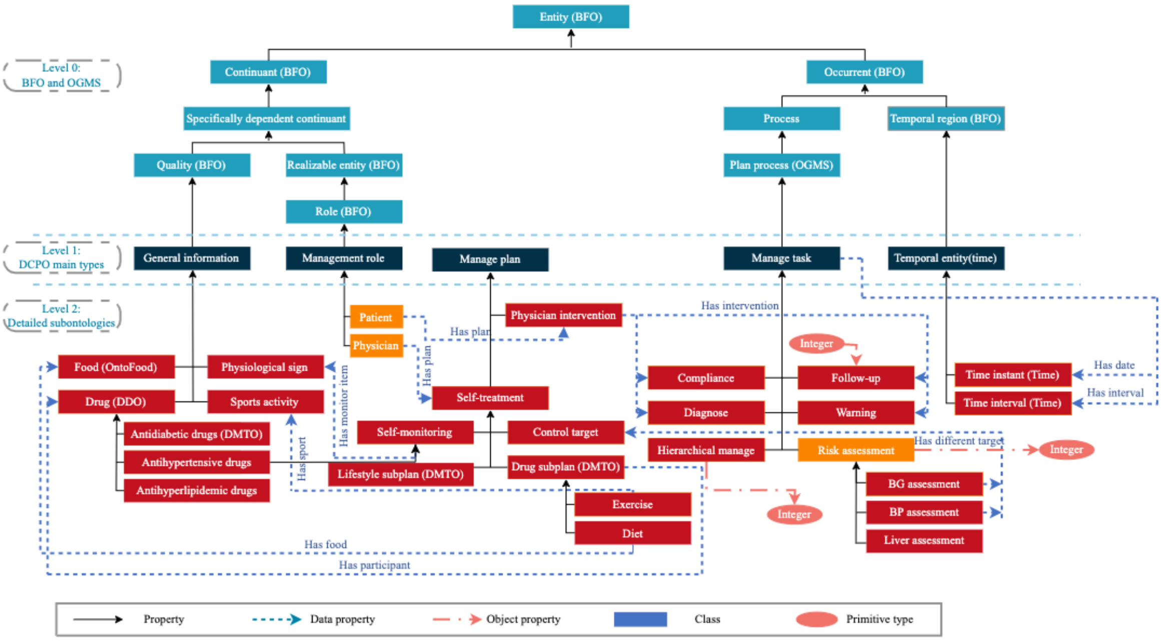
Individuals:

```
Baseball
Basketball
BloodFatRiskAssessment
BloodPressureRiskassessment
Bowling
BPMeasure
Climbing
CMWork
Cooking
Dancing
DiabetesDiagnosis
Dualtherapy
Football
GBU
Glenn
Glenn_nye
gliclazide
gliclazide_tablets
gliipizide
glipizide_tablets
Gls
glu_abnormal
GLUMeasure
Golfing
GPWork
Gymnastics
HDL-C
HDM
HDMPatient
Houseworking
HRMeasure
```

Table 1. Semantic Web Rule Language rule construction results.

Rules module	Derived from clinical expert experience (n=44), n (%)	Derived from medical guidelines (n=83), n (%)	Derived from both clinical expert experience and medical guidelines (n=65), n (%)
Diagnosis patterns	0 (0)	10 (12)	0 (0)
Risk assessment	11 (25)	12 (14)	28 (43)
Control objectives	0 (0)	6 (7)	3 (5)
Hierarchical management	18 (41)	0 (0)	0 (0)
Self-monitoring	0 (0)	0 (0)	18 (28)
Regular follow-up	0 (0)	0 (0)	8 (12)
Abnormal attention	15 (34)	0 (0)	0 (0)
Medication guidance	0 (0)	45 (54)	0 (0)
Lifestyle guidance	0 (0)	10 (12)	3 (5)
Compliance management	0 (0)	0 (0)	5 (8)

Figure 5. The class diagram of Diabetes Care Pathway Ontology (DCPO)'s main core. BFO: Basic Formal Ontology; BP: blood pressure; BG: blood glucose; DMTO: Diabetes Mellitus Treatment Ontology; OGMS: Ontology for General Medical Science.



System Interaction

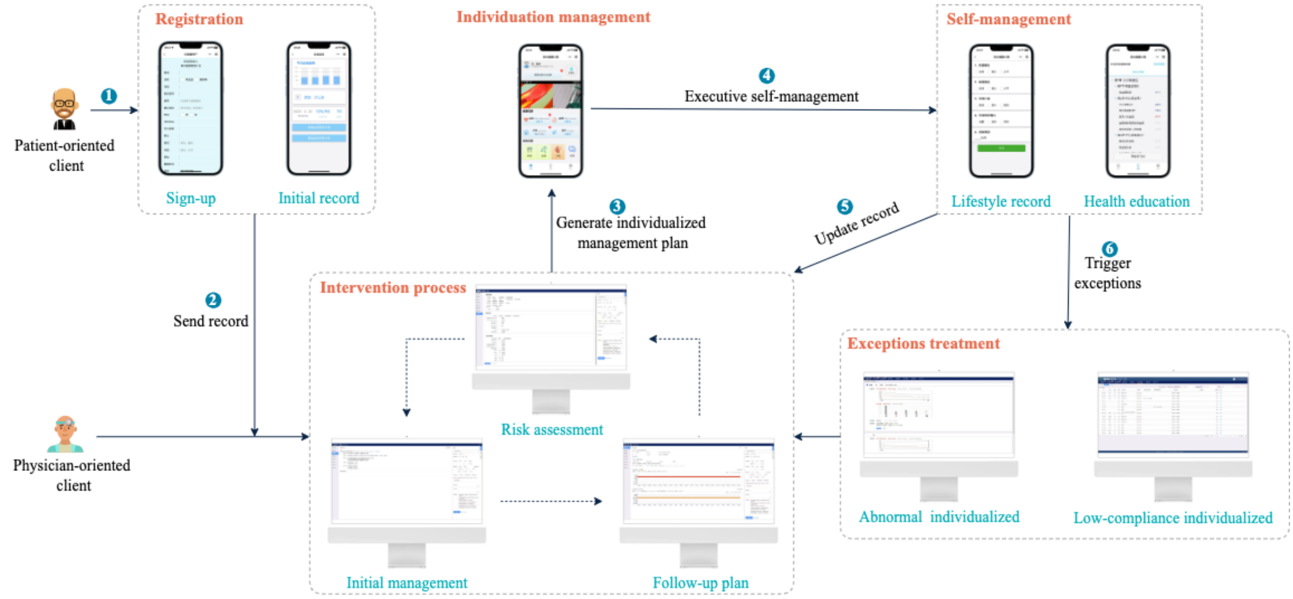
The physician-oriented client of the T2DMMS is operated by the medical service provider on the web page, and the patient-oriented client is displayed in the form of a WeChat applet. The 2 clients are coordinated by the diabetes management path in the intelligent service engine to efficiently complete the management work and improve patient compliance.

As shown in Figure 6, the physician-oriented client assists physicians in diagnosis and treatment tasks, including risk assessment, initial management, regular follow-up, abnormal attention, and compliance management for patients. A closed-loop path is formed between risk assessment, initial management, and follow-up, which can provide long-term management for patients. The physician only needs to operate in a certain order on the client terminal to manage the patient. In addition, the physician-oriented clients can modify the process, timing, and content of physician interventions to guide patients in the diabetes management path.

The patient-oriented client receives the personalized management plan analyzed by the intelligent service engine and displays the corresponding generated individualized management plan to the patients in the form of daily tasks, reminding the patient to follow the physician's instructions to complete the management plan. According to the prompts of daily tasks, patients can record their own health data every day through the WeChat applet; they can not only obtain real-time feedback and intervention from the engine but also receive feedback and intervention guidance from physicians.

Patient health data records mainly include BG levels, BP levels, discomfort symptom, weight, diet, and medication guidance records. The patient's health data records will be used as the data input for the intelligent service engine to evaluate and analyze the patient. When an abnormal attention occurs, the engine will provide real-time intervention guidance push and send the emergency situation of the patient to the physician to remind the physician to intervene.

Figure 6. Schematic of system interaction.



Retrospective Study Data

Overview

In this study, 272 patients with a mean age of 58.24 (SD 9.81) years were selected, of which 156 (57.4%) patients had diabetes only, and 116 (42.6%) patients had both hypertension and diabetes. According to the guidelines' recommendations, the initial BG levels were assessed in all 100% (272/272) of patients with diabetes. The BG levels were normal in 34.9% (95/272) of patients and abnormal in 65.1% (177/272) of patients. The initial BP was also assessed in 42.6% (116/272) of patients with hypertension and diabetes. The BP was normal in 30.2% (35/116) of patients and abnormal in 69.8% (81/116) of patients. Detailed patient data are presented in Table 2.

The management data of all patients were extracted from the T2DMMS during the 3-month management period. The data were analyzed from the following 2 perspectives: the investigation of physician work and analysis of patients' indicators.

Table 2. Experimental subject details (N=272).

Characteristics	Value
Sex, n (%)	
Male	147 (54)
Female	125 (46)
Disease type, n (%)	
Diabetes only	156 (57.4)
Both diabetes and hypertension	116 (42.6)
Initial assessment, n (%)	
Normal blood glucose level	95 (34.9)
Abnormal blood glucose level	177 (65.1)
Normal blood pressure[a]	35 (30.2)
Abnormal blood pressure[a]	81 (69.8)
Age, mean (SD) years	58.24 (9.81)
Special groups, n (%)	
Adolescent	7 (2.5)
Disabilities	2 (0.7)

[a] n=116, which is the number of patients with hypertension.

Investigating Physician Work

Figure 7 shows overall intervention records and patients' self-monitoring data through the engine and system following the diabetes management care pathway. Compared with the usual patient management approach based on guidelines only, T2DMMP added abnormal data attention and low-compliance management. As shown in Table 3, during the 3-month management cycle, the physician provided 904 follow-up visits in the usual patient management approach based on guidelines, and they followed up patients 939 times in T2DMMP management system, including 317 (33.8%) regular follow-up visits; 303 (32.3%) follow-up visits caused by abnormal attention that included 75 (24.8%) BG warning, 150 (49.5%) BP warning, 33 (10.9%) disorder warning, and 45 (14.9%) heart rate warning, and 319 (34%) follow-up visits for low-compliance management. The results showed that not only the physician's work contents were refined but also the number of follow-ups by the physician were changed in the T2DMMP-based management approach. The changes improved the attention and number of interventions for patients with poorer conditions.

The physician had not fully completed the intervention plan because some of the warnings were repeated, and the regular follow-up plan of patients who had hypertension and diabetes was incorporated.

The actual number of follow-up visits was 939 by physicians based on pathway prompts. Of the 272 patients, the number of follow-up visits was 855 (91.1%) in 242 (89%) patients whose initial BG level or BP was abnormal, and each patient was followed up 3.53 (SD 1.07) times on average. The number of follow-up visits was 84 (18.9%) in 30 (11%) patients whose initial BG level or BP was normal, and each patient was followed up with 2.8 (SD 0.59) times on average. Of the 855 follow-up visits owing to initial abnormalities in BG or BP, the real number of follow-up visits made by physicians to patients was classified into the following 3 levels: the number of follow-up visits >3 times was defined as high-intervention level; the number of follow-up visits equal to 3 times was defined as medium-intervention level; and the number of follow-up visits <3 times was defined as low-intervention level. Of the 242 patients with 855 interventions, there were 55 (22.7%) patients with high-intervention level, (398/855, 46.5%) and an average of 7.2 follow-ups per patient; there were 30 (12.4%) patients with medium-intervention level (154/855, 18%) and an average of 3.1 (SD 0.67) follow-ups per patient, 157 (64.9%) patients with low-intervention level (303/855, 35.4%), and an average of 1.9 (SD 1.25) follow-ups per patient. The guidelines state that physicians should follow-up at least once a month for patients with substandard BG, so physicians should follow-up each substandard patient 3 times within a 3-month management cycle. The above analysis results showed that a series of work can be dynamically adjusted according to the actual condition control of patients in the diabetes management pathway constructed in this study, including the follow-up schedule, prompting physicians to give different interventions and attention to patients with different management statuses. The adjusted changes can help patients whose condition control is poor obtain limited medical resource services more efficiently. Diabetes management pathways can improve physicians' working efficiency compared with the management style of managing patients based on guidelines only.

Figure 7. Overall management records of patients and physicians during the 3-month period.

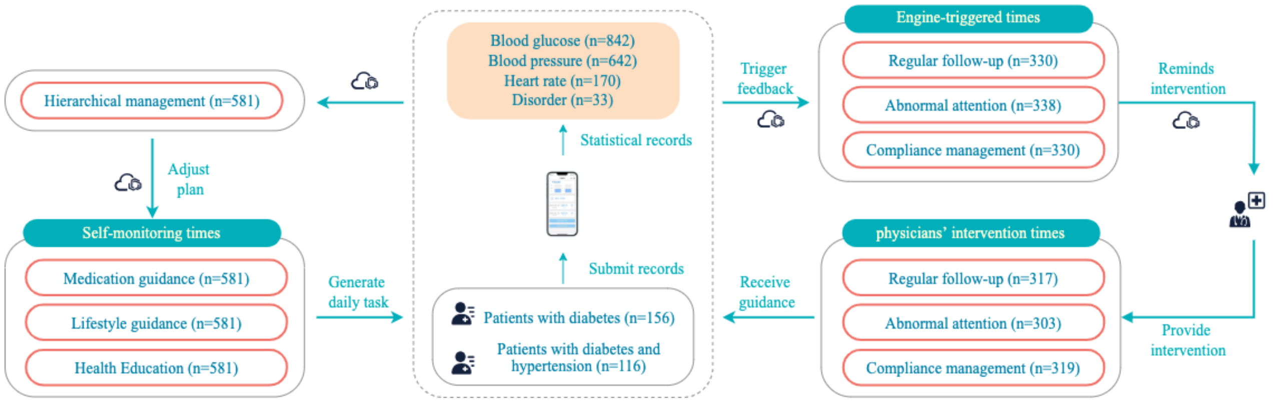

Hierarchical management (n=581)
Adjust plan
Self-monitoring times
Medication guidance (n=581)
Lifestyle guidance (n=581)
Health Education (n=581)
Generate daily task
Blood glucose (n=842)
Blood pressure (n=642)
Heart rate (n=170)
Disorder (n=33)
Statistical records
Submit records
Patients with diabetes (n=156)
Patients with diabetes and hypertension (n=116)
Trigger feedback
Receive guidance
Engine-triggered times
Regular follow-up (n=330)
Abnormal attention (n=338)
Compliance management (n=330)
Reminds intervention
physicians' intervention times
Regular follow-up (n=317)
Abnormal attention (n=303)
Compliance management (n=319)
Provide intervention

Table 3. Comparison of the number of follow-up visits to patients with different management models.

	Follow-up visits using management models	
	Based on diabetes management guidelines (n=904), n (%)	Based on the diabetes management care pathway (n=939), n (%)
Regular follow-up	904 (100)	317 (33.8)
Abnormal attention	0 (0)	303 (32.3)
Compliance management	0 (0)	319 (34)

Analysis of Patient Indicators

The core goal of this study was to achieve effective management of patients with diabetes and comprehensive management of multiple cardiovascular risk factors. Whether the patients' BG or BP decrease is the key metric to evaluate the application effect of the diabetes management model and DCPO designed in this study. The BG data records, which were uploaded by patients, were analyzed in all 100% (272/272) of patients during the 3 months, including FBG and postprandial BG (PBG). In addition, BP data (systolic BP [SBP] and diastolic BP [DBP]) were analyzed in 116 (42.6%) patients with hypertension and diabetes. All BG and BP data were obtained from the T2DMMS database. The FBG level of patients was recorded 2 hours before meals, and the PBG was recorded 2 hours after meals. The patients' BP was recorded between 8 AM and 10 AM. We analyzed the mean trend of all patients over the 3-month management period. Figures 8 and 9 show the patients' monthly average BG and BP records during the 3-month management period. Figure 8 showed the monthly mean FBG level decreased by 1.34 mmol/L (*P*=.003) and the monthly mean PBG level decreased by 2.63 mmol/L (*P*=.003) in all (272/272, 100%) patients with diabetes during the 3-month management period. Figure 9 shows that the monthly mean BP level also decreased significantly and finally reached a stable level in all (116/272, 42.6%) patients with both diabetes and hypertension during the 3-month management period in which the monthly mean SBP decreased by 11.84 mmHg (*P*=.02), and the monthly mean DBP decreased by 8.8 mmHg (*P*=.02).

We screened patients who received physician interventions during the management cycle owing to abnormal attention or low-compliance warnings from the trigger system. In patients with diabetes only, trends in FBG and PBG levels were analyzed. For patients with both hypertension and diabetes, trends in SBP and DBP were analyzed. As can be seen from the statistics in Figures 10 and 11, the mean monthly BG and BP values of the patients decreased significantly with respect to additional physician interventions. The monthly mean FBG level decreased by 2.45 mmol/L (*P*=.003) and the monthly mean PBG level decreased by 2.89 mmol/L (*P*=.003) in all (272/272, 100%) patients with diabetes during the 3-month management period; the monthly mean SBP decreased by 20.06 mmHg (*P*=.02) and the monthly mean DBP decreased by 17.37 mmHg (*P*=.02) in patients (116/272, 42.6%) with both hypertension and diabetes during the 3-month management period.

The analysis of the above results proves that the DCPO constructed in this study positively stabilizes the patient's condition by defining the management responsibilities corresponding to different management roles and adding the physician's intervention plan. It can also effectively help physicians manage patients with diabetes comprehensively.

Figure 8. The records of patient's blood glucose changes during the 3-month management period. BG: blood glucose; FBG: fasting blood glucose; PBG: postprandial blood glucose.

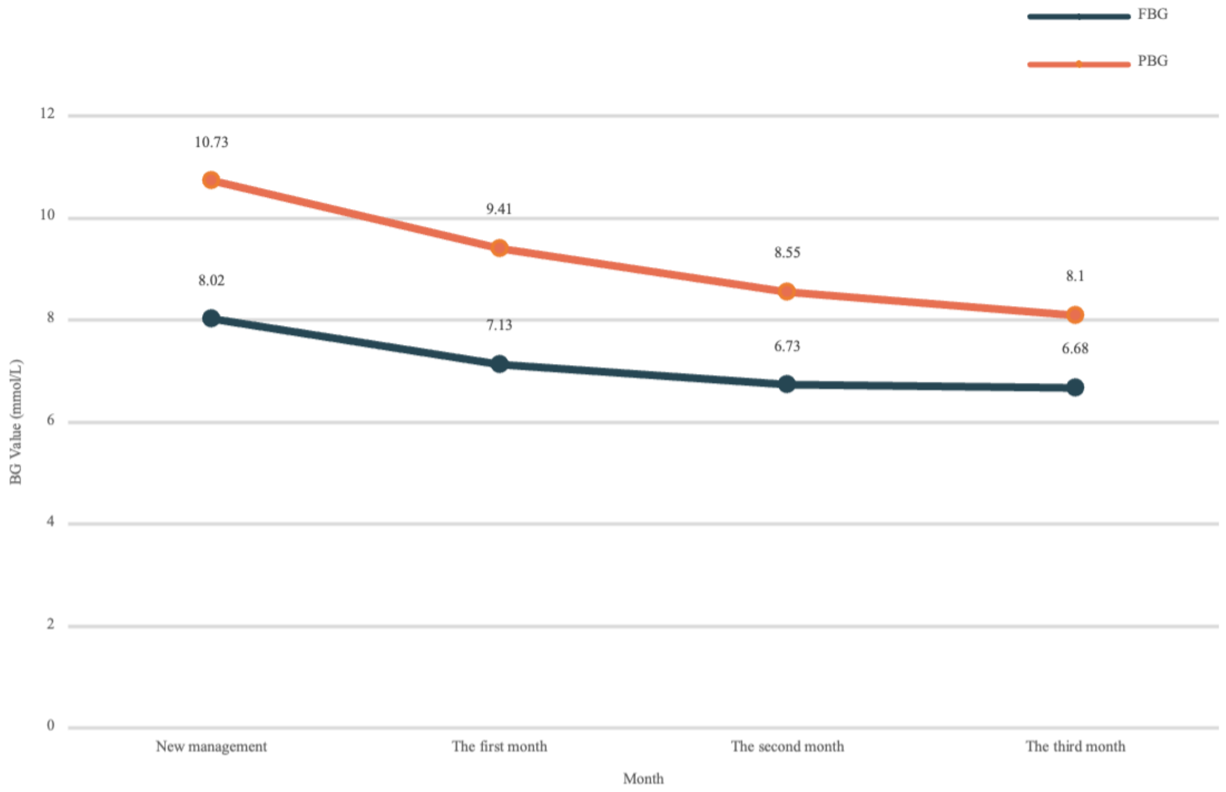


Figure 9. The records of patient's blood pressure changes during the 3-month management period. BP: blood pressure; DBP: diastolic blood pressure; SBP: systolic blood pressure.

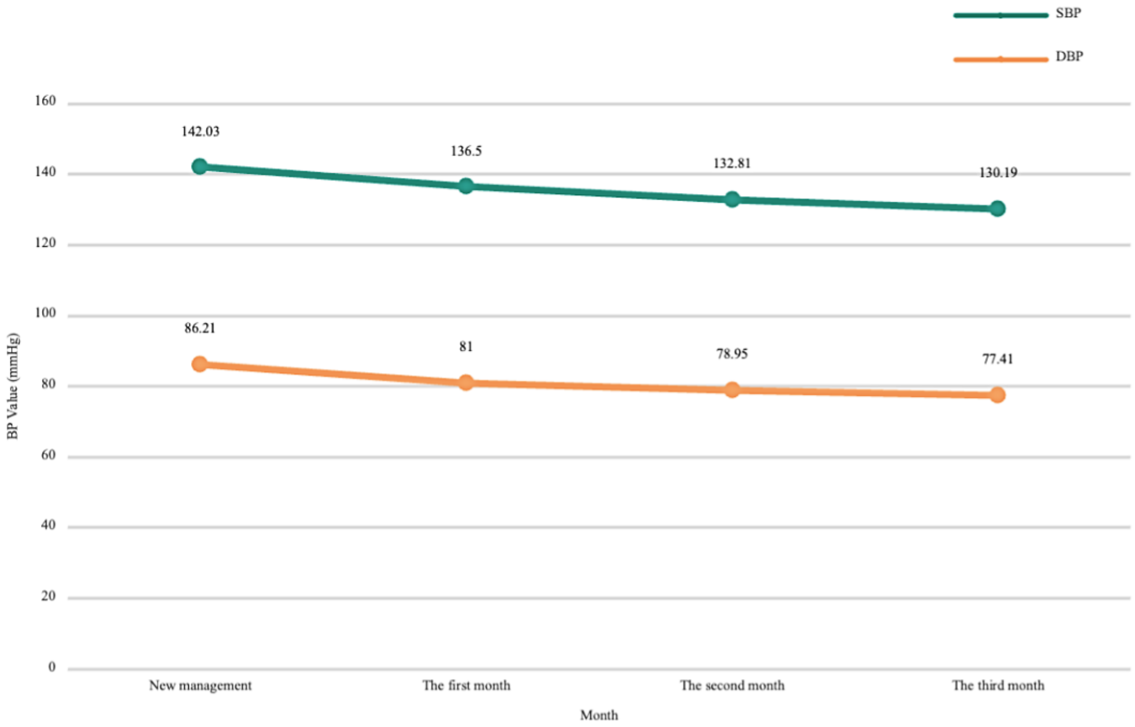


Figure 10. The records of intervened patient's blood glucose during the 3-month management period. BG: blood glucose; FBG: fasting blood glucose; PBG: postprandial blood glucose.

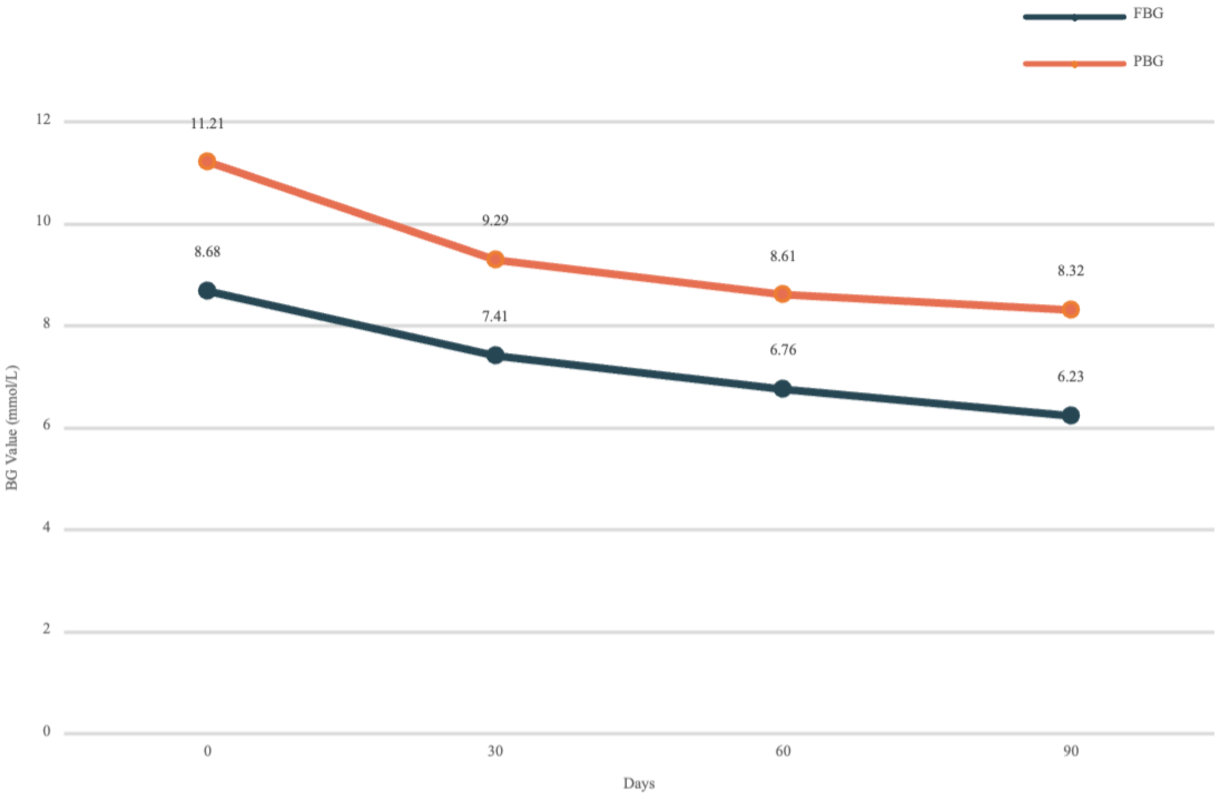
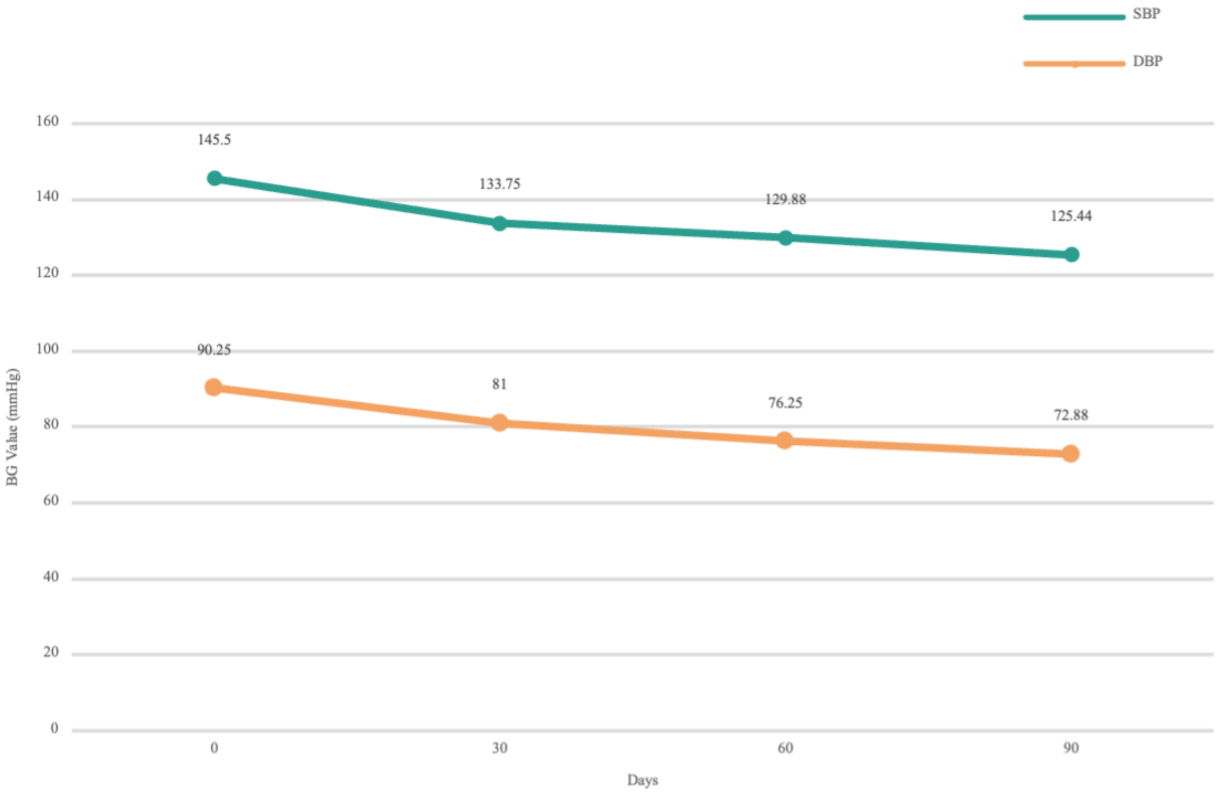


Figure 11. The records of intervened patient’s blood pressure during the 3-month management period. BG: blood glucose; DBP: diastolic blood pressure; SBP: systolic blood pressure.



Discussion

Principal Findings

In this study, we proposed and constructed the T2DMMS for monitoring and managing patients with type 2 diabetes. The T2DMMS could realize the comprehensive management and physician-patient interaction and intervention by defining different management roles corresponding to different management responsibilities proposed and defined in the diabetes management pathway. In addition, we implemented a telehealth system based on the T2DMMS and applied it to actual diabetes management. According to the retrospective analysis of the patient profile data records, the T2DMMS could realize the comprehensive evaluation and management of patients with diabetes and positively impact the comprehensive indicators of patients. Patients can self-manage according to the diabetes management care pathway and receive intervention guidance from physicians. The results indicated that the comprehensive indicators of patients certainly improved after the intervention.

To construct the T2DMMS, we first designed and implemented the diabetes management care pathway (T2DMMP) model. We summarized the diabetes management process into 9 core management tasks based on the CDMP concept and integrated these into a sequential and closed-loop diabetes management pathway model. Meanwhile, we defined the different management roles of physicians and patients, clarified their respective responsibilities, and realized an excellent physician-patient interaction intervention mechanism. Then, we referred to the top-down ontology construction method to construct the DCPO and realized the digitization of the T2DMMP expression process and construction results. By combining the top-level ontology and the existing standard ontology, we defined the main terms of the DCPO as subclasses of these top-level ontologies and finally built the DCPO with a 3-level hierarchy. To achieve more complete diabetes management, we formulated the SWRL rules in the standard coding method to achieve the timing between the management tasks in the T2DMMP.

From the system’s retrospective evaluation results, we identified several aspects and features. First, the monthly mean BG levels of all patients under regular management showed a downward trend. For patients with diabetes and hypertension, the monthly mean BP levels also decreased substantially. Second, the work content of physicians was defined by the diabetes management care pathway, including regular follow-up, abnormal warning, and compliance follow-up intervention. The system will remind physicians to implement interventional guidance. The data analysis showed that patients who had received additional interventions such as abnormal warning or compliance intervention had a more significant reduction in monthly mean BG and BP than patients who only received intervention guidance from regular follow-up.

Comparison With Previous Work

Table 4 provides a comparison between the DCPO and 3 other diabetes management ontologies based on the 10 dimensions. As shown in the table, all 3 compared ontologies were limited to patients’ comprehensive diabetes management, and the management plan mainly focused on patients, ignoring the responsibility of physicians and importance of physicians’ intervention. The DCPO can standardize and guide the comprehensive and complete management of patients with diabetes in primary care institutions in China. In addition, the DCPO performs well in terms of reusability, extensibility, and semantic interoperability.

Compared with previous work, our study is considered innovative in the following aspects: (a) For management, we innovatively introduced the concept of the pathway and then combined it with the information system to establish a standardized and executable management model. This system provides a one-stop platform for physicians and a fully functional terminal for patients. Physicians can perform almost all the daily work on the platform, and patients are able to monitor their BP and BG and receive management advice from their physicians. These 2 clients are connected by an engine that provides automatic decision support during the management process. To the best of our knowledge, such a fully functional and highly usable system for diabetes management in China has not yet been reported in the literature. (b) For the trial design, we proposed 2 perspective outcome measures as follows: physician work content analysis and control of comprehensive patient indicators. Controlling comprehensive patient indicators is an effective treatment for patients with diabetes.

Table 4. Comparison of Diabetes Care Pathway Ontology and other diabetes treatment ontologies.

Dimension	DCPO[a]	DMTO[b]	OMDP[c]	DDO[d]
Purpose	T2DM[e] treatment	T2DM treatment	T2DM treatment	T2DM treatment
Based on top-level ontology	Yes	Yes	Yes	Yes
Integration of the pathway tasks	Diagnosis, risk assessment, hierarchical management, regular follow-up, abnormal warning, medication guidance, lifestyle guidance, health education, and compliance management	Diagnosis and treatment by drug, food, exercise, and education	Prognosis, diagnosis, and treatment plan	Diagnosis
Treatment decision-making	Based on the risk results, management level, and past and current index of patient	Based on the patient's whole profile, including laboratory tests	Based on the patient's laboratory test results	Based on the patient's blood glucose and other laboratory tests
Modeling temporal semantics	Yes	Yes	No	No
Modeling comorbidities and complications	Hypertension, hyperlipidemia, hypoglycemia, and hyperglycemia	Diabetic nephropathy, retinopathy, and other complications	No clear definition	Diabetic ketoacidosis and coronary heart disease
Model of management roles	Defining both physician and patient	No management role defined	No management role defined	No management role defined
Application in telehealth environments	Supported by an intelligent service engine	No application	No application	No application
Drug modeling in the ontology	Antidiabetes drugs, hypertension drugs, and lipid drugs	Antidiabetes drugs, diabetes complication drugs, and the drug-drug interactions	Antidiabetes drugs and other drugs used for complications	Drugs affecting blood glucose
SWRL[f] rule sources in the ontology	Combining clinical experts and medical guidelines	No SWRL rules	Separate use of medical guidelines	No SWRL rules

[a] DCPO: Diabetes Care Pathway Ontology.

[b] DMTO: Diabetes Mellitus Treatment Ontology.

[c] OMDP: Ontology-Based Model for Diagnosis and Treatment of Diabetes Patients.

[d] DDO: Diabetes Diagnosis Ontology.

[e] T2DM: type 2 diabetes mellitus.

[f] SWRL: Semantic Web Rule Language.

Conclusions

In this study, a diabetes management pathway model was constructed, a diabetes management ontology for comprehensive diabetes management was developed to achieve physician-patient intervention, and a telehealth system based on this ontology was developed by summarizing the important process and core content of diabetes management. The DCPO was constructed on the basis of the general semantic definition of the standard top-level ontology, which contained 119 newly added classes, 28 object properties, 58 data properties, 81 individuals, 426 axioms, and 192 SWRL rules; this covered the entire process of diabetes management and managed multiple cardiovascular risk factors for patients with diabetes. Further research should be considered to deal with the ambiguity of medical semantics using fuzzy ontology; enhance the accuracy and reasoning ability of the system; introduce data-driven technology, considering the semantic interoperability with the electronic health record system; and obtain more clinical information from patient information to achieve a more personalized management plan.

Acknowledgments

This work was supported by the Key Research and Development Program of Zhejiang, China (grant 2021C03111) and the National Key Research and Development Program of China (grant 2020YFC2006405).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semantic Web Rule Language rules for Diabetes Care Pathway Ontology.
[DOCX File , 70 KB - medinform_v10i12e42664_app1.docx]

References

1. WHO: China's diabetes prevalence explodes. 21st Century Business Herald. 2016 Apr 7. URL: https://m.21jingji.com/article/20160407/herald/5294d7e89489f0c73dd7cb4bd204c431.html [accessed 2020-10-17]
2. Hu C, Jia W. Diabetes in China: epidemiology and genetic risk factors and their clinical utility in personalized medication. Diabetes 2018 Jan;67(1):3-11. [doi: 10.2337/dbi17-0013] [Medline: 29263166]
3. International Diabetes Federation. URL: https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html [accessed 2021-07-17]
4. Chatterjee S, Davies MJ. Current management of diabetes mellitus and future directions in care. Postgrad Med J 2015 Nov;91(1081):612-621. [doi: 10.1136/postgradmedj-2014-133200] [Medline: 26453594]
5. Reynolds R, Dennis S, Hasan I, Slewa J, Chen W, Tian D, et al. A systematic review of chronic disease management interventions in primary care. BMC Fam Pract 2018 Jan 09;19(1):11 [FREE Full text] [doi: 10.1186/s12875-017-0692-3] [Medline: 29316889]
6. Zwar N, Harris M, Griffiths R, Roland M, Dennis S, Powell Davies G, et al. A systematic review of chronic disease management. The Australian National University. 2006. URL: http://nceph.anu.edu.au/files/approved_3_zwar_pdf_16757.pdf [accessed 2020-03-17]
7. Wagner EH, Davis C, Schaefer J, Von Korff M, Austin B. A survey of leading chronic disease management programs: are they consistent with the literature? Manag Care Q 1999;7(3):56-66. [Medline: 10620960]
8. Harris MF, Zwar NA. Care of patients with chronic disease: the challenge for general practice. Med J Aust 2007 Jul 16;187(2):104-107. [doi: 10.5694/j.1326-5377.2007.tb01152.x] [Medline: 17635094]
9. Brannick B, Dagogo-Jack S. Prediabetes and cardiovascular disease: pathophysiology and interventions for prevention and risk reduction. Endocrinol Metab Clin North Am 2018 Mar;47(1):33-50 [FREE Full text] [doi: 10.1016/j.ecl.2017.10.001] [Medline: 29407055]
10. Marcolino MS, Oliveira JA, D'Agostino M, Ribeiro AL, Alkmim MB, Novillo-Ortiz D. The impact of mHealth interventions: systematic review of systematic reviews. JMIR Mhealth Uhealth 2018 Jan 17;6(1):e23 [FREE Full text] [doi: 10.2196/mhealth.8873] [Medline: 29343463]
11. Wyne K. Information technology for the treatment of diabetes: improving outcomes and controlling costs. J Manag Care Pharm 2008 Mar;14(2 Suppl):S12-S17. [Medline: 18331115]
12. Quinn CC, Shardell MD, Terrin ML, Barr EA, Ballew SH, Gruber-Baldini AL. Cluster-randomized trial of a mobile phone personalized behavioral intervention for blood glucose control. Diabetes Care 2011 Sep;34(9):1934-1942 [FREE Full text] [doi: 10.2337/dc11-0366] [Medline: 21788632]
13. El-Sappagh S, Ali F. DDO: a diabetes mellitus diagnosis ontology. Appl Inform 2016 Aug 25;3(1):5 [FREE Full text] [doi: 10.1186/s40535-016-0021-2]
14. El-Sappagh S, Kwak D, Ali F, Kwak KS. DMTO: a realistic ontology for standard diabetes mellitus treatment. J Biomed Semantics 2018 Feb 06;9(1):8 [FREE Full text] [doi: 10.1186/s13326-018-0176-y] [Medline: 29409535]
15. El-Sappagh S, Ali F, Hendawi A, Jang JH, Kwak KS. A mobile health monitoring-and-treatment system based on integration of the SSN sensor ontology and the HL7 FHIR standard. BMC Med Inform Decis Mak 2019 May 10;19(1):97 [FREE Full text] [doi: 10.1186/s12911-019-0806-z] [Medline: 31077222]
16. Sherimon PC, Krishnan R. OntoDiabetic: an ontology-based clinical decision support system for diabetic patients. Arab J Sci Eng 2016;41(3):1145-1160 [FREE Full text] [doi: 10.1007/s13369-015-1959-4]
17. Chen L, Lu D, Zhu M, Muzammal M, Samuel OW, Huang G, et al. OMDP: an ontology-based model for diagnosis and treatment of diabetes patients in remote healthcare systems. Int J Distrib Sens Netw 2019 May 03;15(5):155014771984711. [doi: 10.1177/1550147719847112]

18. Mengying J. Design and Application of Management Pathway for Hypertension. Zhejiang University. 2017. URL: https://cdmd.cnki.com.cn/Article/CDMD-10335-1017178226.htm [accessed 2022-12-08]
19. Kaining D. Research on Modeling and Implementation of Chronic Disease Management Pathway. Zhejiang University. 2018. URL: https://cdmd.cnki.com.cn/Article/CDMD-10335-1018172703.htm [accessed 2022-12-08]
20. Yan C. Evaluation Study on Implementation Effect of the Management Pathway for Hypertension. Zhejiang University. 2019. URL: https://cdmd.cnki.com.cn/Article/CDMD-10335-1019076042.htm [accessed 2022-12-08]
21. Wang Z, Li C, Huang W, Chen Y, Li Y, Huang L, et al. Effectiveness of a pathway-driven eHealth-based integrated care model (PEICM) for community-based hypertension management in China: study protocol for a randomized controlled trial. Trials 2021 Jan 22;22(1):81 [FREE Full text] [doi: 10.1186/s13063-021-05020-2] [Medline: 33482896]
22. Weiping J, Nong J, Jianping W. Guidelines for the prevention and control of type 2 diabetes in China. Chinese J Pract Intern Med 2018;38(4):292 [FREE Full text]
23. Chinese Diabetes Society, National Offic for Primary Diabetes Care. [National guidelines for the prevention and control of diabetes in primary care(2018)]. Zhonghua Nei Ke Za Zhi 2018 Dec 01;57(12):885-893. [doi: 10.3760/cma.j.issn.0578-1426.2018.12.003] [Medline: 30486556]
24. Garber AJ, Handelsman Y, Grunberger G, Einhorn D, Abrahamson MJ, Barzilay JI, et al. Consensus statement by the American Association of Clinical Endocrinologists and American College of Endocrinology on the comprehensive type 2 diabetes management algorithm - 2020 executive summary. Endocr Pract 2020 Jan;26(1):107-139. [doi: 10.4158/CS-2019-0472] [Medline: 32022600]
25. Davies MJ, D'Alessio DA, Fradkin J, Kernan WN, Mathieu C, Mingrone G, et al. Management of hyperglycemia in type 2 diabetes, 2018. A consensus report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetes Care 2018 Dec;41(12):2669-2701 [FREE Full text] [doi: 10.2337/dci18-0033] [Medline: 30291106]
26. Diabetes Canada Clinical Practice Guidelines Expert Committee. Diabetes Canada 2018 clinical practice guidelines for the prevention and management of diabetes in Canada. Can J Diabetes 2018;42(Suppl 1):S1-325.
27. Maedche A, Staab S. Ontology learning for the Semantic Web. IEEE Intell Syst 2001 Mar;16(2):72-79. [doi: 10.1109/5254.920602]
28. Kumar N, Khunger M, Gupta A, Garg N. A content analysis of smartphone-based applications for hypertension management. J Am Soc Hypertens 2015 Feb;9(2):130-136. [doi: 10.1016/j.jash.2014.12.001] [Medline: 25660364]
29. Pileggi SF, Fernandez-Llatas C. Semantic Interoperability: Issues, Solutions, and Challenges. Boca Raton, FL, USA: CRC Press; 2012.
30. Grenon P, Smith B, Goldberg L. Biodynamic ontology: applying BFO in the biomedical domain. Stud Health Technol Inform 2004;102:20-38. [Medline: 15853262]
31. Stenzhorn H, Beisswanger E, Schulz S. Towards a top-domain ontology for linking biomedical ontologies. Stud Health Technol Inform 2007;129(Pt 2):1225-1229. [Medline: 17911910]
32. Gu Q, Kai Y, Niu B, Tan L, Li L. BFO with information communicational system based on different topologies structure. In: Proceedings of the 9th International Conference on Intelligent Computing Theories and Technology. 2013 Presented at: ICIC '13; July 28-31, 2013; Nanning, China p. 633-640. [doi: 10.1007/978-3-642-39482-9_73]
33. Button K, van Deursen RW, Soldatova L, Spasić I. TRAK ontology: defining standard care for the rehabilitation of knee conditions. J Biomed Inform 2013 Aug;46(4):615-625 [FREE Full text] [doi: 10.1016/j.jbi.2013.04.009] [Medline: 23665300]
34. Wang Z, Huang H, Cui L, Chen J, An J, Duan H, et al. Using natural language processing techniques to provide personalized educational materials for chronic disease patients in China: development and assessment of a knowledge-based health recommender system. JMIR Med Inform 2020 Apr 23;8(4):e17642 [FREE Full text] [doi: 10.2196/17642] [Medline: 32324148]
35. Noy NF, McGuinness DL. Ontology Development 101: A Guide to Creating Your First Ontology. Stanford University. URL: https://protege.stanford.edu/publications/ontology_development/ontology101.pdf [accessed 2016-01-01]

Abbreviations

BG: blood glucose
BP: blood pressure
CDMP: Chronic Disease Management Pathway
DBP: diastolic blood pressure
DCPO: Diabetes Care Pathway Ontology
FBG: fasting blood glucose
PBG: postprandial blood glucose
SBP: systolic blood pressure
SWRL: Semantic Web Rule Language
T2DM: type 2 diabetes mellitus
T2DMMP: Type 2 Diabetes Mellitus Management Pathway

T2DMMS: Type 2 Diabetes Mellitus Management System

Edited by C Lovis; submitted 13.09.22; peer-reviewed by L Min, L Wang, D Hu; comments to author 09.11.22; revised version received 22.11.22; accepted 23.11.22; published 19.12.22.

Please cite as:
Fan Z, Cui L, Ye Y, Li S, Deng N
Telehealth System Based on the Ontology Design of a Diabetes Management Pathway Model in China: Development and Usability Study
JMIR Med Inform 2022;10(12):e42664
URL: https://medinform.jmir.org/2022/12/e42664
doi:10.2196/42664
PMID:36534448

Original Paper

Construction of Cohorts of Similar Patients From Automatic Extraction of Medical Concepts: Phenotype Extraction Study

Christel Gérardin[1], MA, MD; Arthur Mageau[2], MD; Arsène Mékinian[3], MD, PhD; Xavier Tannier[4], PhD; Fabrice Carrat[1,5], MD, PhD

[1]Institute Pierre Louis Epidemiology and Public Health, Institut National de la Santé et de la Recherche Médicale, Sorbonne Université, Paris, France

[2]Institut National de la Santé et de la Recherche Médicale, Unité Mixte de Recherche 1137 Infection Antimicrobials Modelling Evolution, Team Decision Sciences in Infectious Diseases, Université Paris Cité, Paris, France

[3]Service de Médecine Interne, Inflammation-Immunopathology-Biotherapy Department, Hôpital Saint-Antoine, Sorbonne Université, Assistance Publique–Hôpitaux de Paris, Paris, France

[4]Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances pour la e-Santé, Institut National de la Santé et de la Recherche Médicale, Université Sorbonne, Paris, France

[5]Public Health Department, Hopital Saint-Antoine, Assistance Publique–Hôpitaux de Paris, Paris, France

Corresponding Author:
Christel Gérardin, MA, MD
Institute Pierre Louis Epidemiology and Public Health
Institut National de la Santé et de la Recherche Médicale, Sorbonne Université
27 rue de Chaligny
Paris, 75012
France
Phone: 33 678148466
Email: christel.ducroz-gerardin@iplesp.upmc.fr

Abstract

Background: Reliable and interpretable automatic extraction of clinical phenotypes from large electronic medical record databases remains a challenge, especially in a language other than English.

Objective: We aimed to provide an automated end-to-end extraction of cohorts of similar patients from electronic health records for systemic diseases.

Methods: Our multistep algorithm includes a named-entity recognition step, a multilabel classification using medical subject headings ontology, and the computation of patient similarity. A selection of cohorts of similar patients on a priori annotated phenotypes was performed. Six phenotypes were selected for their clinical significance: P1, osteoporosis; P2, nephritis in systemic erythematosus lupus; P3, interstitial lung disease in systemic sclerosis; P4, lung infection; P5, obstetric antiphospholipid syndrome; and P6, Takayasu arteritis. We used a training set of 151 clinical notes and an independent validation set of 256 clinical notes, with annotated phenotypes, both extracted from the Assistance Publique-Hôpitaux de Paris data warehouse. We evaluated the precision of the 3 patients closest to the index patient for each phenotype with precision-at-3 and recall and average precision.

Results: For P1-P4, the precision-at-3 ranged from 0.85 (95% CI 0.75-0.95) to 0.99 (95% CI 0.98-1), the recall ranged from 0.53 (95% CI 0.50-0.55) to 0.83 (95% CI 0.81-0.84), and the average precision ranged from 0.58 (95% CI 0.54-0.62) to 0.88 (95% CI 0.85-0.90). P5-P6 phenotypes could not be analyzed due to the limited number of phenotypes.

Conclusions: Using a method close to clinical reasoning, we built a scalable and interpretable end-to-end algorithm for extracting cohorts of similar patients.

(JMIR Med Inform 2022;10(12):e42379) doi:10.2196/42379

KEYWORDS

natural language processing; similar patient cohort; phenotype; systemic disease; NLP; algorithm; automatic extraction; automated extraction; named entity; MeSH; medical subject heading; data extraction; text extraction

Introduction

Background

Extracting clinical phenotypes from large electronic health record (EHR) databases, also known as clinical data warehouses, is a key step for several medical applications from epidemiological research [1] to prognosis prediction [2,3] and therapeutic decision support [4,5]. Reliable automatic extraction of patient phenotypes from large EHR databases remains a challenge, especially in languages other than English [6]. The actual identification of patients' phenotypes is still largely done via the International Classification of Diseases, Ninth/Tenth Revision (ICD-9/ICD-10) code extraction, reading of clinical notes, or extraction of entities via regular expressions. However, as shown by Farzandipour et al [7] on more than 300 EHR ICD-10 codes, 22.7% presented errors in principal diagnosis codes, of which 33.3% were major errors. Benkhaial et al [8] also showed in a study of 200 patients, ICD allergy codes were present for 18 patients, while 51 had allergy information in a written note, indicating that only 35% of the allergies were correctly coded. These identification methods thus lack precision and require important human control.

With the improvement of natural language processing over the last 10 years, new language models such as Word2vec [9], GloVe [10], FastText [11] and, more recently, Bidirectional Encoder Representations from Transformers (BERT) [12] have allowed significant progress for various natural language processing tasks such as translation, question-answering, and named-entity recognition via an efficient word representation. Named-entity recognition corresponds to the extraction of certain classes of entities in a raw text. In the medical domain, it can be "signs and symptoms," "disorders," "chemicals and drugs," etc.

Many research teams have developed new algorithms based on these word models to allow automatic patient phenotyping. De Freitas et al [13] proposed Phe2vec, a data-driven, unsupervised disease phenotyping algorithm. In their study, disease phenotypes correspond to the word representation of ICD-10 core concepts (or seed concepts) and their closest neighbors. A patient's clinical history is summarized by aggregating all the word vector representations of the medical concepts. Mapping a patient to a disease is then done by computing a cosine distance between the patient with each disease phenotype. In their method, the medical concept extraction step from clinical notes is performed based on 1 ontology [14]. Ferté et al [15] also proposed an algorithm for automatic phenotyping of EHRs by using ICD-10 codes and a dictionary-based entity recognition tool to extract interesting terms from clinical notes. Extracted terms were then mapped to their unified medical language system concept unique identifier as a feature for classification to provide an interpretable parametric predictor. Their work showed particularly interesting results for chronic conditions.

In this work, we extracted similar patients by focusing on 4 systemic diseases as a proof of concept: systemic lupus erythematosus (SLE), systemic sclerosis, antiphospholipid syndrome (APS), and Takayasu arteritis. SLE is an autoimmune disease that can affect a large number of organs: the skin (specific malar rash, photosensitivity, etc), kidneys (nephrotic syndrome and glomerular nephropathy), joints (most often without deformation), brain (with neuropsychiatric forms), etc. It is a rare disease that affects 41 in 100,000 people in France [16], and 9 women for 1 man in generally young (18-30 years old) adults. Systemic sclerosis can also involve various organs: the skin (sclerosis leading to significant functional impotence), the lungs (interstitial lung disease [ILD], fibrosis, and hypertension), the digestive system (reflux and chronic intestinal obstruction), etc. Its frequency is 1/5000 in France, and it preferentially affects women (4 women for 1 man) aged between 40 and 50 years. APS is a disease that causes venous and arterial thrombosis as well as obstetrical complications. Approximately 20%-30% of patients with lupus develop APS. Its frequency is approximately 1 in 12,000 [16]. Takayasu arteritis is an inflammatory disease that affects large vessels in young people. It is a very rare disease affecting 1.2 to 2.6 cases/million/year in France. It affects 4.8 women for 1 man between 20 and 40 years of age [17]. These 4 diseases were chosen because of their large spectrum of signs and symptoms and their similarity (especially for lupus and APS in terms of apparition frequency and APS and Takayasu for their arterial manifestations).

Goal of This Study

In this study, we aimed to develop an automated end-to-end extraction of similar patient cohorts from electronic medical records. Specifically, we place ourselves in the following use case: we have a patient to treat with clinical information in a text document (mentioned as index patient in this paper), and we automatically search for the set of patients with similar symptoms and diseases mentioned in their hospitalization reports. To evaluate our method, we extracted cohorts of similar patients from index patients with certain phenotypes described in their textual reports, arbitrarily selected, and manually annotated by a clinician. Our main contribution in this paper is the development of an algorithm for the automatic construction of similar patient cohorts by a method close to clinical reasoning, as we argue in the Discussion section.

Methods

Algorithm Steps

In this section, we detail the main steps of our algorithm. Similarity is defined here as a patient with identical or closely related signs, symptoms, and disorders. The key steps for extracting these events from the text are a named-entity recognition step to extract medical concepts, a multilabel classification on each extracted term, and an average distance computation on an appropriate representation of all the terms on each label. We validated our interpatient distance by clustering 6 a priori defined phenotypes of interest: osteoporosis, nephritis in SLE, ILD in systemic sclerosis, lung infection, obstetric APS, and Takayasu arteritis. With the same interpatient distance, we then constructed similarity cohorts from index patients for each of these phenotypes.

Overview of the Algorithm

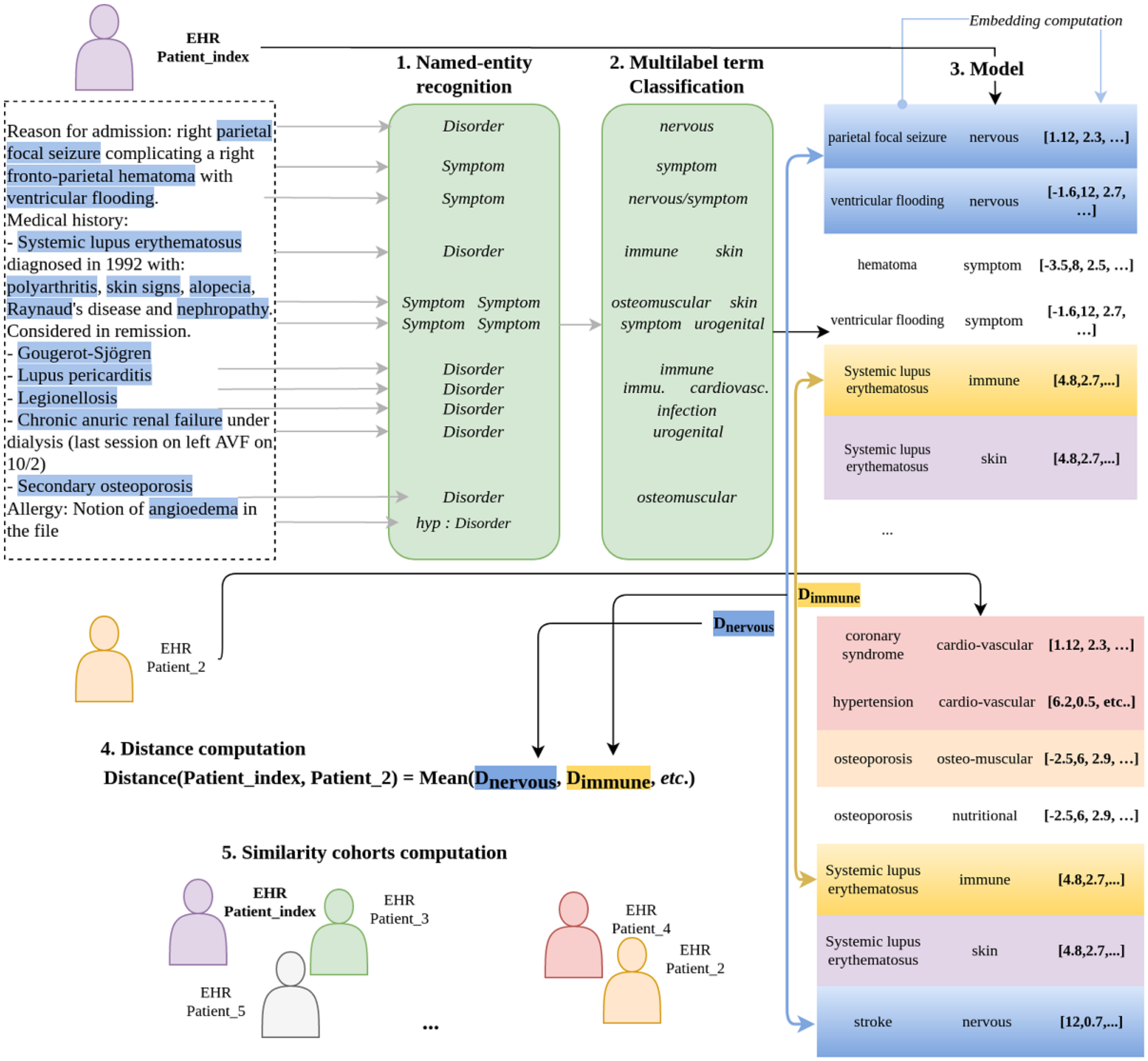
For readability, in the remainder of this paper, we use the term "patient" to refer to the "hospitalization report related to the patient."

The main steps of the algorithms are shown in Figure 1, considering an index patient:

1. Symptoms and diseases were extracted from a raw text while filtering out all negated, hypothetical, and belonging to family terms.
2. All extracted terms were classified into broad organ categories, that is, cardiovascular, immune, ophthalmologic, digestive, etc, by a multilabel classification step using our previously developed algorithm [18].
3. A vector (embedding) representation for all extracted terms was obtained leading to the index patient representation.
4. From this representation for other patients, the distance for each label of the index patient to the other patients was computed. Then, the average of the distances of all the labels was determined.
5. A cohort of similar patients was built from the patients closest to the index patient for each annotated phenotype.

We will refer to this patient's hospitalization report (Figure 1, index_patient) as a running example throughout the steps described below.

Figure 1. Overview of the algorithm to obtain a representation of the patients' electronic health records and to compute a distance from other patients' electronic health records. First, a named-entity recognition step is performed on a patient's electronic health record (to extract symptoms and disorders and filter all negated, hypothetical, and someone else's terms). Second, a multilabel classification step is performed for each extracted term to allow more clinical interpretation. Third, this leads to an electronic health record model containing all the extracted terms with their respective labels and embedding representations (last column of the model). Fourth, this allows a distance computation on each of the 22 labels (Dnervous corresponds to the distance between embeddings of all terms labelled nervous, Dimmune on the immune label, etc). Fifth, a similarity cohort computation is performed. EHR: Electronic Health Record.


EHR
Patient_index
Embedding computation
1. Named-entity recognition
2. Multilabel term Classification
3. Model
Reason for admission: right parietal focal seizure complicating a right fronto-parietal hematoma with ventricular flooding.
Medical history:
- Systemic lupus erythematosus diagnosed in 1992 with: polyarthritis, skin signs, alopecia, Raynaud's disease and nephropathy.
Considered in remission.
- Gougerot-Sjögren
- Lupus pericarditis
- Legionellosis
- Chronic anuric renal failure under dialysis (last session on left AVF on 10/2)
- Secondary osteoporosis
Allergy: Notion of angioedema in the file
Disorder
Symptom
Symptom
Disorder
Symptom Symptom
Symptom Symptom
Disorder
Disorder
Disorder
Disorder
Disorder
hyp : Disorder
nervous
symptom
nervous/symptom
immune skin
osteomuscular skin
symptom urogenital
immune
immu. cardiovasc.
infection
urogenital
osteomuscular
parietal focal seizure nervous [1.12, 2.3, …]
ventricular flooding nervous [-1.6,12, 2.7, …]
hematoma symptom [-3.5,8, 2.5, …]
ventricular flooding symptom [-1.6,12, 2.7, …]
Systemic lupus erythematosus immune [4.8,2.7,...]
Systemic lupus erythematosus skin [4.8,2.7,...]
...
EHR
Patient_2
Dimmune
Dnervous
coronary syndrome cardio-vascular [1.12, 2.3, …]
hypertension cardio-vascular [6.2,0.5, etc..]
osteoporosis osteo-muscular [-2.5,6, 2.9, …]
osteoporosis nutritional [-2.5,6, 2.9, …]
Systemic lupus erythematosus immune [4.8,2.7,...]
Systemic lupus erythematosus skin [4.8,2.7,...]
stroke nervous [12,0.7,...]
4. Distance computation
Distance(Patient_index, Patient_2) = Mean(Dnervous, Dimmune, etc.)
5. Similarity cohorts computation
EHR
Patient_index
EHR
Patient_3
EHR
Patient_5
...
EHR
Patient_4
EHR
Patient_2

Data Sets and Annotation Rules

The data set of this study was obtained from the Assistance Publique-Hôpitaux de Paris (AP-HP) data warehouse. Patients were informed that their EHR information could be reused after an anonymization process, and those who objected to the reuse of their data were excluded. All methods were carried out in accordance with relevant guidelines (reference methodology MR-004 of the Commission Nationale de l'Informatique et des Libertés [19]).

The data set contained all hospitalization reports, consultation reports, test results, prescriptions, etc of all patients older than 15 years with lupus, scleroderma, APS, and Takayasu arteritis who made at least one visit to AP-HP hospitals since 2017. The data set constitutes a set of 2 million pseudonymized clinical records. It was extracted using only the ICD-10 codes of the principal diagnosis for lupus (M320, M321, M328, M329, L930, L931, corresponding to 5176 patients), systemic sclerosis (M340, M341, M348, M349, corresponding to 2833 patients), APS (D686 corresponding to 1250 patients), and Takayasu arteritis (M314, corresponding to 287 patients).

An internist physician annotated a training subset of 151 clinical notes (40 lupus, 35 APS, 37 systemic sclerosis, and 39 Takayasu) with symptoms or disorders by using specific attributes "negated," "hypothetical," and "belonging to family" when relevant. Guided by a clinical logic, we chose not only to annotate the negated terms as negation (eg, no fever, no diabetes) but also all the physiological descriptions (eg, peripheral pulse present, vesicular breath sounds present and symmetrical, regular heart sounds). All of these physiological findings were annotated as negative, because in clinical reasoning, we focus primarily on pathological signs. We adopted this approach also because the language models we use are able to capture both the syntactic and semantic levels of language. The medical subject heading (MeSH) category C [20] head chapters (eg, cardiovascular, immune, digestive) were also annotated at the entity level. This annotated data set was used to train both the named-entity recognition step with the symptoms and disorders labels and the multilabel classification step with MeSH [20] category C chapter head labels. Another test set of 256 hospitalization reports was annotated with one or more of the 6 phenotypes of interest, that is, osteoporosis, nephritis in SLE, ILD in systemic sclerosis, lung infection, obstetric APS, and Takayasu arteritis by another internist physician with no common patients between the training and testing data sets.

The annotation rules were defined before starting. First, a phenotype was only positively annotated if it was explicitly written, and no interpretation was made of signs and test results to guess the phenotype. For example, for osteoporosis, neither bone mineral density nor the number of vertebral fractures was interpreted, and the only terms retained positively were osteoporosis and corticosteroid-induced osteoporosis. Detailed examples can be found in Figure S1 of Multimedia Appendix 1. We selected these phenotypes for their clinical significance both in the 4 pathologies of interest studied and globally in terms of osteoporosis and lung infection phenotypes. These phenotypes were selected as an example, but our algorithm can be generalized to handle very different phenotypes.

Word Representations

Two word representation models were used for this work. First, a French BERT model [12], camemBERT, trained by Martin et al [21] on a wide variety of French documents was used for the named-entity recognition and multilabel classification steps. Second, a FastText model developed by Bojanowski et al [11] was used for the patient model to calculate the interpatient distance. Both methods convert words into vectors of real numbers (called embeddings). BERT produces embeddings that take into account the context (other words in the phase), while FastText produces fixed embeddings (a word corresponds to a vector independently of the surrounding text). For our study, we had 2 million documents of all types (consultation records, hospitalization records, discharge summaries, etc), which correspond to a volume of 5 gigabytes of text. These data allowed us to train the FastText model from scratch. The camemBERT model was too large to train from scratch, but we fine-tuned it on our data, that is, we retrained its final layers. As a result, it was able to learn a context-appropriate vector representation (particularly effective for the feature extraction step 1); nevertheless, its initial vocabulary did not contain all the medical concepts, unlike the FastText model, which we used for the patient representation for the interpatient distance calculation.

Named-entity Recognition

This first step enables us to extract positive symptoms (pathologic signs) and disorders, filtering all terms corresponding to hypothetical, negated, and family-related elements. For instance, in Figure 1 (index_patient), the extracted terms were "parietal focal status epilepticus," "frontoparietal hematoma," and "systemic lupus erythematosus," whereas "angioedema" was not kept since it was only hypothetical. The algorithm used for this first step is based on an encoder (with BERT layers) and a bidirectional long short-term memory decoder. This neural named-entity recognition model, described in [18], obtains an exact F-measurement of 0.931 on the English CoNLL data set [22], using the BERT-large embeddings [12], and 0.784 on GENIA [23], using the BioBERT-large model [24].

Multilabel Classification

To improve clinical interpretability and to analyze patients along several medical dimensions (ie, labels), we chose to perform a multilabel classification of all the terms. The corresponding class is all the MeSH-C head chapters, corresponding to 22 medical fields: infections, ophthalmologic, stomatology, cardiovascular, digestive, respiratory, nervous, etc. A BERT model for the sequence classification was used and trained on all annotated entities and all MeSH terms and their synonyms. Synonyms of MeSH terms were obtained by extracting all the French terms sharing the same code unique identifier in the unified medical language system defined by their authors as a "set of files and software that brings together many health and biomedical vocabularies to enable interoperability between computer systems" [25]. This multilabel classifier has been described in our previous study and evaluated on an external challenge with an F1-score from 0.809 to 0.811 depending on the model used [18]. For instance, for our index_patient in Figure 1, parietal focal status epilepticus is labelled as nervous, and systemic lupus erythematosus is labelled as immune and skin.

Distance Computation

We used FastText to obtain an embedding representation of each extracted term. With all the patients represented as a list of embeddings for each label, the distance between the patients can be computed based on one particular label of interest (cardiovascular, urogenital, etc), or several, or all. However, 2 patient records may contain different numbers of terms (ie, vectors) per label. For example, index_patient on Figure 1 only presents 1 term on the cardiovascular label (lupus pericarditis), whereas patient_2 may present many cardiovascular terms such as coronary syndrome, hypertension, and stroke.

Following Kusner et al’s [26] idea, we decided to use the earth mover’s distance, a distance that minimizes the cost to be paid to transform one distribution into another. We compute this distance for each label. In our case, the distributions correspond to the set of terms per label, and each term corresponds to a point. The size of the point corresponds to the frequency of occurrence of the term, and the distance between the points corresponds to the cosine distance between the FastText embeddings of the terms. In our example, the immune label for index_patient is made of the terms SLE (1 occurrence), Raynaud (1 occurrence), Gougerot-Sjögren (1 occurrence), and lupus pericarditis (1 occurrence).

Having a distance, we are now able to compare patients’ clinical notes on each label (provided that the patient’s record has at least one term present for this label) or globally. To compare 2 patients globally, we summed the earth mover’s distances of the 2 patients across each label and weighted them with the corresponding number of terms for each label. Equations (1) and (2) below specify the weighting term, where HR_1 and HR_2 denote 2 different hospitalization reports, and EMD() denotes the earth mover’s distance between the 2 notes for a specific label i.

$D(HR_1, HR_2) = (1/nlabels)*\Sigma\ (\lambda_i\ EMD(HR_1(label_i), HR_2\ (label_i))$ **(1)**

with $\lambda_i = (nHR_1(label_i) + nHR_2(label_j))\ /\ (nHR_1 + nHR_2)$ **(2)**

where $HR_j(label_j)$ is the list of terms from HR_i involving $label_j$ and nHR is the number of terms in the term subset HR.

Evaluation

We evaluate our approach with the 6 use cases described earlier, each being associated with specific MeSH-C labels. For example, to obtain similar patients for the osteoporosis phenotype (labelled musculoskeletal and nutritional according to MeSH classification), we computed the earth mover’s distance of the hospitalization reports only on these 2 labels. Similarly, for ILD in systemic sclerosis, we focused on the respiratory and immune labels. For lung infection, we focused on the respiratory and infections labels, and so on. However, our algorithms can be applied to any new use case and to any set of MeSH-C labels.

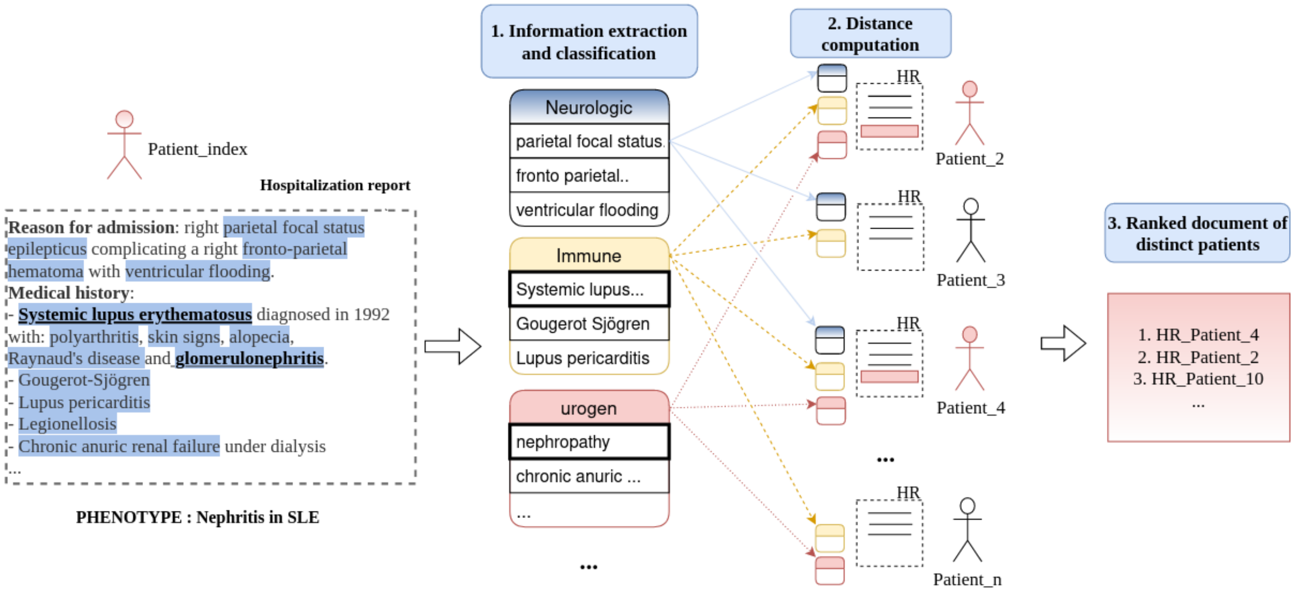
Clustering

To visualize our results and to confirm the relevance of our approaches, we performed an unsupervised hierarchical clustering of all patients in the training data set on each label and globally, checking if patients with similar phenotypes belonged to the same clusters. We used agglomerative hierarchical clustering (each hospitalization report is initialized as a singleton cluster, and clusters are merged two-by-two) with Ward’s criterion, which minimizes the variance of the clusters. The same method was used for our 6 use case phenotypes. We used the SciPy library [27].

Selection of a Cohort of Similar Patients From an Index Patient

We approach the problem of building a cohort of similar patients as an information retrieval problem, where the patient’s document (index patient) is a query. We then evaluate the ability of the system to return a ranked list of documents, with the most relevant/similar at the top of the list. Figure 2 gives an overview of this selection on the example of a patient with the phenotype “Nephritis in SLE.” We evaluate the precision-at-k (percentage of correct phenotype prediction in the first k closest documents of distinct patients), the recall (percentage of all correct phenotypes that are selected in the first n closest patients, n being the number of patients in each phenotype), and the average precision. The average precision computes the average value of the precision for recall values over 0 to 1. It considers the order in which the patients are selected and corresponds to an estimate of the area under the precision-recall curve. For each phenotype, each patient from the test set is chosen in turn as an index patient, and the final results are an average over all patients. Confidence intervals were calculated using the normal distribution approximation.

Figure 2. Example of document selection for the phenotype "Nephritis in systemic lupus erythematosus." First, from the clinical observation of the index patient, symptoms and diseases are extracted and classified according to medical subject heading-C chapter headings (step 1). Then, the distance is calculated on the UroGen and immune classes (specifically for this phenotype, step 2). Finally, the closest documents are those with the same written phenotype, corresponding to the patients in red in the figure, leading to a ranked list of the closest documents of distinct patients (step 3). SLE: Systemic lupus erythematosus; HR: Hospitalization report.


Patient_index
Hospitalization report
Reason for admission: right parietal focal status epilepticus complicating a right fronto-parietal hematoma with ventricular flooding.
Medical history:
- Systemic lupus erythematosus diagnosed in 1992
with: polyarthritis, skin signs, alopecia,
Raynaud's disease and glomerulonephritis.
- Gougerot-Sjögren
- Lupus pericarditis
- Legionellosis
- Chronic anuric renal failure under dialysis
...
PHENOTYPE : Nephritis in SLE
1. Information extraction and classification
Neurologic
parietal focal status
fronto parietal..
ventricular flooding
Immune
Systemic lupus...
Gougerot Sjögren
Lupus pericarditis
urogen
nephropathy
chronic anuric ...
...
...
2. Distance computation
HR
Patient_2
HR
Patient_3
HR
Patient_4
...
HR
Patient_n
3. Ranked document of distinct patients
1. HR_Patient_4
2. HR_Patient_2
3. HR_Patient_10
...

Visualization

A distance-based search result was also constructed to select the most similar patient to an index patient, with clickable labels where clinicians can choose any labels of interest they want to select (as in our phenotype examples). This search result returns the most similar patients on the selected labels in the descending order of similarity. A demonstration can be found in this following link [28], with 4 use cases with word clouds of medical terms enabling the similarity decision. All our codes are available on GitHub [29].

Ethics Approval

The results shown in this study are derived from the analysis of the AP-HP data warehouse. This study and its experimental protocol was approved by the AP-HP Scientific and Ethical Committee (IRB00011591 decision CSE 20-0093). All methods were carried out in accordance with relevant guidelines (reference methodology MR-004 of the Commission Nationale de l'Informatique et des Libertés [19]). All medical records have been pseudonymized. Patients are informed by the AP-HP data warehouse that the data are pseudonymized and that they can object to their sharing. Their consent was therefore collected prior to our study.

Results

Clustering

The results of the unsupervised hierarchical clustering on our training data set of 151 EHRs are shown in Figure 3, Figure 4, and Figure 5. Each cluster is enhanced with its corresponding word cloud (highlighting the frequencies of occurrence of terms within each cluster). Interestingly, on the immune label (Figure 3), we were able to properly separate patients with scleroderma (left, orange cluster) from patients with lupus or lupus with APS (green clusters). As mentioned earlier, 30% of APS is secondary to systemic lupus, and indeed, several patients with APS in our data set also had lupus. Similarly, on the digestive label (Figure 4), we were able to separate upper digestive manifestations (left cluster) from liver issues (left clusters). With regard to the global clustering (using equations 1 and 2 above), we obtained 4 different clusters, as shown in Figure 5. Scleroderma is clustered separately with forms of cutaneous lupus (right, purple cluster) from lupus with thromboembolic manifestations and APS (middle, red cluster) from Takayasu (second left, green cluster). Interestingly, scleroderma with pulmonary arterial hypertension (left, little orange cluster) is close to the Takayasu cluster with arterial complications. The test set included 100 patients with lupus, 87 with scleroderma, 51 with APS, and 18 with Takayasu arteritis. Only 4 Takayasu stroke were labelled and 7 obstetrical APS, which did not allow us to perform clustering or other performance computations. The clustering results for phenotypes osteoporosis and lung infection with ground truth labelled documents are shown as examples in Figure 6 and Figure 7, respectively.

Figure 3. Unsupervised hierarchical clustering based on electronic health record earth mover's distance on the “immune” label. Word clouds of electronic health records words are plotted on each respective cluster. Interestingly, patients with systemic scleroderma all belong to the same cluster (orange). Only patients who were labelled “immune” are clustered; we thus represent 129 patients out of 151.

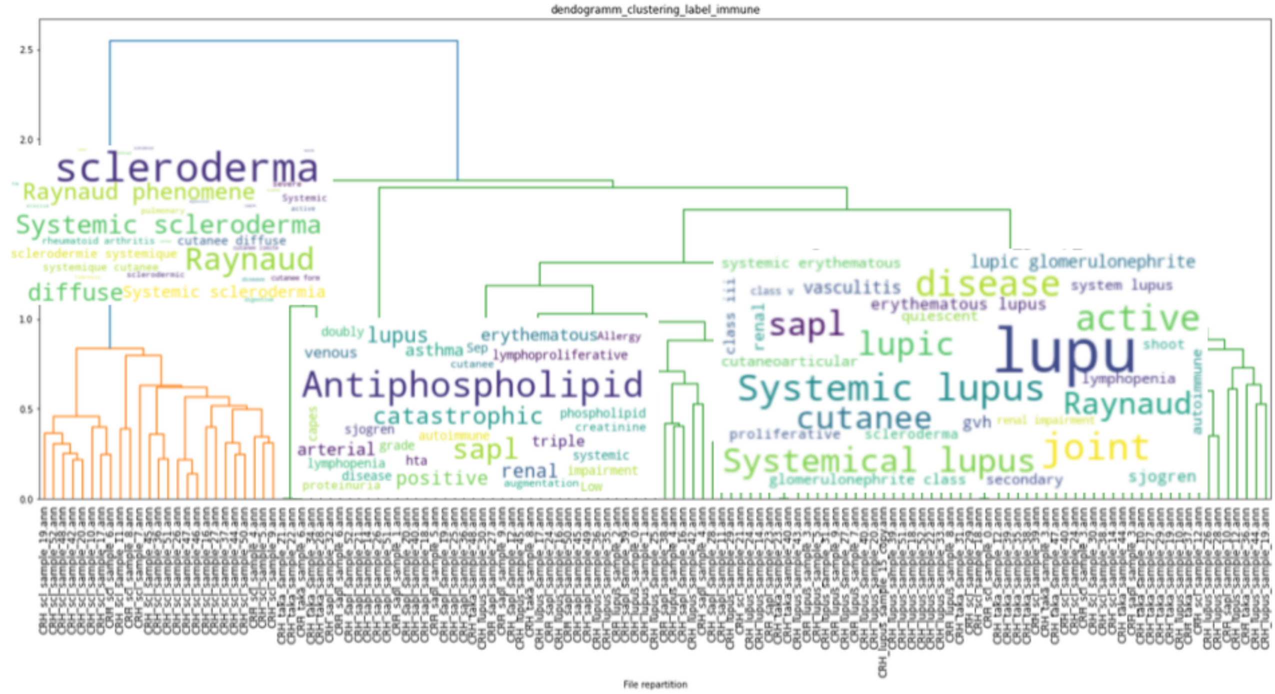


Figure 4. Unsupervised hierarchical clustering based on earth mover's distance of electronic health records on the label “digestive.” The word cloud of the electronic health records is shown on each respective cluster. Interestingly, the left cluster reports upper digestive manifestations (oesophagitis, gastroesophageal reflux or RGO in French), and the rightmost cluster represents patients with liver diseases (brown cluster: cytolysis, hepatitis, hepatic), whereas the middle cluster represents patients with both conditions. Only patients who were labelled digestive are clustered; we thus represent 89 patients out of 151.

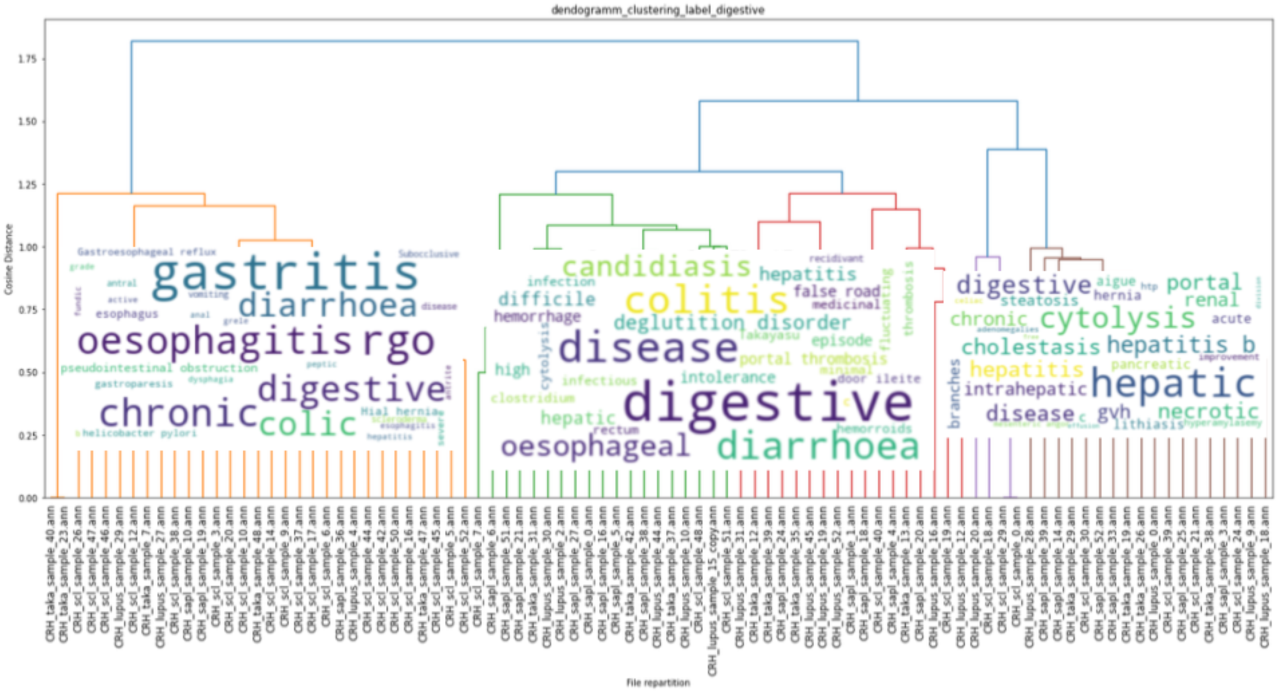


Figure 5. Unsupervised ascending hierarchical clustering based on the overall earth mover's distance of the electronic health records from equations (1) and (2). Word clouds of term frequency in the electronic health records are plotted on each respective cluster.

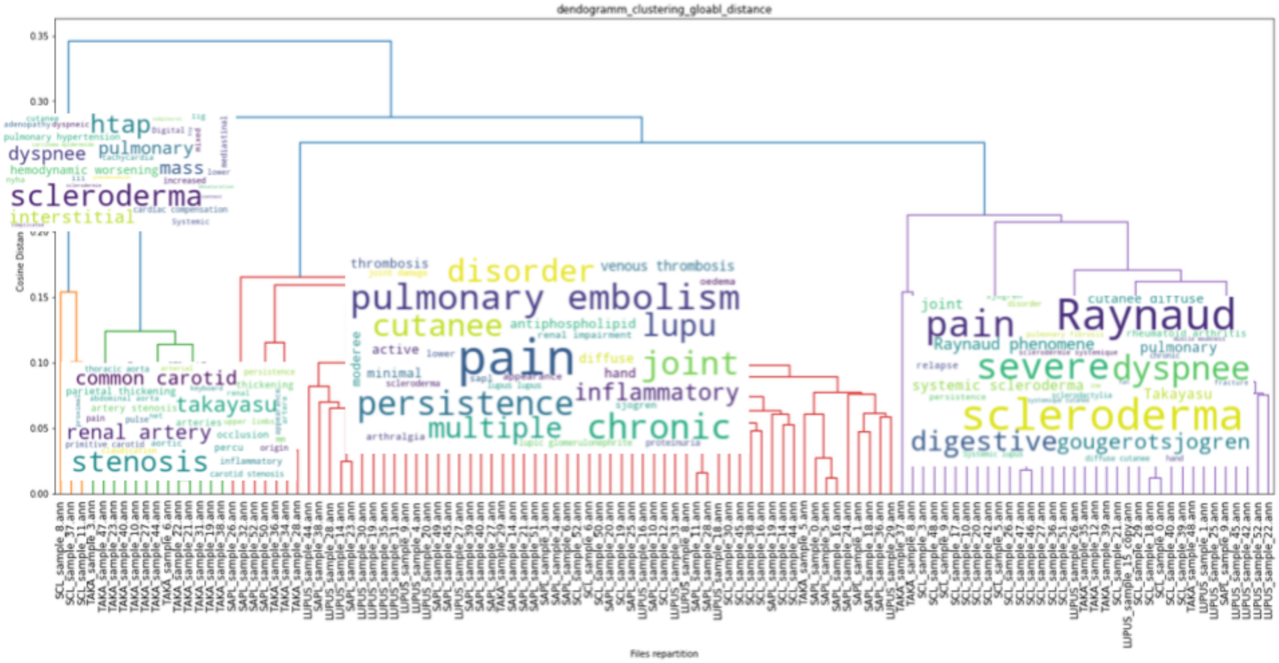


Figure 6. Unsupervised ascending hierarchical clustering based on earth mover's distance of electronic health records on the "osteomuscular" and "nutritional" labels (derived from the medical subject heading classification); only patients having the labels "osteomuscular" and "nutritional" are represented here (corresponding to 119 patients, not 256). All patients with osteoporosis were labelled "OSTEO" in the orange cluster. Other patients present in this cluster without explicitly written osteoporosis present "osteopenia" (all 4 first patients) of several vertebral fractures.

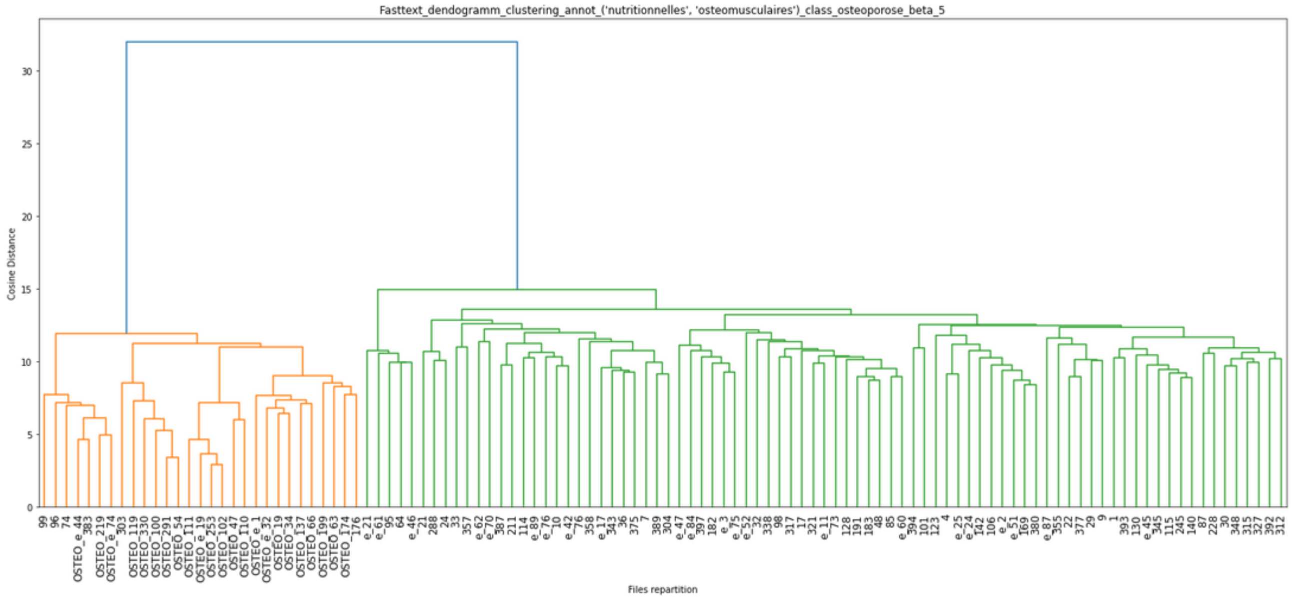
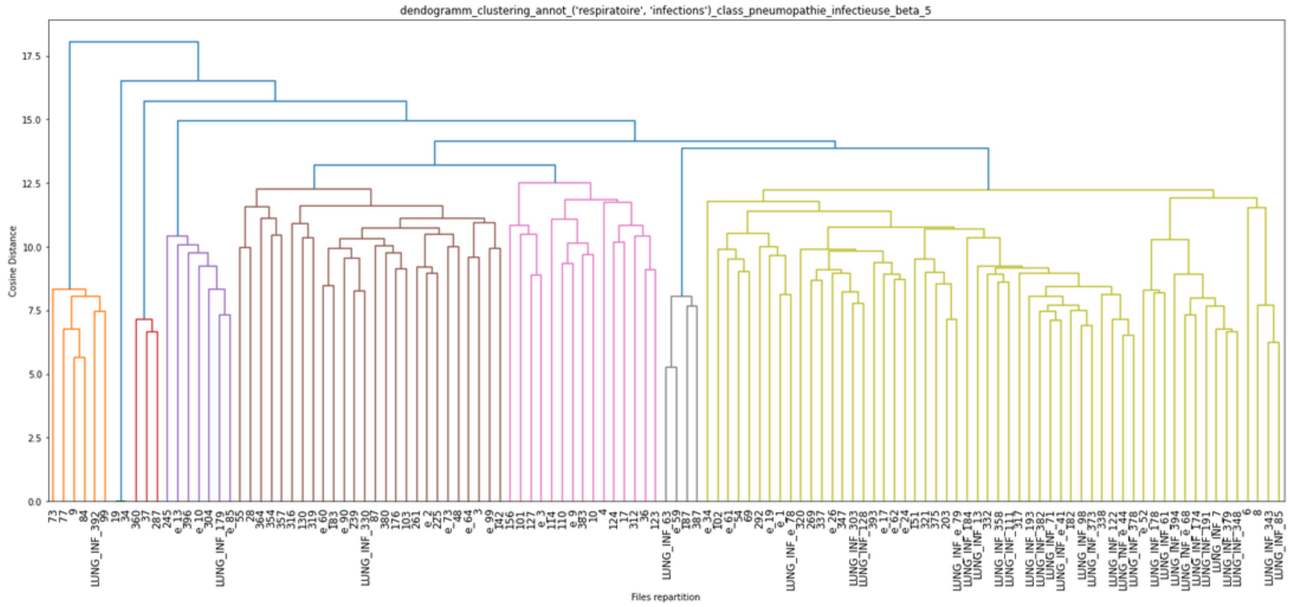


Figure 7. Unsupervised ascending hierarchical clustering based on earth mover's distance of electronic health records on the respiratory and infection axes (derived from the medical subject heading classification). All patients with lung infections were labelled “LUNG_INF” in the green cluster. Some outliers may be noticed; on the very left, the patient had purulent pleurisy, and one had pulmonary tuberculosis. The remaining patients on the left of the green cluster all had other linked manifestations such as bronchitis, parainfluenza infection, and bronchoalveolar lavage positive for *Klebsiella pneumoniae* and oropharyngeal flora.



Selection of a Cohort of Similar Patients From an Index Patient

The performance of cohort construction for the first 4 phenotypes is presented in Table 1. The last 2 phenotypes (P5-P6) could not be analyzed due to a limited number of phenotypes at the annotation stage (7 and 4, respectively).

Overall, we obtained an average precision ranging from 0.58 to 0.88, precision@10 from 0.65 to 0.98, and recall from 0.53 to 0.83. However, the average precision was lower for P3 (ILD in systemic sclerosis) owing to the higher diversity of terms used to describe the lung condition, that is, fibrosis, ILD, scleroderma with pulmonary involvement, etc, and to the fact that the phenotype annotations were very specific. As an example, sclerodermatomyositis or mixed connective tissue disease with lung involvement, which are very close to this phenotype were not annotated positively. An error analysis with mention encountered on close patients can be found in Table S1 of Multimedia Appendix 1. For the 4 phenotypes P1-P4, the precision-recall curves (means for all patients within each phenotype) were computed and are shown in Figure S1 of Multimedia Appendix 1, which is another way of showing the average precision performances. We showed very good results for the P1-P2 and P4 phenotypes and satisfactory results for the P3 phenotype since the patients had to present exactly the same disease.

Table 1. Performance results for phenotype similarity (mean and 95% CI) for all patients of a phenotype. For each phenotype, each patient in the test set is chosen in turn as an index patient, and the final results are an average of all patients.

	P1, osteoporosis (n=23)	P2, nephritis in systemic lupus erythematosus (n=48)	P3, interstitial lung disease in systemic sclerosis (n=20)	P4, lung infections (n=33)
Precision@3[a]	0.97 (0.91-1.0)	0.99 (0.98-1.0)	0.85 (0.75-0.95)	0.92 (0.84-0.99)
Precision@10	0.95 (0.91-0.99)	0.98 (0.97-0.99)	0.65 (0.58-0.72)	0.86 (0.81-0.92)
Average precision	0.88 (0.85-0.90)	0.85 (0.83-0.87)	0.58 (0.54-0.62)	0.72 (0.69-0.75)
Recall[b]	0.83 (0.81-0.84)	0.79 (0.77-0.80)	0.53 (0.50-0.55)	0.66 (0.64-0.68)

[a] Precision@3 patients (precision@10) is presented, which represents the obtained precision calculated on the 3 (or 10) patients closest to the index patient (ie, with the minimum distance).

[b] Recall is the recall calculated for all patients to be found with the same phenotype (ie, recall calculated on the 23 closest patients for osteoporosis, the 48 closest patients for nephritis in systemic lupus erythematosus, etc). Precision-recall curves for the 4 phenotypes are shown in Figure S1 of Multimedia Appendix 1.

Visualization

As an illustration, Figures 8 and 9 below show the search results described earlier for a patient with ILD in systemic sclerosis and nephritis in SLE, respectively. We see that for an index patient with ILD in systemic sclerosis (Figure 8), choosing the immune and respiratory labels led to the finding of 10 patients out of the 15 first, having the same condition. Interestingly, among these 15 samples, the 5 unlabeled patients had a disease very close to the expected one: “ILD evolving to fibrosis” and a “mixed connective tissue disease” for the first one (note_98, rank 4) and “sclerodermatomyositis” and “interstitial lung disease” for the second (note_182, rank 5). Further analysis of the errors is presented in Table S1 of Multimedia Appendix 1. A more extensive error analysis can be found in Table S1 of Multimedia Appendix 1. Figure 9 shows the search results for an index patient with nephritis in SLE. All the 21st closest patients on labels “immune” and “urogenital” showed nephritis in SLE.

Figure 8. Search results of an index patient with interstitial lung disease; the darker the color is, the closer the patients are to that particular label. Here, the selected labels “immune” and “respiratory” in 8 of the 10 first patients are labelled with “PINS_Sclerodermie” (in French, ie, interstitial lung disease in systemic sclerosis).

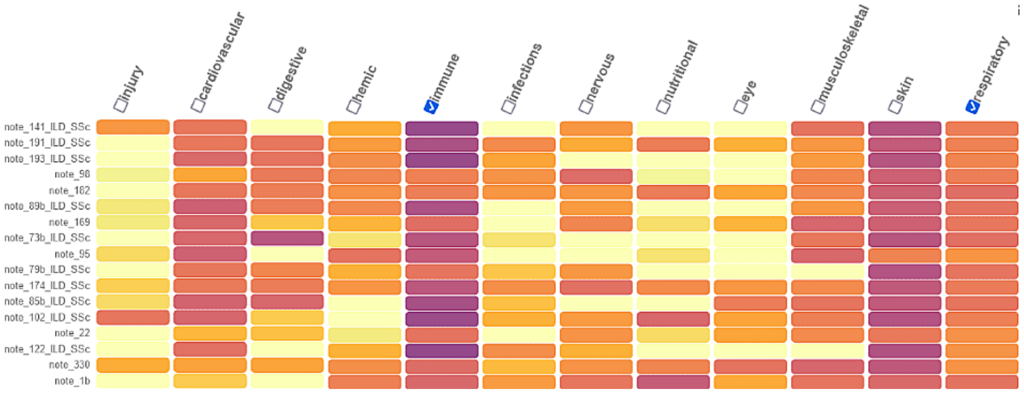
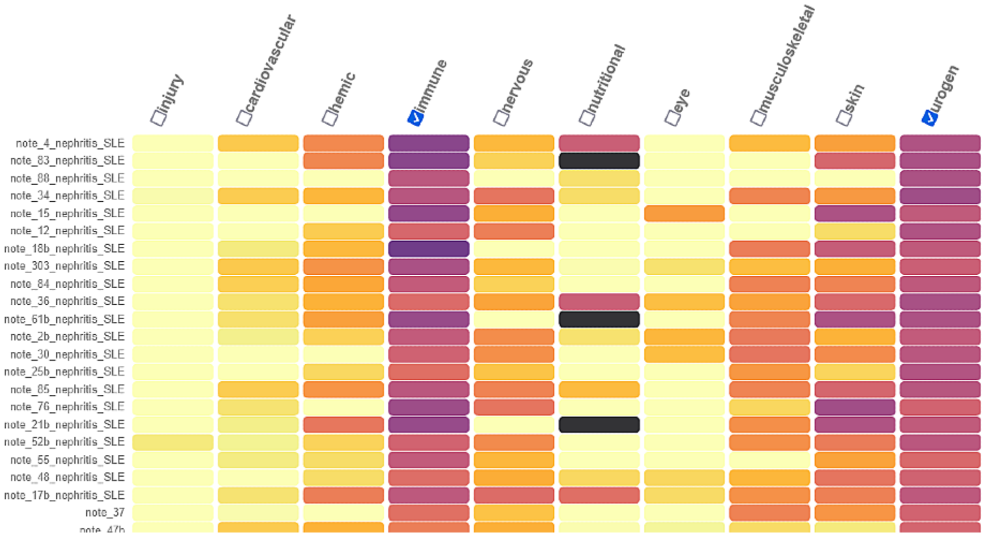


Figure 9. Search results of a patient with nephritis in systemic lupus erythematosus. The darker the color is, the closer the patients are to that particular label. Here, the selected labels “immune” and “urogenital” in all the 20 first closest patients are labelled with the right phenotype nephro_lupus.



Discussion

Summary

In this study, we developed a novel end-to-end algorithm from raw clinical notes to cohort similarity extraction. We have shown that we can cluster very specific phenotypes on an annotated data set and build similarity cohorts with good mean average precision results. These phenotypes and diseases were chosen as a proof of concept, with 2 general phenotypes such as osteoporosis and lung infection and 2 very specific phenotypes with nephritis in SLE and ILD in scleroderma. However, our algorithm can be applied to other phenotypes or diseases as well. Furthermore, our system can be applied to any other data warehouse and does not contain any handcrafted rules. An interactive demo is available online [28], and all our codes are available on GitHub [29].

Advantages of Our Approach

The main advantage of our approach is the proximity to clinical reasoning—the named-entity recognition step focusing on the distinction between physiological and pathological signs and the observations of the patients on the 22 main medical domains (cardiovascular, pulmonary, hemic, immune)—thereby allowing clinicians to choose on which aspect patients should be similar. This analysis provides interpretable results to clinicians as well as high modularity, which is essential in the field of therapeutic decision support. In clinical practice, this algorithm would enable the physician to automatically extract similar patients, evaluate their clinical evolution, and extrapolate them to the patient they want to treat. Our algorithm focuses on 1 patient’s hospitalization report rather than on the entire patient’s record (EHR), as we want to extract patients with similar conditions and similar acute complications at a time. This algorithm is also able to compare along very fine-grained characteristics. For example, 2 patients with osteoporosis complicated by a bone fracture will be closer than 2 patients with osteoporosis without a fracture. In addition, although our algorithm does not directly consider biological results in a quantitative manner, the clinician's interpretation of these results in the text is systematically integrated and analyzed as a symptom, for example, anemia, hypoalbuminemia, and positive antibodies. Similarly, the pathological description of imaging reports, such as an alveolar condensation in radiology images or an abnormal left ventricular ejection fraction in echocardiograms will be taken into account in our algorithm. We show very good results in terms of precision and average precision for selecting similar patient cohorts. The robustness of the algorithm is demonstrated on the one hand by the evaluation of the precision-to-3, which is calculated here not for the construction of the cohort but rather to show that there is, as expected, a gradient of similarity from the closest to the most distant patients, and on the other hand, as shown in the error analysis, patients close to a given index patient had very similar disease, even if the exact phenotype was not encountered.

Comparison With Previous Work

Other studies have focused on patient similarity cohorts; for instance, in the French language, Garcelon et al [30] used a patient representation and a similarity measure to try to find patients with rare diseases in the Dr Warehouse database [31]. Although their objective is quite similar to ours, they used a different representation based on the term frequency–inverse document frequency weights of the extracted concept in each clinical note, and the concept extraction is based on handcrafted rules. They obtained a percentage of 71%-99% of indexed patients returning at least one similar true-positive patient within the first 30 similar patients, and the average number of patients with exactly the same disease among the 30 patients was 51%. In a second study based on the same term frequency–inverse document frequency similarity metric, they evaluated the association between clinical phenotypes and rare disease and measured the relevance of the first 50 similar patients by a domain expert a posteriori; they obtained average precision from 0.55 to 0.91 on 6 phenotypes with mean average precision of 0.79 [32]. The main differences from our method are that we focus on clinical interpretability, and our metric computation is based on one of the most recent and performant language models [12]. Moreover, in our case, the test set was annotated a priori. Jia et al [33] also proposed an interesting algorithm for diagnostic prediction based on patient similarity, but unlike our method, their named-entity recognition step is based on a dictionary of symptoms, while disorders are extracted from ICD-10 coding. The similarity regarding symptoms is binary: 1 if the symptom is shared by both patients and 0 if otherwise. The similarity of diseases is based on their respective ICD-10 similarity (using the ICD-10 coding tree structure).

Ng et al [34] presented an insightful method based on a precision cohort (ie, patient-similarity cohorts) to help clinicians make treatment decisions for chronic diseases. They trained a global similarity model on a set of thousands of predefined variables (disease variables were constructed using their ICD-9 and ICD-10 codes, laboratory variables with their Logical Observation Identifiers Names and Codes, etc) that learns a disease-specific distance (for the 3 chronic diseases presented: hypertension, type 2 diabetes mellitus, and hyperlipidemia), with significant manual work to build the training data set. The authors did not compute direct measures of similarity cohorts but the direct impact of their method, with 75%, 74%, and 85% of decision points in hypertension, diabetes, and hyperlipidemia, respectively, and with at least one significantly better treatment. In contrast, our method focused on the performance of the similarity cohorts with metrics used in the information retrieval field, does not rely on manual variable definition, and does not learn disease-specific distance but a completely generic distance. One of the main advantages of our work is the original calculation of distance per class between patients; to the best of our knowledge, there is no similar work in the literature to compare our work to. However, we show that the named-entity recognition algorithm obtained state-of-the-art results, and the multilabel classification obtained the same performance as the best team of a French national challenge [18].

Limitations

Our work has several limitations. First, it does not cover mental health diseases, which are a completely different branch of the MeSH classification. However, training the multilabel classifier with a new label for mental health diseases with MeSH terms and synonyms can be done fairly directly based on our framework. In addition, due to time constraints, the data used in this paper were labeled by only 1 internist, and the quality of the data labeling cannot be assessed. In addition, one could argue that we did not compare our clustering and cohort similarity extraction with an ICD-10 extraction. However, because we built our initial data set with ICD-10 codes for our 4 main pathologies, we had an initial bias that we could not overcome for fair comparison. In addition, nephritis in SLE, ILD in systemic sclerosis, and lung infections do not have direct ICD-10 codes used in clinical practice. For example, "glomerular disease with SLE" has the ICD-10 "M3214" but in the entire database of 39 different hospitals, no patient had this particular code. This is because the coding is primarily done to describe the severity of the patient being managed, and this last code, in particular, does not reflect the severity of the renal involvement (in our case, codes for nephritis usually used would be N03, N04, or N05 and M320, M321, M328, and M329 for SLE). Similarly, scleroderma with pulmonary involvement has an ICD-10 code M348 that also does not appear in our database.

Assuming that an important clinical fact is repeated several times in a clinical report (eg, a patient hospitalized for acute coronary syndrome will have many cardiovascular terms linked to his/her cardiac condition), our distance computation from equations 1 and 2 depends on the number of terms in the document. Hence, 2 patients with the same major (repeated) problem would be relatively close. However, sometimes, repeated terms are not directly derived from a major clinical fact (for instance, medical history may be repeated several times without clinical relevance).

Conclusion

In this work, we have presented a novel end-to-end interpretable algorithm to automatically extract similar patients from an index patient based on clinical note analysis. Our algorithm shows good performance results for 4 specific phenotypes in the context of 4 systemic diseases. In this work, we focused only on pathological signs, but in clinical practice, one could also be interested in negative signs (for instance, the absence of Raynaud syndrome is very atypical in systemic sclerosis). This will be added in our future work, thereby adding a new physiological dimension to patients. In future work, the drug information will also be added for patient comparison, and similar to our presented approach, the clinician will then be able to focus only on treatments or on treatments and signs and symptoms. Finally, we will consider patients as a set of multiple longitudinal hospitalization reports (EHRs). An important perspective of this work is also to evaluate this tool in clinical practice.

Acknowledgments

The authors would like to thank the Assistance Publique-Hôpitaux de Paris data warehouse, which provided the data and the computing power to carry out this study under good conditions. We would like to thank all the medical colleges, including the college of internal medicine, especially Prof Jean-Emmanuel Kahn, Dr Guillaume Bussone, Prof Sébastien Abad, Dr Virginie Zarrouk, Dr Noémie Chanson, Dr Antoine Dossier, Prof Luc Mouthon, and Dr Geoffrey Cheminet from the department of rheumatology. We would also like to thank Dr Augustin Latourte, Dr Florent Eymard, Prof Xavier Mariette, Dr Gaétane Nocturne, Prof Raphaele Serror, Prof Sébastien Ottaviani, Prof Francis Berenbaum, Prof Jérémie Sellam, Prof Yannick Allanore, Prof Jérôme Avouac, Prof Maxime Breban, Dr Félicie Costantino, and doctors from the dermatology, nephrology, pneumology, hepato-gastroenterology, hematology, endocrinology, gynecology, infectiology, cardiology, oncology, emergency, and intensive care units, who gave their agreements for the use of the clinical data.

Data Availability

The data sets generated during this study (anonymized similarity measures between patients for the 4 use cases described in this paper) are available in the data repository at this link [35]. The data sets analyzed in this study are not publicly available due the confidentiality of data from patient records, even after deidentification. However, access to the Assistance Publique-Hôpitaux de Paris data warehouse's raw data can be granted following the process described on its website [36] by contacting the Ethical and Scientific Community at secretariat.cse@aphp.fr. A prior validation of the access by the local institutional review board is required. In the case of researchers who are not from the Assistance Publique-Hôpitaux de Paris, the signature of a collaboration contract is mandatory.

Authors' Contributions

CG was involved in conceptualization, data curation, formal analysis, investigation, methodology, software validation, writing the original draft, reviewing, and editing. Arthur M was involved in data curation, methodology, annotation, and writing the original draft. Arsène M was involved in designing the methodology and writing the original draft. XT was involved in conceptualization, formal analysis, methodology design, writing the original draft, reviewing, and editing. FC was involved in conceptualization, methodology, project administration, supervision, writing the original draft, reviewing, and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Examples of terms extracted from a hospitalization report close to an index patient with interstitial lung disease.

[DOCX File , 63 KB - medinform_v10i12e42379_app1.docx]

References

1. Savova GK, Danciu I, Alamudun F, Miller T, Lin C, Bitterman DS, et al. Use of Natural Language Processing to Extract Clinical Cancer Phenotypes from Electronic Medical Records. Cancer Res 2019 Nov 01;79(21):5463-5470 [FREE Full text] [doi: 10.1158/0008-5472.CAN-19-0579] [Medline: 31395609]
2. Celi LA, Galvin S, Davidzon G, Lee J, Scott D, Mark R. A Database-driven Decision Support System: Customized Mortality Prediction. J Pers Med 2012 Sep 27;2(4):138-148 [FREE Full text] [doi: 10.3390/jpm2040138] [Medline: 23766893]
3. Lieu TA, Herrinton LJ, Buzkov DE, Liu L, Lyons D, Neugebauer R, et al. Developing a Prognostic Information System for Personalized Care in Real Time. EGEMS (Wash DC) 2019 Mar 25;7(1):2 [FREE Full text] [doi: 10.5334/egems.266] [Medline: 30937324]
4. Frankovich J, Longhurst CA, Sutherland SM. Evidence-based medicine in the EMR era. N Engl J Med 2011 Nov 10;365(19):1758-1759. [doi: 10.1056/NEJMp1108726] [Medline: 22047518]
5. Callahan A, Polony V, Posada JD, Banda JM, Gombar S, Shah NH. ACE: the Advanced Cohort Engine for searching longitudinal patient records. J Am Med Inform Assoc 2021 Jul 14;28(7):1468-1479 [FREE Full text] [doi: 10.1093/jamia/ocab027] [Medline: 33712854]

6. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical Natural Language Processing in languages other than English: opportunities and challenges. J Biomed Semantics 2018 Mar 30;9(1):12 [FREE Full text] [doi: 10.1186/s13326-018-0179-8] [Medline: 29602312]
7. Farzandipour M, Sheikhtaheri A, Sadoughi F. Effective factors on accuracy of principal diagnosis coding based on International Classification of Diseases, the 10th revision (ICD-10). International Journal of Information Management 2010 Feb;30(1):78-84 [FREE Full text] [doi: 10.1016/j.ijinfomgt.2009.07.002]
8. Benkhaial A, Kaltschmidt J, Weisshaar E, Diepgen TL, Haefeli WE. Prescribing errors in patients with documented drug allergies: comparison of ICD-10 coding and written patient notes. Pharm World Sci 2009 Aug;31(4):464-472. [doi: 10.1007/s11096-009-9300-5] [Medline: 19412703]
9. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. ArXiv 2013:1-12 [FREE Full text] [doi: 10.48550/arXiv.1301.3781]
10. Pennington J, Socher. Global vectors for word representation. Glove; 2014 Presented at: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October; Doha, Qatar URL: http://www.aclweb.org/anthology/D14-1162 [doi: 10.3115/v1/d14-1162]
11. 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]
12. Devlin J, Chang M, Lee. BERT: Pre-training of deep bidirectional transformers for language understanding. ACL Anthology. URL: https://aclanthology.org/N19-1423.pdf [accessed 2018-10-11]
13. De Freitas JK, Johnson KW, Golden E, Nadkarni GN, Dudley JT, Bottinger EP, et al. Phe2vec: Automated disease phenotyping based on unsupervised embeddings from electronic health records. Patterns (N Y) 2021 Sep 10;2(9):100337 [FREE Full text] [doi: 10.1016/j.patter.2021.100337] [Medline: 34553174]
14. Jonquet C, Shah NH, Musen MA. The open biomedical annotator. Summit Transl Bioinform 2009 Mar 01;2009:56-60 [FREE Full text] [Medline: 21347171]
15. Ferté T, Cossin S, Schaeverbeke T, Barnetche T, Jouhet V, Hejblum BP. Automatic phenotyping of electronical health record: PheVis algorithm. J Biomed Inform 2021 May;117:103746 [FREE Full text] [doi: 10.1016/j.jbi.2021.103746] [Medline: 33746080]
16. FAI2R. URL: https://www.fai2r.org/ [accessed 2022-11-25]
17. Takayasu Arteritis. URL: https://www.has-sante.fr/upload/docs/application/pdf/2020-01/pnds_takayasu_fair_-_favamulti.pdf [accessed 2022-11-25]
18. Gérardin C, Wajsbürt P, Vaillant P, Bellamine A, Carrat F, Tannier X. Multilabel classification of medical concepts for patient clinical profile identification. Artif Intell Med 2022 Jun;128:102311. [doi: 10.1016/j.artmed.2022.102311] [Medline: 35534148]
19. CNIL. URL: https://www.cnil.fr/en/home [accessed 2018-05-10]
20. MeSH. National Center for Biotechnology Information. URL: https://www.ncbi.nlm.nih.gov/mesh/ [accessed 2017-02-10]
21. Martin L, Muller B, Suárez P, Dupont Y, Romary L, de La Clergerie E. CamemBERT: a tasty French language model. ACL Anthology. URL: https://aclanthology.org/2020.acl-main.645.pdf [accessed 2020-07-01]
22. Sang, Erik F, Fien De Meulder. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. ACL Anthology. 2003. URL: https://aclanthology.org/W03-0419.pdf [accessed 2003-06-12]
23. Kim J, Ohta T, Tateisi Y, Tsujii J. GENIA corpus--semantically annotated corpus for bio-textmining. Bioinformatics 2003;19 Suppl 1:i180-i182. [doi: 10.1093/bioinformatics/btg1023] [Medline: 12855455]
24. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
25. Unified medical language system. National Library of Medicine. URL: https://www.nlm.nih.gov/research/umls/index.html [accessed 2022-11-25]
26. Kusner M, Sun Y, Kolkin. From word embeddings to document distances. 2015 Presented at: ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning; July 6-11; Lille, France p. 957-966 URL: https://dl.acm.org/doi/10.5555/3045118.3045221
27. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, SciPy 1.0 Contributors. Author Correction: SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020 Mar;17(3):352 [FREE Full text] [doi: 10.1038/s41592-020-0772-5] [Medline: 32094914]
28. Patient similarity demo. Xavier Tannier. 2022. URL: http://xavier.tannier.free.fr/misc/patient_similarity/demo.html [accessed 2022-05-20]
29. Gérardin C. Cohort similarity. GitHub. 2022. URL: https://github.com/ChristelDG/cohort-similarity [accessed 2022-05-20]
30. Garcelon N, Neuraz A, Benoit V, Salomon R, Kracker S, Suarez F, et al. Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack. J Biomed Inform 2017 Sep;73:51-61 [FREE Full text] [doi: 10.1016/j.jbi.2017.07.016] [Medline: 28754522]

31. Garcelon N, Neuraz A, Salomon R, Faour H, Benoit V, Delapalme A, et al. A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. J Biomed Inform 2018 Apr;80:52-63 [FREE Full text] [doi: 10.1016/j.jbi.2018.02.019] [Medline: 29501921]
32. Garcelon N, Neuraz A, Salomon R, Bahi-Buisson N, Amiel J, Picard C, et al. Next generation phenotyping using narrative reports in a rare disease clinical data warehouse. Orphanet J Rare Dis 2018 May 31;13(1):85 [FREE Full text] [doi: 10.1186/s13023-018-0830-6] [Medline: 29855327]
33. Jia Z, Zeng X, Duan H, Lu X, Li H. A patient-similarity-based model for diagnostic prediction. Int J Med Inform 2020 Mar;135:104073 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.104073] [Medline: 31923816]
34. Ng K, Kartoun U, Stavropoulos H, Zambrano JA, Tang PC. Personalized treatment options for chronic diseases using precision cohort analytics. Sci Rep 2021 Jan 13;11(1):1139 [FREE Full text] [doi: 10.1038/s41598-021-80967-5] [Medline: 33441956]
35. Gérardin C. Cohort similarity main data. GitHub. 2022. URL: https://github.com/ChristelDG/cohort-similarity/tree/main/data [accessed 2022-05-23]
36. Entrepot de données de Santé de l'AP-HP. Citrix Gateway. URL: https://www.eds.aphp.fr [accessed 2022-05-20]

Abbreviations

AP-HP: Assistance Publique-Hôpitaux de Paris
APS: antiphospholipid syndrome
BERT: Bidirectional Encoder Representations from Transformers
EHR: electronic health record
ICD-9/ICD-10: International Classification of Diseases, Ninth/Tenth Revision
ILD: interstitial lung disease
MeSH: medical subject heading
SLE: systemic lupus erythematosus

Edited by C Lovis; submitted 01.09.22; peer-reviewed by Y Xiong, J Candeliere, C Gaudet-Blavignac; comments to author 09.10.22; revised version received 17.10.22; accepted 22.10.22; published 19.12.22.

Please cite as:
Gérardin C, Mageau A, Mékinian A, Tannier X, Carrat F
Construction of Cohorts of Similar Patients From Automatic Extraction of Medical Concepts: Phenotype Extraction Study
JMIR Med Inform 2022;10(12):e42379
URL: https://medinform.jmir.org/2022/12/e42379
doi:10.2196/42379
PMID:36534446

Original Paper

Synchronous Teleconsultation and Monitoring Service Targeting COVID-19: Leveraging Insights for Postpandemic Health Care

Milena Soriano Marcolino[1,2,3], MD, MSc, PhD; Clara Sousa Diniz[2], MD; Bruno Azevedo Chagas[4], MSc; Mayara Santos Mendes[2], MSc; Raquel Prates[4], MSc, PhD; Adriana Pagano[5], BA, MA, PhD; Thiago Castro Ferreira[5], MSc, PhD; Maria Beatriz Moreira Alkmim[2], MSc, MD; Clara Rodrigues Alves Oliveira[1,2], MD, MSc, PhD; Isabela Nascimento Borges[1,2], MD, PhD; Magda César Raposo[6], BSc; Zilma Silveira Nogueira Reis[2], MD, MSc, PhD; Maria Cristina Paixão[2], BSc; Leonardo Bonisson Ribeiro[2], BSc; Gustavo Machado Rocha[7], MD, MSc, PhD; Clareci Silva Cardoso[2,6,7], MSc, PhD; Antonio Luiz Pinho Ribeiro[1,2,3], MD, PhD

[1]Department of Internal Medicine, Medical School, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[2]Telehealth Center, University Hospital and Telehealth Network of Minas Gerais, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[3]Institute for Health Technology Assessment, Porto Alegre, Brazil

[4]Department of Computer Science, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[5]Arts Faculty, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[6]Telehealth Center, Universidade Federal de São João Del-Rei, Divinópolis, Brazil

[7]Medical School, Campus Centro Oeste, Universidade Federal de São João del-Rei, Divinópolis, Brazil

Corresponding Author:
Milena Soriano Marcolino, MD, MSc, PhD
Telehealth Center
University Hospital and Telehealth Network of Minas Gerais
Universidade Federal de Minas Gerais
Avenida Professor Alfredo Balena, 110 Room 107 Ala Sul
Santa Efigênia
Belo Horizonte, 30130-100
Brazil
Phone: 55 31 33079201
Email: milenamarc@gmail.com

Abstract

Background: Although a great number of teleconsultation services have been developed during the COVID-19 pandemic, studies assessing usability and health care provider satisfaction are still incipient.

Objective: This study aimed to describe the development, implementation, and expansion of a synchronous teleconsultation service targeting patients with symptoms of COVID-19 in Brazil, as well as to assess its usability and health care professionals' satisfaction.

Methods: This mixed methods study was developed in 5 phases: (1) the identification of components, technical and functional requirements, and system architecture; (2) system and user interface development and validation; (3) pilot-testing in the city of Divinópolis; (4) expansion in the cities of Divinópolis, Teófilo Otoni, and Belo Horizonte for Universidade Federal de Minas Gerais faculty and students; and (5) usability and satisfaction assessment, using Likert-scale and open-ended questions.

Results: During pilot development, problems contacting users were solved by introducing standardized SMS text messages, which were sent to users to obtain their feedback and keep track of them. Until April 2022, the expanded system served 31,966 patients in 146,158 teleconsultations. Teleconsultations were initiated through chatbot in 27.7% (40,486/146,158) of cases. Teleconsultation efficiency per city was 93.7% (13,317/14,212) in Teófilo Otoni, 92.4% (11,747/12,713) in Divinópolis, and 98.8% (4981/5041) in Belo Horizonte (university campus), thus avoiding in-person assistance for a great majority of patients. In total, 50 (83%) out of 60 health care professionals assessed the system's usability as satisfactory, despite a few system instability problems.

Conclusions: The system provided updated information about COVID-19 and enabled remote care for thousands of patients, which evidenced the critical role of telemedicine in expanding emergency services capacity during the pandemic. The dynamic nature of the current pandemic required fast planning, implementation, development, and updates in the system. Usability and satisfaction assessment was key to identifying areas for improvement. The experience reported here is expected to inform telemedicine strategies to be implemented in a postpandemic scenario.

(JMIR Med Inform 2022;10(12):e37591) doi:10.2196/37591

KEYWORDS

COVID-19; telemonitoring; remote consultation; telemedicine; primary health care; delivery of health care; telehealth; text message; mobile health; public health; remote care; digital health; usability

Introduction

The COVID-19 pandemic has brought dramatic transformative changes in economies, societies, and health care, with an unprecedented challenge to public health worldwide [1]. The need to avoid patient crowds in health services and offer alternative ways for patient assistance while preserving physical distancing and isolation, as well as the prioritization of emergency departments and intensive care units, have proven to be important drivers for the urgent need and quick adoption of telemedicine.

Telehealth services were growing exponentially prior to COVID-19 [2]. However, it was during the pandemic that they received a major boost. Governments from different countries were urged to promote telehealth and make provisions to address some of the previously encountered barriers, and they quickly updated law restrictions and reimbursement policies [3]. In Brazil, telehealth has been consolidated over the years, but it was only after the spread of COVID-19 that a legal and regulatory framework emerged, authorizing remote medical and other professional health consultations. The Telehealth Network of the State of Minas Gerais (TNMG) in Brazil—one of the largest public telehealth services in Latin America [4,5]—was quick to implement telemedicine services for the care of patients with suspected novel coronavirus infection soon after the first patient was diagnosed with COVID-19 in the country.

Although a great number of teleconsultation services have been developed, studies assessing usability and satisfaction from the health care provider's perspective are still incipient. Concerns have been raised regarding challenges posed by diagnosing without an actual physical examination and the negative impact on patient-provider rapport [6]. In the aftermath of COVID-19, when telehealth services are expected to remain in use and health care provider satisfaction is a key feature for telehealth sustainability, usability assessment is particularly relevant as a source to be tapped for lessons to be learned.

Our aim was to assess the feasibility of the development, implementation, and expansion of a synchronous teleconsultation service for care provided to patients with symptoms of COVID-19, as well as to perform assessments of usability and health care professionals' satisfaction.

Methods

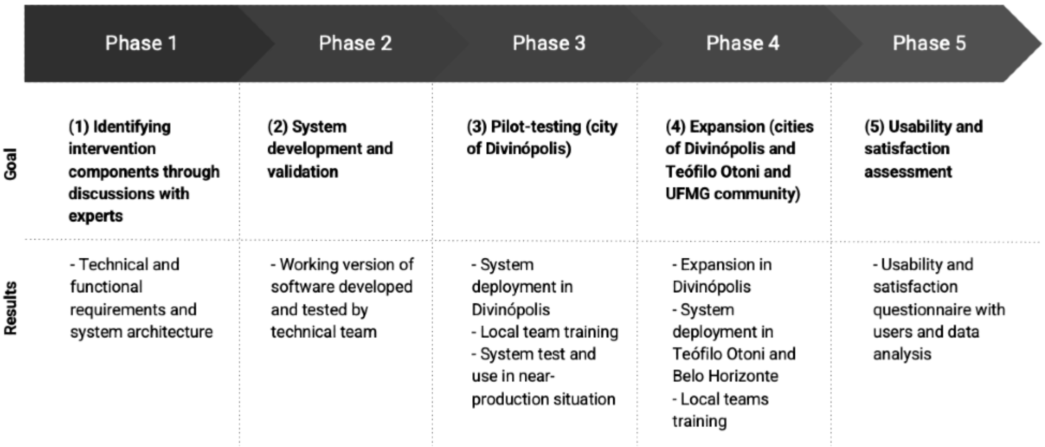
Study Design

This mixed methods study was developed in 5 phases (Figure 1), following guidance from the Medical Council Framework [7]:

1. Identifying intervention components through discussions with experts;
2. System development and validation;
3. Pilot-testing;
4. Expansion; and
5. Usability and satisfaction assessment.

Each phase is briefly explained in the following subsections.

Figure 1. Project phases. UFMG: Universidade Federal de Minas Gerais.



	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5
Goal	(1) Identifying intervention components through discussions with experts	(2) System development and validation	(3) Pilot-testing (city of Divinópolis)	(4) Expansion (cities of Divinópolis and Teófilo Otoni and UFMG community)	(5) Usability and satisfaction assessment
Results	- Technical and functional requirements and system architecture	- Working version of software developed and tested by technical team	- System deployment in Divinópolis - Local team training - System test and use in near-production situation	- Expansion in Divinópolis - System deployment in Teófilo Otoni and Belo Horizonte - Local teams training	- Usability and satisfaction questionnaire with users and data analysis

Identifying Intervention Components Through Discussions With Experts

To identify components in the intervention, information was extracted from guidance issued by the Brazilian Ministry of Health [8,9] and evidence available at the onset of the pandemic, as well as discussions among an interdisciplinary team of IT specialists, clinicians with long-term expertise in telemedicine [4,5], infectious diseases specialists, and nurses.

The workflow suggested by the Brazilian Ministry of Health was adapted to improve the assistance flow, offer agility for the teams and data security, and reduce the burden of patients who need in-person consultations at primary care centers. The municipality where the system was planned to run initially—Divinópolis—had a telephone service dedicated for the general population to answer queries related to COVID-19 and for primary care practitioners, medical university professors, and undergraduate medical students working in a monitoring program. All available resources were used to design an integrated teleassistance flow, which assisted patients from their initial doubt through to clinical assistance and monitoring, at 4 levels: level 1, performed by local health care professionals (nursing technicians, physiotherapists, nutritionists, and psychologists); level 2, performed by nursing staff; level 3, performed by medical staff; and level 4, telemonitoring that was performed by students under medical supervision.

Although this teleassistance flow was not fully integrated into the local emergency departments, upon concluding teleconsultations, when face-to-face assessments were deemed necessary, the patients could be referred to face-to-face medical consultations with a specific clinical report.

An internal medicine specialist, a nurse, a doctor with long-term experience in telemedicine, and an IT specialist identified the main components in the intervention to map the main needs, steps in the process of care, and specificities of each screen and functionality in the system. These health care professionals worked alongside the IT specialist to discuss and propose changes and improvements to the system throughout the iterative development cycle adopted. Unfortunately, due to the necessary urgency of the actions—the platform was offered for use just 2 months after the start of its development—it was not possible to involve patients in the development of the self-assessment tools.

A management model based on the Plan-Do-Check-Act cycle and a monitoring system based on key performance indicators were developed. The nurse and the doctor with long-term experience in telemedicine defined the indicators to be monitored (Table S1 in Multimedia Appendix 1).

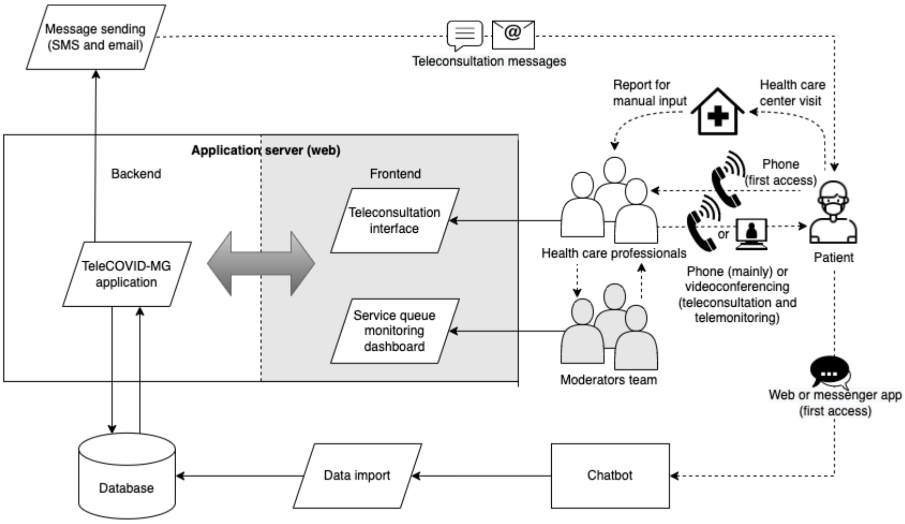
System Development and Validation

The system was developed and validated following an agile software methodology. Its backend was built using the Java programming language (version 1.8) with the *Spring Boot* and *Hibernate* frameworks, whereas the system’s user interface was built using the *Angular* user interface framework.

The system, named TeleCOVID-MG, started being developed in March 2020. Throughout March and April, the team of analysts met weekly with the clinical team to assess new requests that arose. Meanwhile, the development team delivered weekly packages that were internally tested and, on weekends, were validated and approved by professionals from the clinical team. Thus, in May 2020, the first version was released into a production environment. From then on, fortnightly sprints were adopted, generating deliveries for testing and approval.

The software runs on a web environment, which allows the full recording of activities. It is composed of an application server, which runs the main application backend and serves the frontend to the users’ client browsers, and an SQL relational database (Postgres). The frontend has 2 main interfaces: 1 for teleconsultation, which is used by the health care professionals, and another for monitoring the service queue, which is a dashboard used by the team of moderators, who identify the need for additional health care professionals in the shift to reduce the response time. A chatbot, developed using the BLiP platform (Take), was aimed to be a first point of contact for patients with the telehealth service [10,11]. It assisted in screening the severity of respiratory and flu-like symptoms and queuing patients for teleconsultation based on warning-sign severity [10,11]. There is a module for importing data from the chatbot into the database and a module for sending messages to patients (Figure 2).

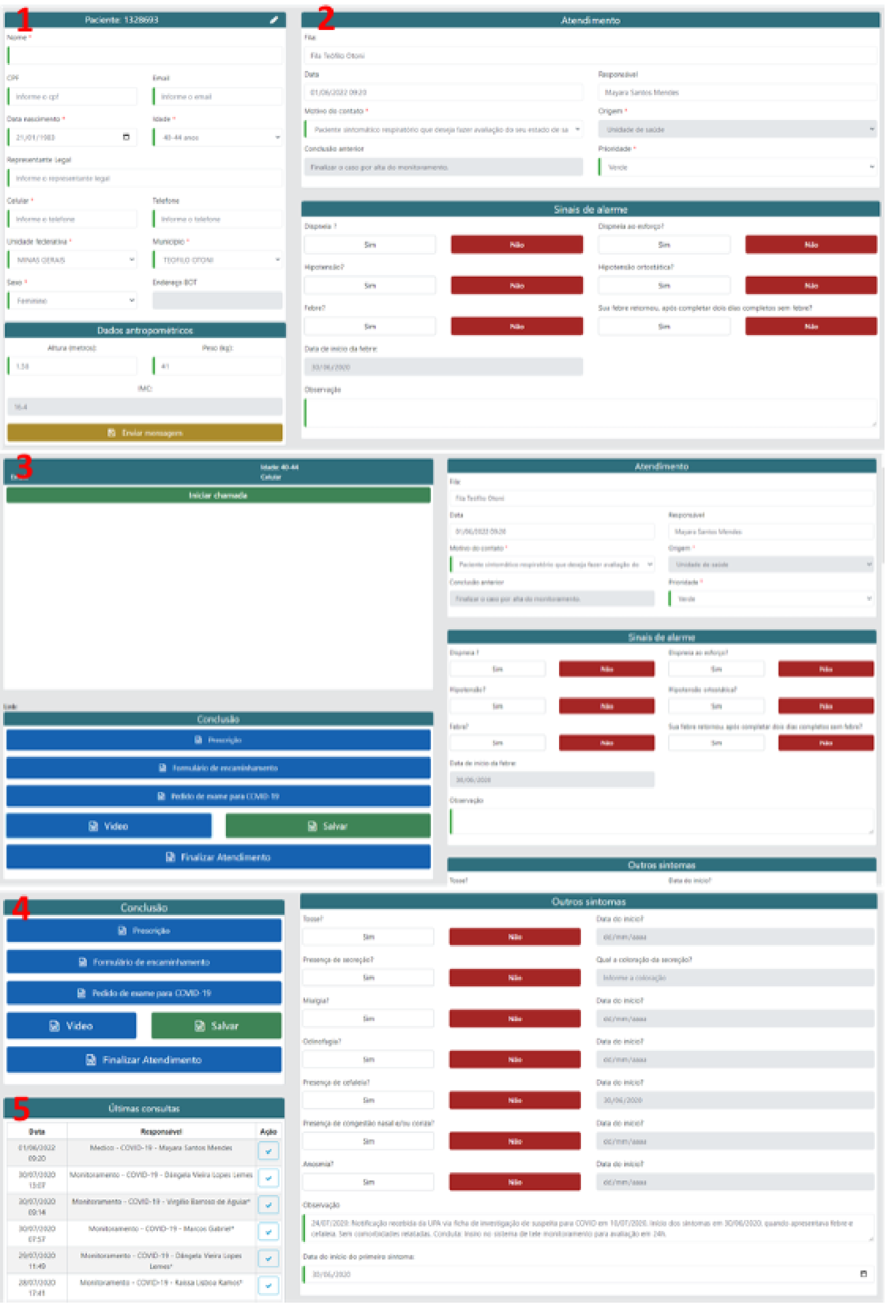
Figure 2. System architecture overview.


Message sending (SMS and email)
Teleconsultation messages
Report for manual input
Health care center visit
Application server (web)
Backend
Frontend
Teleconsultation interface
TeleCOVID-MG application
Service queue monitoring dashboard
Phone (first access)
or
Health care professionals
Patient
Phone (mainly) or videoconferencing (teleconsultation and telemonitoring)
Moderators team
Web or messenger app (first access)
Database
Data import
Chatbot

Teleconsultation and Telemonitoring Services

TeleCOVID-MG has 3 main goals: (1) assessing and managing patients with respiratory or flu-like symptoms, (2) monitoring patients with COVID-19, and (3) providing the general population with updated information about COVID-19. The system enables performing consultations either with or without videoconferencing, issuing medical prescriptions and reports, as well as issuing orders for diagnostic COVID-19 tests (Figure 3) by nurses and physicians from the TeleCOVID-MG teleconsultation team, following the Ministry of Health and local clinical protocols. All these documents generated during the teleconsultation can be easily downloaded as PDF files by the patients. The software also enables the generation of the compulsory report of COVID-19 cases, in compliance with requirements by the Brazilian Health Ministry, as well as teleconsultations scheduling, patient referral to telemonitoring services, or face-to-face consultations at other levels of care (Figure 4 and "TeleCOVID-MG service workflow" in Multimedia Appendix 1 [6,9,12]).

Figure 3. Screenshots of TeleCOVID-MG. User registration form: (1) patient personal information tab; (2) patient clinical condition tab recording warning signs; and (3) video call tab; (4) form tab (for prescriptions, reports, and test orders); and (5) record of past teleconsultations.



Paciente: 1328693

Nome *

CPF: Informe o cpf	Email: Informe o email
Data nascimento *: 21/01/1983	Idade *: 40-44 anos
Representante Legal: Informe o representante legal	
Celular *: Informe o telefone	Telefone: Informe o telefone
Unidade federativa *: MINAS GERAIS	Município *: TEOFILO OTONI
Sexo *: Feminino	Endereço BOT

Dados antropométricos

Altura (metros): 1.58	Peso (kg): 41

IMC: 16.4

Enviar mensagem

Atendimento

Fila: Fila Teófilo Otoni

Data: 01/06/2022 09:20	Responsável: Mayara Santos Mendes
Motivo do contato *: Paciente sintomático respiratório que deseja fazer avaliação do seu estado de sa	Origem *: Unidade de saúde
Conclusão anterior: Finalizar o caso por alta do monitoramento.	Prioridade *: Verde

Sinais de alarme

Dispneia ?	Sim	Não	Dispneia ao esforço?	Sim	Não
Hipotensão?	Sim	Não	Hipotensão ortostática?	Sim	Não
Febre?	Sim	Não	Sua febre retornou, após completar dois dias completos sem febre?	Sim	Não

Data de início da febre: 30/06/2020

Observação

Idade 40-44
Celular

Iniciar chamada

Conclusão

Prescrição

Formulário de encaminhamento

Pedido de exame para COVID-19

Vídeo Salvar

Finalizar Atendimento

Outros sintomas

Tosse?	Sim	Não	Data do início?	dd/mm/aaaa
Presença de secreção?	Sim	Não	Qual a coloração da secreção?	Informe a coloração
Mialgia?	Sim	Não	Data do início?	dd/mm/aaaa
Odinofagia?	Sim	Não	Data do início?	dd/mm/aaaa
Presença de cefaleia?	Sim	Não	Data do início?	30/06/2020
Presença de congestão nasal e/ou coriza?	Sim	Não	Data do início?	dd/mm/aaaa
Anosmia?	Sim	Não	Data do início?	dd/mm/aaaa

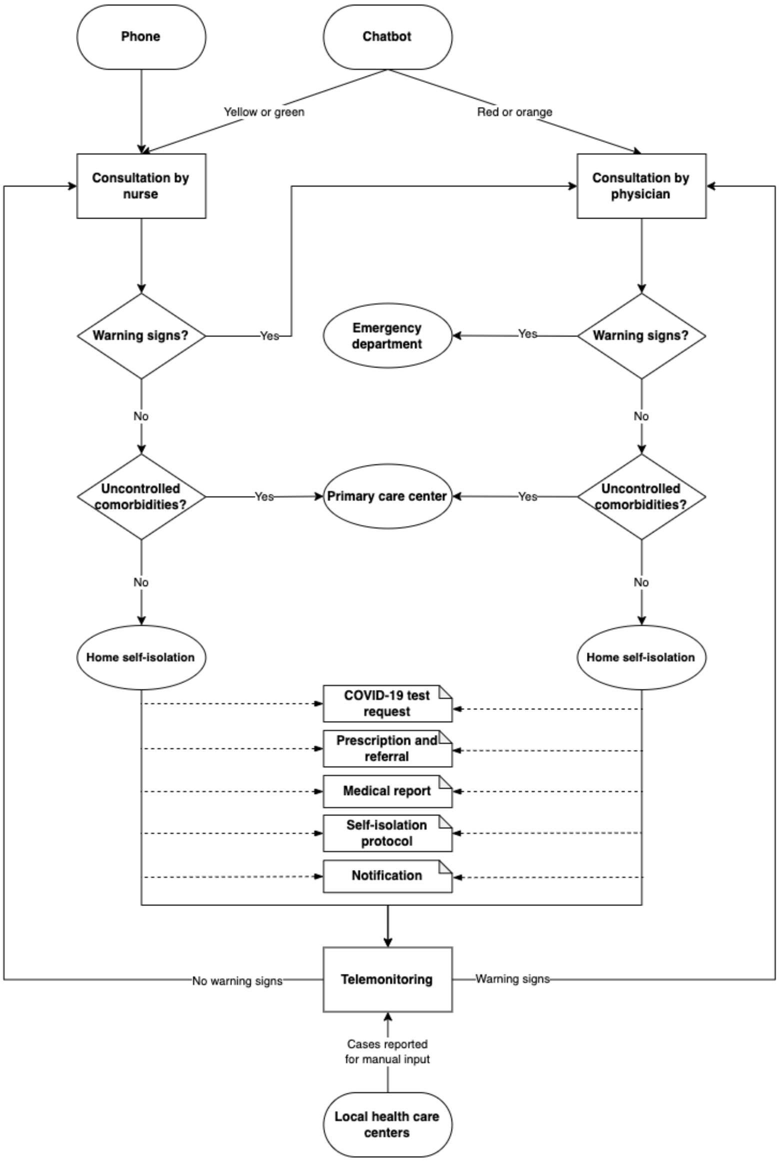
Observação: 24/07/2020: Notificação recebida da UPA via ficha de investigação de suspeita para COVID em 10/07/2020. Início dos sintomas em 30/06/2020, quando apresentava febre e cefaleia. Sem comorbidades relatadas. Conduta: Insiro no sistema de tele monitoramento para avaliação em 24h.

Data do início do primeiro sintoma: 30/06/2020

Últimas consultas

Data	Responsável	Ação
01/06/2022 09:20	Medico - COVID-19 - Mayara Santos Mendes	✓
30/07/2020 15:07	Monitoramento - COVID-19 - Dângela Vieira Lopes Lemes	✓
30/07/2020 09:14	Monitoramento - COVID-19 - Virgílio Barroso de Aguiar*	✓
30/07/2020 07:57	Monitoramento - COVID-19 - Marcos Gabriel*	✓
29/07/2020 11:49	Monitoramento - COVID-19 - Dângela Vieira Lopes Lemes*	✓
28/07/2020 17:41	Monitoramento - COVID-19 - Raissa Lisboa Ramos*	✓

Figure 4. TeleCOVID-MG service workflow.



Pilot-testing

The pilot study was carried out in Divinópolis, a 213,016-inhabitant city with a human development index of 0.76, from May 18 to 28, 2020 [12]. The team responsible for teleconsultations and telemonitoring in Divinópolis comprised physicians, nurses, and students from Universidade Federal de São João del-Rei. An instruction manual was prepared, and participating health care professionals received web-based clinical training, based on the best available scientific evidence at that time. They were also trained to use the system before starting the activities.

Expansion

The project was expanded to the Teófilo Otoni, a 140,937-inhabitant city with a human development index of 0.70 [11], and subsequently to faculty and students at Universidade Federal de Minas Gerais (UFMG), a federal university where the coordination center of the TNMG is located. It has over 45,000 students and 7400 faculty members. The team responsible for the teleconsultations comprised physicians and nurses from the TNMG, and, in the case of Teófilo Otoni, also included nurses from that city. The team responsible for telemonitoring comprised medical students supervised by medical professors and nurses. All of them received technical training to operate the system and theoretical training, as aforementioned. Weekly meetings were held with local coordinators to discuss indicators, identify deviations from planned targets, and plan and implement corrective actions.

Once patients entered the system, after the initial teleconsultation, a follow-up plan that involved monitoring or new consultations was defined based on the assessment of their situation. Teleconsultation efficiency was calculated as the number of patients who were provided with consultation and did not need to be referred to face-to-face consultations divided by the total number of patients who were provided with consultation.

For reporting expansion results, all records of patients who were provided with consultation at the 3 locations from May 2020 to April 2022 were eligible.

Usability and Satisfaction Assessment

A questionnaire was developed to assess health care professionals' satisfaction and the usability of the system, as they were the primary immediate users. It included eight 5-point Likert-scale questions that focused on aspects regarding user satisfaction and usability and open-ended questions that focused on the perceived strengths and weaknesses of the system, features to be improved, and comments about their experience with the system. All health care professionals who worked in the service and used the TeleCOVID-MG system were eligible (n=60). A thematic analysis was conducted for the open-ended questions.

Ethics Approval

Ethical approval was obtained from the UFMG Research Ethics Committee (CAAE: 35953620.9.0000.5149). Informed consent was obtained from study participants.

Results

Through the study, the system served 31,966 patients, totaling 146,158 teleconsultations covering the first and subsequent consultations performed for each patient, since the same patient could be assessed more than once by nurses, physicians, and the telemonitoring team. The accumulated number of teleconsultations and patients assisted by location and service efficiency are displayed in Figure 5. Other indicators that were monitored weekly and monthly are shown in Table S2 in Multimedia Appendix 1. The real-time analysis of these data allowed system and service workflow adjustments as necessary.

As shown in Figure 2, both the chatbot and telephone number were the gateway to the program. The telephone was primarily used, and teleconsultations were initiated through the chatbot in 27.7% (40,486/146,158) of cases. Additionally, the main method used to carry out the teleconsultations was via telephone call, with videoconferencing showing a very low usage rate (only 192 [0.13%] videoconferencing teleconsultations in total). When carried out, videoconferencing was performed via smartphone, using patients' preferred software that was previously installed on their device. The main challenge faced during the expansion phase in Teófilo Otoni was the difficulty in reaching patients, even by telephone call, which may be due to the instability of the local telephone network and the fact that part of the population lives in rural areas and lack familiarity with telecommunication tools. Due to these same logistical and cultural reasons, the use of videoconferencing and other technologies such as chatbot was even more challenging.

Another difficulty was aligning the clinical guideline developed for remote care with the practice carried out in the city. To face this challenge, training meetings were held with local health teams, and several seminars addressing theoretical issues related to the management of patients with COVID-19 were carried out. The number of assessments initiated via chatbot was low in Teófilo Otoni, which may be due to the low socioeconomic level of the population and their limited digital literacy when dealing with new technologies. However, the number of calls via chatbot was extensive among the university community at UFMG, which is consistent with college users who have the digital skills needed to deal with chatbots.

Of the 60 health care professionals who used the system when the assessment was performed, 50 (83%) answered the questionnaire (age: median 35, IQR 31-40 years; women: n=43, 86%). Of these professionals, 42% (n=21) were physicians and 54% (27) were nurses (Table S3 in Multimedia Appendix 1).

Overall, the system was evaluated as satisfactory (Table 1 and Figure S1 in Multimedia Appendix 1). The only exception was for the statement "The system is stable, and no errors occur during use," which had a median score of 4 (IQR 2-4). We believe this score reflects technical infrastructure problems and the short time taken to put the system into production, which prevented debugging. With regard to the open-ended questions, 35 participants answered at least one question, 44 commented on the system's strengths, 41 mentioned weaknesses, 37 made suggestions, and 16 commented on their experience with the system (see "Responses to the open-ended questions" in Multimedia Appendix 1).

Figure 5. Percentage of teleconsultation distribution efficiency among the 3 cities in our study. UFMG: Universidade Federal de Minas Gerais.

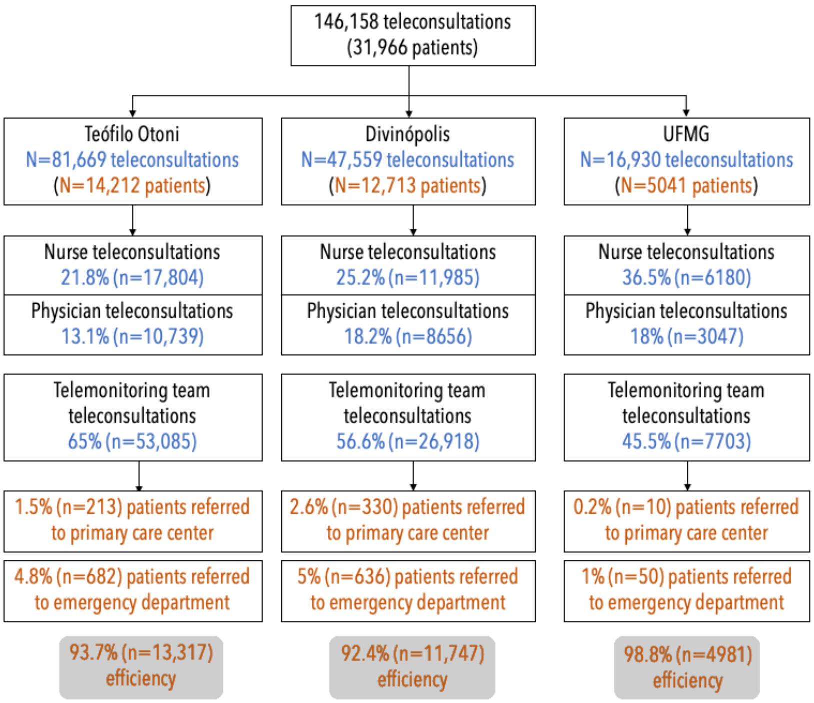

146,158 teleconsultations (31,966 patients)
Teófilo Otoni N=81,669 teleconsultations (N=14,212 patients)
Divinópolis N=47,559 teleconsultations (N=12,713 patients)
UFMG N=16,930 teleconsultations (N=5041 patients)
Nurse teleconsultations 21.8% (n=17,804)
Physician teleconsultations 13.1% (n=10,739)
Nurse teleconsultations 25.2% (n=11,985)
Physician teleconsultations 18.2% (n=8656)
Nurse teleconsultations 36.5% (n=6180)
Physician teleconsultations 18% (n=3047)
Telemonitoring team teleconsultations 65% (n=53,085)
Telemonitoring team teleconsultations 56.6% (n=26,918)
Telemonitoring team teleconsultations 45.5% (n=7703)
1.5% (n=213) patients referred to primary care center
2.6% (n=330) patients referred to primary care center
0.2% (n=10) patients referred to primary care center
4.8% (n=682) patients referred to emergency department
5% (n=636) patients referred to emergency department
1% (n=50) patients referred to emergency department
93.7% (n=13,317) efficiency
92.4% (n=11,747) efficiency
98.8% (n=4981) efficiency

Table 1. Usability and satisfaction assessment (n=50). Likert-scale responses range from 1 to 5: 1=totally disagree, 2=partially disagree, 3=indifferent, 4=partially agree, and 5=totally agree.

Item	Median (IQR)	Mean
System screens can be easily understood.	5 (5-5)	4.9
The system allows recording all relevant data on patient consultation.	5 (4-5)	4.64
By following the screen prompts, I was able to provide patient care with quality.	5 (5-5)	4.86
The system fields are easy to fill out.	5 (5-5)	4.84
The system is intuitive to use.	5 (4-5)	4.42
I believe that the system can be useful in clinical practice, for the care of patients with suspected COVID-19.	5 (5-5)	4.96
The system is stable, and no errors occur during use.	4 (2-4)	3.3
I was satisfied with the use of the system.	5 (4-5)	4.64

Discussion

Principal Findings

Our study presents a novel telehealth tool from its conceptualization through its development, validation, implementation, and rapid expansion. When planning the teleconsultation system, 2 major barriers to the implementation were identified. First, the lack of local experience with the functionalities needed for synchronous teleconsultation. Despite the TNMG’s long experience with other telehealth tools, it was only after the spread of COVID-19 that a legal and regulatory framework emerged, authorizing remote medical and other professional health consultations in Brazil [13]. Second, as COVID-19 was a new disease, information about it was still scarce. Due to the successive emergence of new evidence, the system’s initial matrix had to be progressively changed over time. Thus, a continuous development and validation process was of utmost importance to guarantee that the system was kept in line with updated evidence.

The use of the TeleCOVID-MG system made it possible to clarify queries about the novel coronavirus and deliver remote care to thousands of patients, thus reducing the circulation of individuals with respiratory or flu-like symptoms, minimizing the burden on health services, and increasing patient access to care in places with scarce health resources; together, these possibilities evidence the critical role of telemedicine in expanding emergency services capacity during a pandemic. The system also contributed to the updating of several health care professionals on the main topics related to COVID-19.

Doubts and concerns about the use of teleconsultations, especially teleconsultations performed by telephone calls (which was the most frequently used medium in our context), were already present even before the pandemic. Impossibility to perform the physical examination, compromised physician-patient relationship, difficulty in performing a global assessment of the patient that only focus on acute complaints, and uncertainty in the quality of information are some of the challenges reported in studies that evaluated the perception of health care professionals about the use of teleconsultations [14-17]. With the pandemic, providers and patients were forced into a new normal that included communicating with each other through video and audio [18], and studies have demonstrated that the COVID-19 pandemic affected the way physicians use and perceive telehealth and increased telehealth activities use both in type and frequency [19] despite the aforementioned limitations. Through their experience during the pandemic, physicians became more convinced of the efficacy, efficiency, and safety of telemedicine, as well as their ability to meet their patients' needs remotely. Although there was a shift in physicians' activities and perceptions, concerns about the effectiveness of remote consults and the lack of adequate legal frameworks remain [17]. Negative aspects related to teleconsultations reported in the literature include concerns about the absence of visual clues, inability to perform a physical examination, and thus the lack of comprehensive assessments [17,18].

Health care professionals with no experience in telehealth needed to quickly develop skills in web-based rapport building [19]; therefore, assessing the usability and provider satisfaction of each implemented system is of utmost importance. The analysis of usability and satisfaction of health care professionals with our system showed that most of them agreed that the system is intuitive and easy to understand and operate; allows them to provide care with quality; and is useful for evaluating patients with COVID-19. The social function of the systems was highlighted for the way it guaranteed the expansion of access to health care and decreased the burden in local health care. In addition, the systems allow interdisciplinarity and the development of a continuum of care until the patient's complete recovery.

Our results are in line with other studies [4,20-22], which showed high levels of satisfaction with telemedicine implemented during the pandemic. A recent integrative review has found 5 studies assessing provider satisfaction, all of them in outpatient clinics for specialized care during the pandemic for other conditions, and satisfaction ranged from 78% to 93% among the studies [23]. The evidence presented here suggests the feasibility of incorporating synchronous teleconsultations for the management of other health conditions. For this application to be possible, we emphasize the need for constant improvements in the systems and the importance of integrating remote care with face-to-face care.

Bearing in mind that the uptake and sustainability of telehealth interventions are the ultimate goals when implementing them, we highlight the following as takeaway lessons:

- Previous expertise is important for the successful development of a new system, particularly when implementation within a short amount of time is needed;
- The engagement of end users, in this case health care professionals, in system design and development is of utmost importance to ensure the fulfillment of user needs and usability;
- Health care professionals' perception of telehealth was positively impacted by the pandemic setting, as shown by their reported high levels of satisfaction; and
- In remote or resource-constrained locations with unstable internet, having an alternative way to perform teleconsultation (such as using telephone calls) is of utmost importance.

The main challenges faced in the usability of the TeleCOVID-MG system were related to the instability of the local telephone network, the need to align the clinical guideline developed for remote care with the practices carried out in the municipalities, and continually adjusting the system to the new scientific evidence and practices arising through the course of the pandemic.

As limitations of the TeleCOVID-MG system, we should remark that the lack of integration with data from face-to-face assistance were reported. The need for an interoperable health care system became blatantly evident worldwide during the COVID-19 pandemic to avoid duplicating work and improve decision-making. Although not designed for interoperability, the system architecture allowed the on-demand generation of customized queries and reports.

With great growth in the use of teleconsultations as a way to fight the pandemic, several entities have published guidelines to help health care professionals in remote patient care [24]. Furthermore, studies have been published focusing on evaluating the use of this telehealth tool and proposing adjustments for expansion in the postpandemic period [18,20].

Limitations

With regards to the efficiency assessment, although the team performed a thorough assessment of referrals and nonreferrals, there might be cases in which patients did seek face-to-face care despite not having been recommended to do so. Patient and caregiver experience, as well as patient digital literacy and satisfaction with the TeleCOVID-MG service, has not been formally addressed yet. We opted to restrict our analysis to health care professionals due to 2 main reasons: (1) they had to adapt their work routine very quickly due to the pandemic; and (2) they were the primary users of the teleconsultation system, as they had to fill out the patients' electronic record and issue medical prescriptions, reports, and orders for diagnostic COVID-19 tests through the system. Despite the lack of formal assessment with patients, the assisting health care professionals reported spontaneous comments from patients on how they felt welcomed and listened to in a better way than in face-to-face consultation, as they had time to report everything they wanted, without the time constraints present in face-to-face consultations. This finding supports the idea that it is indeed possible to provide humanized care in telehealth. We are currently conducting a formal patient satisfaction analysis for TeleCOVID-MG.

Due to the pandemic scenario and the goal of including as many health care professionals who were using the system as possible in our usability and satisfaction assessment, our analysis was centered mainly on our questionnaire. A more thorough analysis about satisfaction drivers using a more in-depth qualitative study could provide additional lessons.

Conclusion

This paper described the rapid development, implementation, and expansion of the TeleCOVID-MG system, as well as the results of our usability and satisfaction assessment with health care professionals. The system made it possible to answer queries about COVID-19 and provide remote care to thousands of patients, showing the critical role of telemedicine in expanding emergency services capacity during a pandemic. The dynamic nature of the current pandemic required regular updates in the system and frequent monitoring of the implemented actions. The experience reported here is expected to inform telemedicine strategies to be implemented in a postpandemic scenario, not only to deal with eventual new pandemics but also, and most importantly, explore the affordances of telemedicine to enhance public policies aimed at promoting health care prevention, treatment, and education. Furthermore, our experience illustrates the local and cultural challenges and specificities that need to be dealt with in the development of such systems, which indicate that even if there were "off-the-shelf" solutions available, they might not be able to address local community needs.

Acknowledgments

This study is supported in part by the Brazilian research agencies Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, grant 88887.516155 / 2020-00) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG; APQ-00208-20). ALPR was supported in part by the Brazilian research agencies Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; grant 310790/2021-2 and IATS 465518/2014-1) and FAPEMIG (Programa Pesquisador Mineiro PPM-00428-17 and Rede de Teleassistência de Minas Gerais RED-00081-16). MS Marcolino was supported in part by CNPq (grant 310561/2021-3). AR and MS Marcolino are members of the National Institute of Science and Technology for Health Technology. AP has a CNPq research grant (313103/2021-6) and a FAPEMIG grant (APQ-01.461-14). TCF has a CAPES postdoctoral grant (88887.508597/2020-00). BAC was supported by a Pró-Reitoria de Pesquisa (PRPq)/Universidade Federal de Minas Gerais (UFMG)/Secretaria de Educação Superior (SESu)/Ministério da Educação (MEC) grant (23072.211119/2020-10). MCR has a CAPES Master's scholarship (88887.516155 / 2020-00).

Data Availability

Data is available upon reasonable request.

Authors' Contributions

ALPR, MS Marcolino, MBMA, MCP, MS Mendes, and CRAO were responsible for the research protocol and coordinated the study. LBR, MS Marcolino, MS Mendes, and MBMA participated in the TeleCOVID-MG development and testing. MBMA, MS Mendes, LBR, MCP, GMR, CSC, and INB participated in the TeleCOVID-MG implementation. RP and BAC were responsible for the coordination of the research team for the application of the questionnaire. MS Marcolino, AP, TCF, RP, BAC, CRAO, CSC, MS Mendes, ZSNR, and MCR performed the data analysis and drafted the manuscript. All authors reviewed and edited the manuscript. MCR organized the manuscript references. All authors approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary file.
[DOCX File , 152 KB - medinform_v10i12e37591_app1.docx]

References

1. Ibn-Mohammed T, Mustapha KB, Godsell J, Adamu Z, Babatunde KA, Akintade DD, et al. A critical analysis of the impacts of COVID-19 on the global economy and ecosystems and opportunities for circular economy strategies. Resour Conserv Recycl 2021 Jan;164:105169 [FREE Full text] [doi: 10.1016/j.resconrec.2020.105169] [Medline: 32982059]
2. Barbosa W, Zhou K, Waddell E, Myers T, Dorsey ER. Improving access to care: telemedicine across medical domains. Annu Rev Public Health 2021 Apr 01;42(1):463-481 [FREE Full text] [doi: 10.1146/annurev-publhealth-090519-093711] [Medline: 33798406]
3. 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]
4. Marcolino MS, Alkmim MB, Santos TADQ, Riberio AL. The Telehealth Network of Minas Gerais: a large-scale Brazilian public telehealth service improving access to specialised health care. Policy in Focus 2016;13(1):59-61.

5. Alkmim MB, Figueira RM, Marcolino MS, Cardoso CS, Pena de Abreu M, Cunha LR, et al. Improving patient access to specialized health care: the Telehealth Network of Minas Gerais, Brazil. Bull World Health Organ 2012 May 01;90(5):373-378 [FREE Full text] [doi: 10.2471/BLT.11.099408] [Medline: 22589571]
6. Garcia-Huidobro D, Rivera S, Valderrama Chang S, Bravo P, Capurro D. System-wide accelerated implementation of telemedicine in response to COVID-19: mixed methods evaluation. J Med Internet Res 2020 Oct 06;22(10):e22146 [FREE Full text] [doi: 10.2196/22146] [Medline: 32903195]
7. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061 [FREE Full text] [doi: 10.1136/bmj.n2061] [Medline: 34593508]
8. Protocolo de manejo clínico do coronavírus (COVID-19) na atenção primária à saùde. Coronavirus Clinical Management Protocol (COVID-19) in Primary Health Care. Article in Portuguese. Ministry of Health, Brazil. 2020 Apr. URL: https://saude.rs.gov.br/upload/arquivos/202004/14140606-4-ms-protocolomanejo-aps-ver07abril.pdf [accessed 2021-08-02]
9. Ministry of Health, Minister's Office, Brazil. Portaria nº 467, de 20 de março de 2020. Ordinance No. 467, of March 20, 2020. Article in Portuguese. Diário Oficial da União. 2020 Mar 20. URL: https://www.in.gov.br/en/web/dou/-/portaria-n-467-de-20-de-marco-de-2020-249312996 [accessed 2021-08-02]
10. Cateb GF, Amaral S, Gonçalves SCL, Oliveira IJR, Prates RO, Chagas BA, et al. Estudo piloto de validação de um chatbot de rastreamento, implementado para direcionar a teleassistência em COVID-19. 2021 Jun 15 Presented at: SBCAS 2021: XXI Simpósio Brasileiro de Computação Aplicada à Saúde; June 15-18, 2021; Online event p. 97-102. [doi: 10.5753/sbcas.2021.16108]
11. Chagas BA, Ferreguetti K, Ferreira TC, Prates RO, Ribeiro LB, Pagano AS, et al. Chatbot as a telehealth intervention strategy in the COVID-19 pandemic: lessons learned from an action research approach. CLEI Electronic Journal 2021 Dec 13;24(3):6 [FREE Full text] [doi: 10.19153/cleiej.24.3.6]
12. Brazilian Census 2010. Brazilian Institute of Geography and Statistics. 2010. URL: https://cidades.ibge.gov.br/brasil/mg/divinopolis/panorama [accessed 2021-08-02]
13. Presidência da República (Brasil). Lei nº 989, de 15 de abril de 2020: dispõe sobre o uso da telemedicina durante a crise causada pelo coronavírus (SARS-CoV-2). Diário Oficial da União. 2020 Apr 15. URL: https://www.in.gov.br/en/web/dou/-/lei-n-13.989-de-15-de-abril-de-2020-252726328 [accessed 2020-12-19]
14. Caetano R, Silva AB, Guedes ACCM, Paiva CCN, Ribeiro GDR, Santos DL, et al. Challenges and opportunities for telehealth during the COVID-19 pandemic: ideas on spaces and initiatives in the Brazilian context. Cad Saude Publica 2020;36(5):e00088920 [FREE Full text] [doi: 10.1590/0102-311x00088920] [Medline: 32490913]
15. Derkx HP, Rethans JJE, Maiburg BH, Winkens RA, Muijtjens AM, van Rooij HG, et al. Quality of communication during telephone triage at Dutch out-of-hours centres. Patient Educ Couns 2009 Feb;74(2):174-178. [doi: 10.1016/j.pec.2008.08.002] [Medline: 18845413]
16. McKinstry B, Hammersley V, Burton C, Pinnock H, Elton R, Dowell J, et al. The quality, safety and content of telephone and face-to-face consultations: a comparative study. Qual Saf Health Care 2010 Aug 29;19(4):298-303. [doi: 10.1136/qshc.2008.027763] [Medline: 20430933]
17. Hjelm NM. Benefits and drawbacks of telemedicine. J Telemed Telecare 2005 Jun 24;11(2):60-70. [doi: 10.1258/1357633053499886] [Medline: 15829049]
18. Hasani SA, Ghafri TA, Al Lawati H, Mohammed J, Al Mukhainai A, Al Ajmi F, et al. The use of telephone consultation in primary health care during COVID-19 pandemic, Oman: perceptions from physicians. J Prim Care Community Health 2020 Dec 14;11:2150132720976480 [FREE Full text] [doi: 10.1177/2150132720976480] [Medline: 33307943]
19. Mann DM, Chen J, Chunara R, Testa PA, Nov O. COVID-19 transforms health care through telemedicine: evidence from the field. J Am Med Inform Assoc 2020 Jul 01;27(7):1132-1135 [FREE Full text] [doi: 10.1093/jamia/ocaa072] [Medline: 32324855]
20. Helou S, El Helou E, Abou-Khalil V, Wakim J, El Helou J, Daher A, et al. The effect of the COVID-19 pandemic on physicians' use and perception of telehealth: the case of Lebanon. Int J Environ Res Public Health 2020 Jul 06;17(13):4866 [FREE Full text] [doi: 10.3390/ijerph17134866] [Medline: 32640652]
21. Darr A, Senior A, Argyriou K, Limbrick J, Nie H, Kantczak A, et al. The impact of the coronavirus (COVID-19) pandemic on elective paediatric otolaryngology outpatient services - an analysis of virtual outpatient clinics in a tertiary referral centre using the modified paediatric otolaryngology telemedicine satisfaction survey (POTSS). Int J Pediatr Otorhinolaryngol 2020 Nov;138:110383 [FREE Full text] [doi: 10.1016/j.ijporl.2020.110383] [Medline: 33152974]
22. Dobrusin A, Hawa F, Gladshteyn M, Corsello P, Harlen K, Walsh CX, et al. Gastroenterologists and patients report high satisfaction rates with telehealth services during the novel coronavirus 2019 pandemic. Clin Gastroenterol Hepatol 2020 Oct;18(11):2393-2397.e2 [FREE Full text] [doi: 10.1016/j.cgh.2020.07.014] [Medline: 32663521]
23. Andrews E, Berghofer K, Long J, Prescott A, Caboral-Stevens M. Satisfaction with the use of telehealth during COVID-19: an integrative review. Int J Nurs Stud Adv 2020 Nov;2:100008 [FREE Full text] [doi: 10.1016/j.ijnsa.2020.100008] [Medline: 33083791]
24. Greenhalgh T, Koh GCH, Car J. COVID-19: a remote assessment in primary care. BMJ 2020 Mar 25;368:m1182. [doi: 10.1136/bmj.m1182] [Medline: 32213507]

Abbreviations

TNMG: Telehealth Network of the State of Minas Gerais
UFMG: Universidade Federal de Minas Gerais

Edited by C Lovis; submitted 28.02.22; peer-reviewed by M Mars, P Santana-Mancilla; comments to author 22.04.22; revised version received 13.05.22; accepted 06.06.22; published 22.12.22.

Please cite as:
Marcolino MS, Diniz CS, Chagas BA, Mendes MS, Prates R, Pagano A, Ferreira TC, Alkmim MBM, Oliveira CRA, Borges IN, Raposo MC, Reis ZSN, Paixão MC, Ribeiro LB, Rocha GM, Cardoso CS, Ribeiro ALP
Synchronous Teleconsultation and Monitoring Service Targeting COVID-19: Leveraging Insights for Postpandemic Health Care
JMIR Med Inform 2022;10(12):e37591
URL: https://medinform.jmir.org/2022/12/e37591
doi:10.2196/37591
PMID:36191175

Original Paper

A Framework for Modeling and Interpreting Patient Subgroups Applied to Hospital Readmission: Visual Analytical Approach

Suresh K Bhavnani[1], MArch, PhD; Weibin Zhang[1], PhD; Shyam Visweswaran[2], MD, PhD; Mukaila Raji[3], MSc, MD; Yong-Fang Kuo[1], PhD

[1]School of Public and Population Health, University of Texas Medical Branch, Galveston, TX, United States

[2]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, United States

[3]Division of Geriatric Medicine, Department of Internal Medicine, University of Texas Medical Branch, Galveston, TX, United States

Corresponding Author:
Suresh K Bhavnani, MArch, PhD
School of Public and Population Health
University of Texas Medical Branch
Institute for Translational Sciences
301 University Blvd.
Galveston, TX, 77555-0129
United States
Phone: 1 (409) 772 1928
Email: subhavna@utmb.edu

Abstract

Background: A primary goal of precision medicine is to identify patient subgroups and infer their underlying disease processes with the aim of designing targeted interventions. Although several studies have identified patient subgroups, there is a considerable gap between the identification of patient subgroups and their modeling and interpretation for clinical applications.

Objective: This study aimed to develop and evaluate a novel analytical framework for modeling and interpreting patient subgroups (MIPS) using a 3-step modeling approach: *visual analytical* modeling to automatically identify patient subgroups and their co-occurring comorbidities and determine their statistical significance and clinical interpretability; *classification* modeling to classify patients into subgroups and measure its accuracy; and *prediction* modeling to predict a patient's risk of an adverse outcome and compare its accuracy with and without patient subgroup information.

Methods: The MIPS framework was developed using bipartite networks to identify patient subgroups based on frequently co-occurring high-risk comorbidities, multinomial logistic regression to classify patients into subgroups, and hierarchical logistic regression to predict the risk of an adverse outcome using subgroup membership compared with standard logistic regression without subgroup membership. The MIPS framework was evaluated for 3 hospital readmission conditions: chronic obstructive pulmonary disease (COPD), congestive heart failure (CHF), and total hip arthroplasty/total knee arthroplasty (THA/TKA) (COPD: n=29,016; CHF: n=51,550; THA/TKA: n=16,498). For each condition, we extracted cases defined as patients readmitted within 30 days of hospital discharge. Controls were defined as patients not readmitted within 90 days of discharge, matched by age, sex, race, and Medicaid eligibility.

Results: In each condition, the visual analytical model identified patient subgroups that were statistically significant (Q=0.17, 0.17, 0.31; P<.001, <.001, <.05), significantly replicated (Rand Index=0.92, 0.94, 0.89; P<.001, <.001, <.01), and clinically meaningful to clinicians. In each condition, the classification model had high accuracy in classifying patients into subgroups (mean accuracy=99.6%, 99.34%, 99.86%). In 2 conditions (COPD and THA/TKA), the hierarchical prediction model had a small but statistically significant improvement in discriminating between readmitted and not readmitted patients as measured by net reclassification improvement (0.059, 0.11) but not as measured by the C-statistic or integrated discrimination improvement.

Conclusions: Although the visual analytical models identified statistically and clinically significant patient subgroups, the results pinpoint the need to analyze subgroups at different levels of granularity for improving the interpretability of intra- and intercluster associations. The high accuracy of the classification models reflects the strong separation of patient subgroups, despite the size and density of the data sets. Finally, the small improvement in predictive accuracy suggests that comorbidities alone were not strong predictors of hospital readmission, and the need for more sophisticated subgroup modeling methods. Such advances could improve the interpretability and predictive accuracy of patient subgroup models for reducing the risk of hospital readmission, and beyond.

(JMIR Med Inform 2022;10(12):e37239) doi:10.2196/37239

KEYWORDS

visual analytics; Bipartite Network analysis; hospital readmission; precision medicine; modeling; Medicare

Introduction

Overview

A wide range of studies [1-9] on topics ranging from molecular to environmental determinants of health have shown that most humans tend to share a subset of characteristics (eg, comorbidities, symptoms, or genetic variants), forming distinct patient subgroups. A primary goal of precision medicine is to identify such patient subgroups, and to infer their underlying disease processes to design interventions targeted at those processes [2,10]. For example, recent studies on complex diseases such as breast cancer [3,4], asthma [5-7], and COVID-19 [11] have revealed patient subgroups, each with different underlying mechanisms precipitating the disease and therefore each requiring different interventions.

However, there is a considerable gap between the identification of patient subgroups and their modeling and interpretation for clinical applications. To bridge this gap, we developed and evaluated a novel analytical framework called modeling and interpreting patient subgroups (MIPS) using a 3-step modeling approach: (1) identification of patient subgroups, their frequently co-occurring characteristics, and their risk of adverse outcomes; (2) classification of a new patient into one or more subgroups; and (3) prediction of an adverse outcome for a new patient informed by subgroup membership. We evaluated MIPS on 3 data sets related to hospital readmission, which helped pinpoint the strengths and limitations of MIPS. Furthermore, the results provided implications for improving the interpretability of patient subgroups in large and dense data sets, and for the design of clinical decision support systems to prevent adverse outcomes such as hospital readmissions.

Identification of Patient Subgroups

Patients have been divided into subgroups using (1) investigator-selected variables such as race for developing hierarchical regression models [12] or assigning patients to different arms of a clinical trial, (2) existing classification systems such as the Medicare Severity-Diagnosis Related Group [13] to assign patients to a disease category for purposes of billing, and (3) computational methods such as classification [14-16] and clustering [5,17] to discover patient subgroups from data (also referred to as *subtypes* or *phenotypes* depending on the condition and variables analyzed).

Several studies have used a wide range of computational methods to identify patient subgroups, each with critical trade-offs. Some studies have used *combinatorial* approaches [18] (identifying all pairs, all triples, etc), which although intuitive, can lead to a combinatorial explosion (eg, enumerating combinations of the 31 Elixhauser comorbidities would lead to 2^{31} or 2147483648 combinations), with most combinations that do not incorporate the full range of symptoms (eg, the most frequent pair of symptoms ignores which other symptoms exist in the profile of patients with that pair). Other studies have used *unipartite* clustering methods [16,17] (clustering patients or comorbidities but not both together), such as k-means and hierarchical clustering. Furthermore, dimensionality-reduction methods such as principal component analysis used with unipartite clustering methods have been used to identify clusters of frequently co-occurring comorbidities [18-24]. However, such methods have well-known limitations, including the requirement of inputting user-selected parameters (eg, similarity measures and the number of expected clusters) and the lack of a quantitative measure to describe the quality of the clustering (critical for measuring the statistical significance of the clustering). Furthermore, because these methods are unipartite, there is no agreed-upon method for identifying the patient subgroup defined by a cluster of variables, and vice versa.

More recently, bipartite network analysis [25] has been used to address these limitations by automatically identifying *biclusters*, consisting of patients and characteristics simultaneously. This method takes as input any data set, such as patients and their comorbidities, and outputs a quantitative and visual description of biclusters (containing both patient subgroups and their frequently co-occurring comorbidities). The quantitative output generates the number, size, and statistical significance of the biclusters [26-28], and the visual output displays the quantitative information of the biclusters through a network visualization [29-31]. Bipartite network analysis therefore enables (1) the automatic identification of biclusters and their significance and (2) the visualization of the biclusters critical for their clinical interpretability. Furthermore, the attributes of patients in a subgroup can be used to measure the subgroup risk for an adverse outcome, develop classification models for classifying a new patient into one or more of the subgroups, and develop prediction models that use subgroup membership for measuring the risk of an adverse outcome for the classified patient.

However, although several studies [11,28,32-38] have demonstrated the usefulness of bipartite networks for the identification and clinical interpretation of patient subgroups, there has been no systematic attempt to integrate them with classification and prediction modeling, which is a critical step toward their clinical application. Therefore, we leveraged the advantages of a bipartite network to develop the MIPS framework with the goal of bridging the gap between the identification of patient subgroups, and their modeling and interpretation for future clinical applications.

The Need for Modeling and Interpreting Patient Subgroups in Hospital Readmission

An estimated 1 in 5 elderly patients (more than 2.3 million Americans) is readmitted to a hospital within 30 days of discharge [39]. Although many readmissions are unavoidable, an estimated 75% of readmissions are unplanned and mostly preventable [40], imposing a significant burden in terms of mortality, morbidity, and resource consumption. Across all conditions, unplanned readmissions in the United States cost approximately US $17 billion [40], making them an ineffective use of costly resources. Consequently, hospital readmission is closely scrutinized as a marker for poor quality of care by organizations such as the Centers for Medicare & Medicaid Services (CMS) [41].

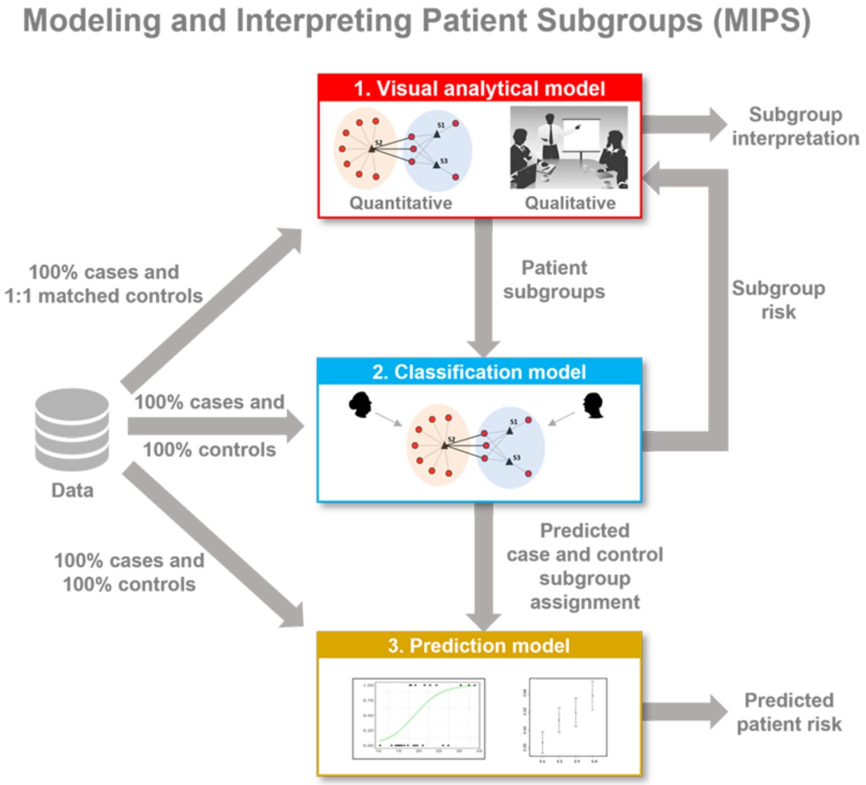
To address this epidemic of hospital readmission, CMS sponsored the development of models to predict the patient-specific risk of readmission in specific index conditions such as chronic obstructive pulmonary disease (COPD) [42], congestive heart failure (CHF) [43], and total hip arthroplasty/total knee arthroplasty (THA/TKA) [44]. As numerous studies have shown that almost two-thirds of older adults have 2 or more comorbid conditions with a heightened risk of adverse health outcomes [18], the independent variables in the CMS models included prior comorbidities (as recorded in Medicare claims data) and demographics (age, sex, and race). However, although prior studies have shown the existence of subgroups among patients with hospital readmission [28], none of the CMS models have incorporated patient subgroups. The identification and inclusion of patient subgroups could improve the accuracy of predicting hospital readmission for a patient, in addition to enabling the design of interventions targeted at each patient subgroup to reduce the risk of readmission. Therefore, we used the MIPS framework to model and interpret patient subgroups in hospital readmission and tested its generality across the 3 index conditions. Furthermore, to enable a head-to-head comparison with existing CMS predictive models, we used the same independent variables as were used in those models, in addition to patient subgroup membership when developing our prediction models.

Methods

Overview

Figure 1 provides a conceptual description of the data inputs and outputs from the 3-step modeling in MIPS. The visual analytical model identifies patient subgroups and visualizes them through a network. The classification model determines the subgroup membership for cases and controls. These subgroup memberships are then used to measure the risk for readmission within each subgroup based on the proportion of cases and juxtaposed with the respective subgroup visualization to enable clinicians to interpret the readmitted patient subgroups. Finally, the prediction model uses the subgroup membership assignment of cases and controls to determine the readmission risk of a patient. Multimedia Appendix 1 [16,23,25-31,45,46] provides a summary of the inputs, methods, and outputs for each model.

Figure 1. Inputs and outputs for the 3-step modeling in MIPS consisting of the visual analytical model, classification model, and prediction model. MIPS: Modeling and Interpreting Patient Subgroups.


Modeling and Interpreting Patient Subgroups (MIPS)
1. Visual analytical model
Quantitative
Qualitative
Subgroup interpretation
100% cases and 1:1 matched controls
Patient subgroups
Subgroup risk
100% cases and
100% controls
Data
2. Classification model
100% cases and 100% controls
Predicted case and control subgroup assignment
3. Prediction model
Predicted patient risk

Data Description

Study Population

We analyzed patients hospitalized for COPD, CHF, or THA/TKA. We selected these 3 index conditions because (1) hospitalizations for each of these conditions are highly prevalent in older adults [39], (2) hospitals report very high variations in their readmission rates [39], and (3) there exist well-tested readmission prediction models for each of these conditions that do not consider patient subgroups [42-44,47,48].

Data for these 3 index conditions were extracted from the Medicare insurance claims data set. In 2019, Medicare provided health insurance to approximately 64.4 million Americans, of whom 55.5 million were older Americans (≥65 years) [49]. Furthermore, 94% of noninstitutionalized older Americans were covered by Medicare [50], with eligible claims received from 6204 medical institutions across the United States, and is therefore one of the few data sets that is highly representative of older Americans and their care.

For each index condition, we used the same inclusion and exclusion criteria that were used to develop the CMS models but with the most recent years (2013-2014) provided by Medicare when we started the project. We extracted all patients who were admitted to an acute care hospital between July 2013 and August 2014 with a principal diagnosis of the index condition, were aged ≥66 years, and were enrolled in both Medicare parts A and B fee-for-service plans 6 months before admission. Furthermore, we excluded patients who were transferred from other facilities, died during hospitalization, or transferred to another acute care hospital. Similar to the CMS models, we selected the first admission for patients with multiple admissions during the study period, and we did not use data from Medicare Part D (related to prescription medications).

Multimedia Appendix 2 [40,44] describes (1) the International Classification of Diseases, Ninth Version, codes for each of the 3 index conditions selected for analysis and (2) the inclusion and exclusion criteria used to extract cases and controls for COPD, CHF, and THA/TKA; the respective numbers of patients extracted at each step; and how we addressed the small incidence of missing data. Each modeling method used relevant subsets of these data, as described in the Analytical and Evaluation Approach section.

Variables

The independent variables consisted of comorbidities and patient demographics (age, sex, and race). Comorbidities common in older adults were derived from 3 established comorbidity indices: Charlson Comorbidity Index [51], Elixhauser Comorbidity Index [52], and the Center for Medicare and Medicare Services Condition Categories used in the CMS readmission models [53] (the variables in the CMS models varied across the index conditions). As these indices had overlapping comorbidities, we derived a union of them, which was verified by the clinician stakeholders. They recommended that we also include the following additional variables, as they were pertinent to each index condition: COPD (history of sleep apnea and mechanical ventilation), CHF (history of coronary artery bypass graft surgery), and THA/TKA (congenital deformity of the hip joint and posttraumatic osteoarthritis). For each patient in our cohort, we extracted these comorbidities and variables from the physicians, outpatient, and inpatient Medicare claims data in the 6 months before (to guard against miscoding) and on the day of the index admission. The dependent variable (outcome) was whether a patient with an index admission (COPD, CHF, or THA/TKA) had an unplanned readmission to an acute care hospital within 30 days of discharge as was recorded in the Medicare Provider Analysis and Review file (inpatient claims) in the Medicare database.

Analytical and Evaluation Approach

Visual Analytical Modeling

The goal of visual analytical modeling was to identify and interpret biclusters of readmitted patients (cases), consisting of patient subgroups and their most frequently co-occurring comorbidities. The data used to build the visual analytical model in each index condition consisted of randomly dividing 100% of the cases into training (50%) and replication (50%) data sets (we use the term *replication* to avoid confusion with the term *validation* typically used in classification and prediction models). For feature selection, we extracted an equal number of 1:1 matched controls based on age, sex, race, and ethnicity, and Medicaid eligibility [45]. These data were analyzed for each index condition using the following steps (Multimedia Appendix 1 provides additional details for each step):

1. *Model training*: to train the visual analytical model, we used feature selection to identify the set of comorbidities that were univariably significant in both the training and replication data sets and used bicluster modularity maximization [26,27] to identify the number, members, and significance of biclusters in the training data set.
2. *Model replication*: to test the replicability of the biclusters, we repeated the bicluster analysis on the replication data set and used the Rand Index (RI) [46] to measure the degree and significance of similarity in comorbidity co-occurrence between the 2 data sets.
3. *Model interpretation*: to enable clinical interpretation of the patient subgroups, we used the *Fruchterman-Reingold* [29] and *ExplodeLayout* [30,31] algorithms to visualize the network. Furthermore, based on a request from our clinician stakeholder team, for each bicluster, we ranked and displayed the comorbidity labels with their univariable odds ratios (ORs) for readmission (obtained from the feature selection mentioned earlier) and juxtaposed the readmission risk of the bicluster (obtained from the classification step discussed in the next section) onto the network visualization. Clinician stakeholders were asked to use the visualization to interpret patient subgroups, their mechanisms, and potential interventions to reduce the risk of readmission.

Classification Modeling

The goal of classification modeling was to classify all cases and controls from the entire Medicare data set into the biclusters identified from the visual analytical model. The resulting bicluster membership for all cases and controls was designed to (1) develop the predictive modeling described in the next section and (2) measure the risk of each subgroup to enable clinical interpretation of the patient subgroups. The training data set in each condition consisted of a random sample of 75% cases with their subgroup membership (output of the visual analytical modeling) and an internal validation data set consisting of randomly selected 25% of the cases (with subgroup membership used to validate the model). These data were used to develop and use classification models for each index condition using the following steps (Multimedia Appendix 1 provides additional details for each step):

1. *Model training*: to train the classifier, we used multinomial logistic regression [16] with independent variables consisting of comorbidities (identified through feature selection). The accuracy of the trained model was measured by calculating the percentage of times the model correctly classified the cases into subgroups using the highest predicted probability across the subgroups.
2. *Model internal validation*: to internally validate the classifier, we randomly split these data into training (75%) and testing (25%) data sets 1000 times. For each iteration, we trained a model using the training data set and measured its accuracy using the testing data set. This was done by predicting subgroup membership using the highest predicted probability among all the subgroups. The overall predicted accuracy was estimated by calculating the mean accuracy across the 1000 models.
3. *Model application*: to generate data for the visual analytical and prediction models, the classifier was used to classify 100% of cases and controls from our entire Medicare data set (July 2013-August 2014). The resulting classified data were used to measure the risk of each subgroup (juxtaposed onto the network visualization to enable clinical interpretation) and to conduct the following prediction modeling.

Prediction Modeling

The goal of prediction modeling was to predict the risk of readmission for a patient, taking into consideration subgroup membership. The data used to build the prediction models consisted 100% of cases and 100% of controls, with subgroup membership generated from the classification modeling. These data were randomly spilt into training (75%) and internal validation (25%) data sets. These data were used to train, internally validate, and compare the prediction models in each index condition using the following steps (Multimedia Appendix 1 provides additional details for each step):

1. *Model training*: to train the prediction model, we used binary logistic regression for developing a Standard Model (without subgroup membership, similar to the CMS models) and a Hierarchical Model (with subgroup membership). The independent variables for both models consisted of comorbidities (identified through feature selection) and demographics, and the outcome was 30-day unplanned readmission (yes vs no).
2. *Model internal validation*: to internally validate the models, we used the internal validation data set to measure discrimination (C-statistic) and calibration (calibration-in-the-large and calibration slope).
3. *Model comparisons*: to compare the accuracy of the Standard and Hierarchical Models, we used the chi-squared test to compare their C-statistics. Furthermore, to examine how the Standard Model was applied to each subgroup, we measured the C-statistics of the Standard Model applied to each subgroup separately. Finally, because both these models used comorbidities selected through feature selection, they differed from the set of comorbidities used in the published CMS models. Therefore, to perform a head-to-head comparison with the published CMS models (COPD [42], CHF [43], and THA/TKA [44]), we developed a logistic regression model using the independent variables from the published CMS model (CMS Standard Model) and compared it to the same model, but which also included subgroup membership (CMS Hierarchical Model). Similar to these comparisons, we used the chi-squared test to compare the C-statistics of the CMS standard and the CMS Hierarchical Models and additionally measured the differences between the models using net reclassification improvement (NRI) and integrated discrimination improvement (IDI).

Ethics Approval

Medicare data were analyzed using a CMS data-use agreement (CMS DUA RSCH-2017-51404) and approved by the University of Texas Medical Branch Institutional Review Board (16-0361).

Results

Data

Table 1 summarizes the number of cases and controls used to develop the 3 models for each condition.

Table 1. Training and replication/validation data sets used to develop the three models in each of the 3 index conditions.

Model	Training	Replication/validation	Total
Visual analytical[a] (cases/controls)			
Chronic obstructive pulmonary disease (COPD)	14,508/14,508	14,508/14,508	29,016/29,016
Congestive heart failure (CHF)	25,775/25,775	25,775/25,775	51,550/51,550
Total hip arthroplasty/total knee arthroplasty (THA/TKA)	8249/8249	8249/8249	16,498/16,948
Classification (cases)			
COPD	10,842	3615	14,457
CHF	19,254	6418	25,672
THA/TKA	5257	1753	7010
Prediction (cases/controls)			
COPD	21,692/117,839	7334/39,176	29,026/157,015
CHF	38,728/183,093	12,845/61,095	51,573/244,188
THA/TKA	12,376/255,203	41,44/85,049	16,520/340,252

[a] The visual analytical models used 1:1 matched controls for the feature selection, and used only cases for the bipartite networks to analyze heterogeneity in readmission. The numbers shown for the visual analytical models are before removing patients with no comorbidities. The resulting cases-only data sets were used for the classification modelling as shown.

Visual Analytical Modeling

Overview

Visual analytical modeling of readmitted patients in all 3 index conditions produced statistically and clinically significant patient subgroups and their most frequently co-occurring comorbidities, which were significantly replicated. We report the results for each index condition.

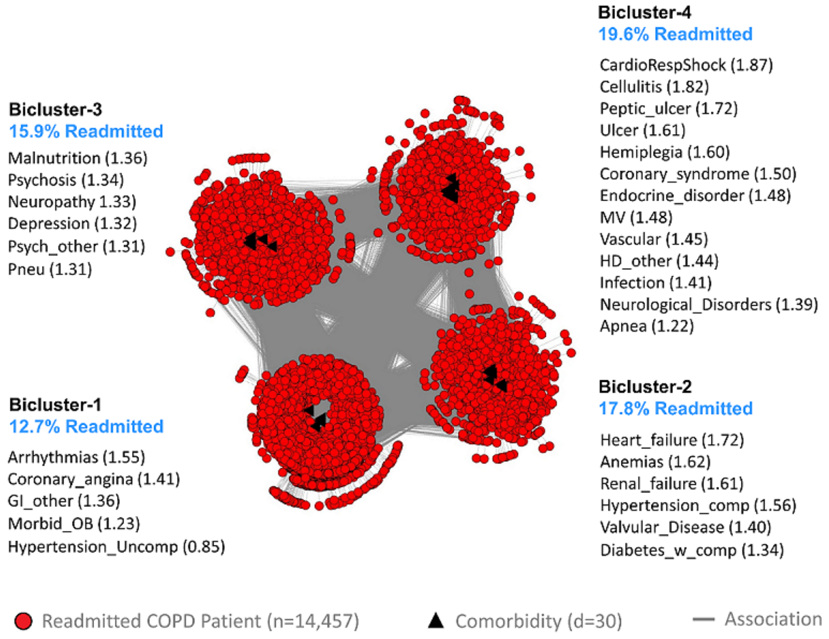
COPD Visual Analytical Model

The inclusion and exclusion selection criteria (Multimedia Appendix 2) resulted in a training data set (n=14,508 matched case-control pairs, of which 51 patient pairs had no dropped comorbidities) and a replication data set (n=14,508 matched case-control pairs, of which 51 patient pairs had no dropped comorbidities), matched by age, sex, race, and Medicaid eligibility (a proxy for economic status). The feature selection method (Multimedia Appendix 3) used 45 unique comorbidities identified from a union of the 3 comorbidity indices, plus 2 condition-specific comorbidities. Of these, 3 were removed because of <1% prevalence. Of the remaining comorbidities, 30 survived significance and replication testing using Bonferroni correction. The visual analytical model used these surviving comorbidities (d=30), and readmitted patients with COPD with at least one of these comorbidities (n=14,457).

As shown in Figure 2, bipartite network analysis identified 4 biclusters, each representing a subgroup of readmitted patients with COPD and their most frequently co-occurring comorbidities. Biclustering had significant modularity (*Q*=0.17; *z*=7.3; *P*<.001) and significant replication (RI=0.92; *z*=11.62; *P*<.001) of comorbidity co-occurrence. Furthermore, as requested by the clinician stakeholders, we juxtaposed a ranked list of comorbidities based on their ORs for readmission in each bicluster, in addition to the risk for each patient subgroup.

The pulmonologist inspected the visualization and noted that the readmission risk of the patient subgroups had a wide range (12.7%-19.6%) with clinical (face) validity. Furthermore, the co-occurrence of comorbidities in each patient subgroup was clinically meaningful with interpretations for each subgroup. Subgroup-1 had a low disease burden, with uncomplicated hypertension leading to the lowest risk (12.7%). This subgroup represented patients with early organ dysfunction and would benefit from using checklists such as regular monitoring of blood pressure in predischarge protocols to reduce the risk of readmission. Subgroup-3 had mainly psychosocial comorbidities, which could lead to aspiration precipitating pneumonia, leading to an increased risk for readmission (15.9%). This subgroup would benefit from early consultation with specialists (eg, psychiatrists, therapists, neurologists, and geriatricians) who have expertise in psychosocial comorbidities, with a focus on the early identification of aspiration risks and precautions. Subgroup-2 had diabetes with complications, renal failure, and heart failure and therefore had higher disease burden, leading to an increased risk of readmission (17.8%) compared with Subgroup-1. This subgroup had metabolic abnormalities with greater end-organ dysfunction and would therefore benefit from case management by advanced practice providers (eg, nurse practitioners) with rigorous adherence to established guidelines to reduce the risk of readmission. Subgroup-4 had diseases with end-organ damage, including gastrointestinal disorders, and therefore had the highest disease burden and risk for readmission (19.6%). This subgroup would also benefit from case management with rigorous adherence to established guidelines to reduce the risk of readmission. Furthermore, as patients in this subgroup typically experience complications that could impair their ability to make medical decisions, they should be provided with early consultation with a palliative care team to ensure that care interventions align with patients' preferences and values.

Figure 2. The chronic obstructive pulmonary disease (COPD) visual analytical model showing 4 biclusters consisting of patient subgroups and their most frequently co-occurring comorbidities (whose labels are ranked by their univariable odds ratios, shown within parentheses) and their risk of readmission (shown in blue text). GI: Gastrointestinal disorders; HD: Heart disease; MV: History of mechanical ventilation.

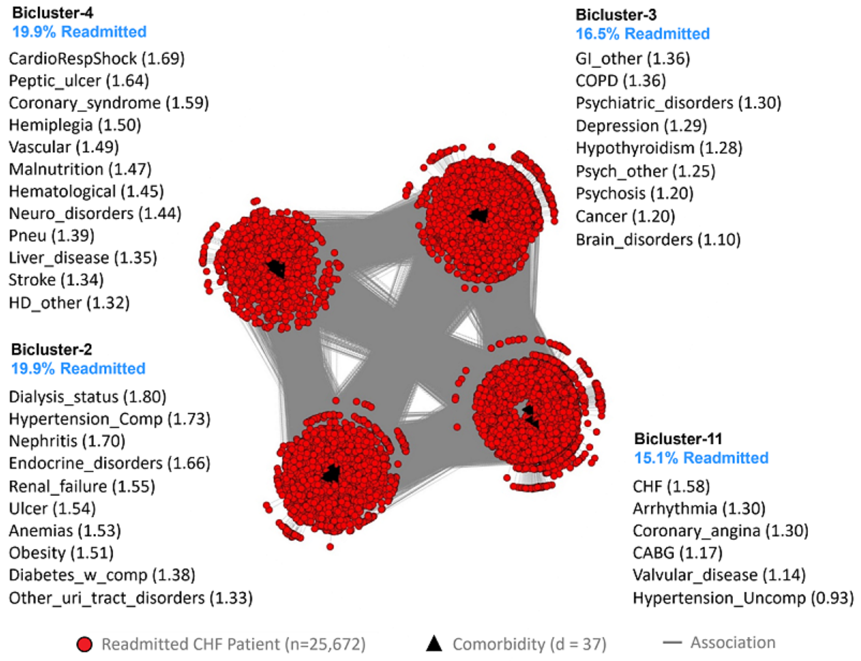


CHF Visual Analytical Model

The inclusion and exclusion selection criteria (Multimedia Appendix 2) resulted in a training data set (n=25,775 matched case-control pairs, of which 103 patient pairs with no dropped comorbidities) and a replication data set (n=25,775 matched case-control pairs, of which 104 patient pairs with no dropped comorbidities), matched by age, sex, race, and Medicaid eligibility (a proxy for economic status). The feature selection method (Multimedia Appendix 3) used 42 unique comorbidities identified from a union of the 3 comorbidity indices plus 1 condition-specific comorbidity. Of these, 1 comorbidity was removed because of <1% prevalence. Of those remaining, 37 survived the significance and replication testing with the Bonferroni correction. The visual analytical model (Figure 3) used these surviving comorbidities (d=37) and cases consisting of readmitted patients with CHF, with at least one of those comorbidities (n=25,672). As shown in Figure 3, the bipartite network analysis of the CHF cases identified 4 biclusters, each representing a subgroup of readmitted patients with CHF and their most frequently co-occurring comorbidities. The analysis revealed that the biclustering had significant modularity (*Q*=0.17; *z*=8.69; *P*<.001) and significant replication (RI=0.94; *z*=17.66; *P*<.001) of comorbidity co-occurrence. Furthermore, as requested by the clinicians, we juxtaposed a ranked list of comorbidities based on their ORs for readmission in each bicluster, in addition to the risk for each of the patient subgroups.

The geriatrician inspected the visualization and noted that the readmission risk of the patient subgroups, ranging from 15.1% to 19.9%, was wide, with clinical (face) validity. Furthermore, the co-occurrence of comorbidities in each patient subgroup was clinically significant. Subgroup-1 had chronic but stable conditions and therefore had the lowest risk for readmission (15.1%). Subgroup-3 had mainly psychosocial comorbidities but was not as clinically unstable or fragile compared with Subgroup-2 and Subgroup-4, and therefore had medium risk (16.6%). Subgroup-2 had severe chronic conditions, making them clinically fragile (with potential benefits from early palliative and hospice care referrals), and were therefore at high risk for readmission if nonpalliative approaches were used (19.9%). Subgroup-4 had severe acute conditions that were also clinically unstable, associated with substantial disability and care debility and therefore at high risk for readmission and recurrent intensive care unit use (19.9%).

Figure 3. The congestive heart failure (CHF) visual analytical model showing 4 biclusters consisting of patient subgroups and their most frequently co-occurring comorbidities (whose labels are ranked by their univariable odds ratios, shown within parentheses) and their risk of readmission (shown in blue text). CABG: History of coronary artery bypass graft surgery; COPD: Chronic obstructive pulmonary disease; GI: Gastrointestinal disorders; HD: Heart disease.



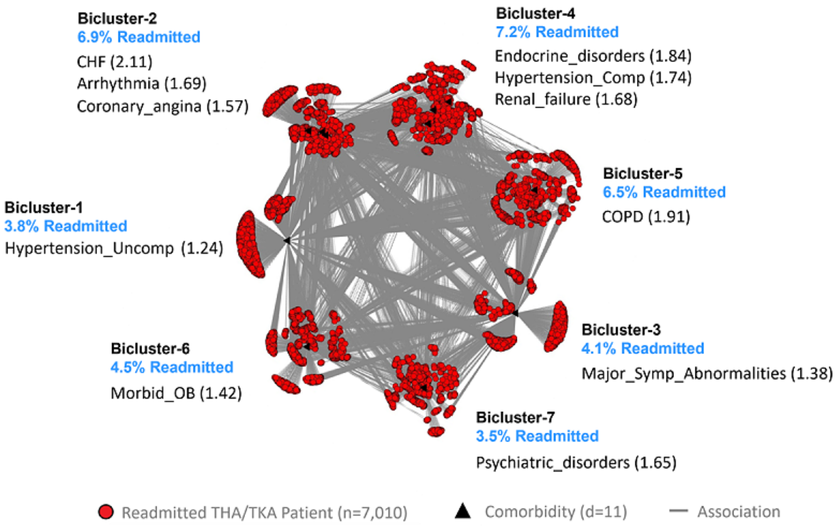
THA/TKA Visual Analytical Model

The inclusion and exclusion selection criteria (Multimedia Appendix 2) resulted in a training data set (n=8249 matched case-control pairs, of which 1239 patient pairs had no dropped comorbidities) and a replication data set (n=8249 matched case-control pairs, of which 1264 patient pairs had no dropped comorbidities), matched by age, sex, race, and Medicaid eligibility (a proxy for economic status). Feature selection (Multimedia Appendix 3) used 39 unique comorbidities identified from the 3 comorbidity indices plus 2 condition-specific comorbidities. Of these, 11 comorbidities were excluded because of <1% prevalence. Of the remaining, 11 comorbidities survived significance and replication testing with the Bonferroni correction. The visual analytical model (Figure 4) used these surviving comorbidities (d=11) and cases consisting of readmitted patients with at least one of those comorbidities (n=7010).

As shown in Figure 4, the bipartite network analysis of THA/TKA cases identified 7 biclusters, each representing a subgroup of readmitted patients with THA/TKA and their most frequently co-occurring comorbidities. The analysis revealed that biclustering had significant modularity (Q=0.31; z=2.52, P=.01), and significant replication (RI=0.89; z=3.15; P=.002) of comorbidity co-occurrence. Furthermore, as requested by the clinician stakeholders, we juxtaposed a ranked list of comorbidities based on their ORs for readmission in each bicluster, in addition to the risk for each patient subgroup.

The geriatrician inspected the network and noted that patients with total knee arthroplasty, in general, were healthier than patients with total hip arthroplasty. Therefore, the network was difficult to interpret when the 2 index conditions were merged together. Although our analysis was constrained because we used the conditions defined by CMS, these results nonetheless suggest that the interpretations did not suffer from a *confirmation bias* (manufactured interpretations to fit the results). However, he noted that the range of readmission risk had clinical (face) validity. Furthermore, Subgroup-2, Subgroup-4, and Subgroup-5 had more severe comorbidities related to the lung, heart, and kidney and therefore had a higher risk for readmission compared with Subgroup-1, Subgroup-6, and Subgroup-7, which had less severe comorbidities and therefore had a lower risk for readmission. In addition, Subgroup-2, Subgroup-5, Subgroup-6, and Subgroup-7 would benefit from chronic care case management from advanced practice providers (eg, nurse practitioners). Finally, Subgroup-2 and Subgroup-5 would benefit from using well-established guidelines for CHF and COPD, Subgroup-7 would benefit from mental health care and management of psychosocial comorbidities, and Subgroup-6 would benefit from care for obesity and metabolic disease management.

Figure 4. The total hip arthroplasty/total knee arthroplasty (THA/TKA) visual analytical model showing 4 biclusters consisting of patient subgroups and their most frequently co-occurring comorbidities (whose labels are ranked by their univariable odds ratios, shown within parentheses) and their risk for readmission (shown in blue text). CHF: Congestive heart failure; COPD: Chronic obstructive pulmonary disease; OB: Obesity.



Classification Modeling

Overview

The classification model used multinomial logistic regression for each index condition (Multimedia Appendix 4 for the model coefficients) to predict the membership of patients using subgroups (identified from the aforementioned visual analytical models). The results revealed that in each index condition, the classification model had high accuracy in classifying all the cases in the full data set (training data set used in the visual analytical modeling). Similarly, the internal validation results using a 75%:25% split of this data set also had a high classification accuracy (Table 2 with classification accuracy divided into quantiles). We report the results for each index condition.

Table 2. Internal validation results showing the percentage of chronic obstructive pulmonary disease (COPD) congestive heart failure (CHF), and total hip arthroplasty/total knee arthroplasty (THA/TKA) patients correctly-assigned to a subgroup by the classification models in each condition.

Models	Quantiles					Summary, mean (SD; range)
	Q 0.025	Q 0.25	Q 0.50	Q 0.75	Q 0.975	
COPD						
Training (n=10842)	99.90	100.00	100.00	100.00	100.00	100 (0.02; 99.7-100)
Testing (n=3615)	99.30	99.40	99.60	99.60	99.80	99.6 (0.15; 99.1-100)
CHF						
Training (n=19254)	99.40	99.50	99.60	99.60	99.80	99.57 (0.11; 99-99.9)
Testing (n=6418)	99.00	99.30	99.30	99.40	99.60	99.34 (0.15; 98.7-99.7)
THA/TKA						
Training (n=5257)	100.00	100.00	100.00	100.00	100.00	100 (0; 100-100)
Testing (n=1753)	99.70	99.80	99.90	99.90	100.00	99.86 (0.09; 99.4-100)

COPD Classification Model

The model correctly predicted subgroup membership for 99.9% (14,443/14,457) of the cases in the full data set. Furthermore, as shown in Table 2, the internal validation results revealed that the percentage of COPD cases correctly assigned to a subgroup in the testing data set ranged from 99.1% to 100%, with a median (Q.50 as shown in Table 2) of 99.6%, and with 95% being in the range of 99.3% to 99.8%.

CHF Classification Model

The model correctly predicted the subgroup membership for 99.2% (25,476/25,672) of the cases in the full data set. Furthermore, as shown in Table 2, the internal validation results revealed that the percentage of CHF cases correctly assigned to a subgroup in the testing data set ranged from 98.7% to 99.7%, with a median (Q.50) of 99.3%, and with 95% being in the range between 99% to 99.6%.

THA/TKA Classification Model

The model correctly predicted subgroup membership in 100% (7010/7010) of the cases in the full data set. Furthermore**,** as shown in Table 2, the internal validation results revealed that the percentage of CHF cases correctly assigned to a subgroup in the testing data set ranged from 99.4% to 100%, with a median (Q.50) of 99.9%, and with 95% being in the range of 99.7% to 100%.

Application of the Classification Model to Generate Information for Other Models

The classification model was used to classify 100% of cases and 100% of controls for use in the prediction model (described in the next section). Furthermore, the proportion of cases and controls classified into each subgroup was used to calculate the risk of readmission for the respective subgroup (Multimedia Appendix 3). As this subgroup risk information was requested by the clinicians to improve the interpretability of the visual analytical model, the risk was juxtaposed next to the respective subgroups in the bipartite network visualizations (see blue text in Figures 2-4).

Prediction Modeling

Overview

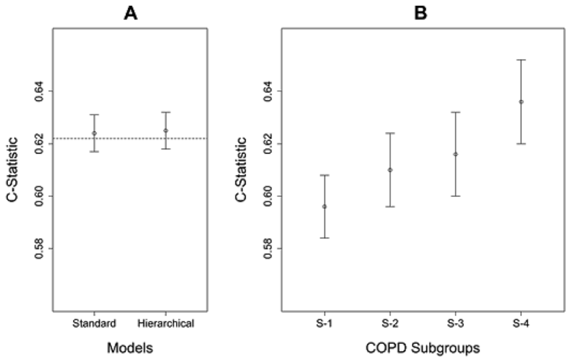
For each of the 3 index conditions, we developed 2 binary logistic regression models to predict readmission, with comorbidities in addition to sex, age, and race: (1) Standard Model representing all patients without subgroup membership, similar to the CMS models and (2) Hierarchical Model with an additional variable that adjusted for subgroup membership.

COPD Prediction Model

The inclusion and exclusion criteria (Multimedia Appendix 2) resulted in a cohort of 186,041 patients (29,026 cases and 157,015 controls). As shown in Figure 5A, the Standard Model had a C-statistic of 0.624 (95% CI 0.617-0.631) which was not significantly (*P*=.86) different from the Hierarchical Model that had a C-statistic of 0.625 (95% CI 0.618-0.632). The calibration plots revealed that both models had a slope close to 1 and an intercept close to 0 (Multimedia Appendix 5 [42-44]).

As shown in Figure 5B, the Standard Model was used to measure the predictive accuracy of patients in each subgroup. The results showed that Subgroup-1 had a lower C-statistic than Subgroup-3 and Subgroup-4. Although the C-statistics in Figures 5A and Figures 5B cannot be compared as they are based on models developed from different populations, these results reveal that the current CMS readmission model for CHF might be underperforming for a COPD patient subgroup, pinpointing which one might benefit from a Subgroup-Specific Model.

Figure 5. Predictive accuracy of the Standard Model compared with the Hierarchical Model in chronic obstructive pulmonary disease (COPD), as measured by the C-statistic. The C-statistic for the Centers for Medicare & Medicaid Services Standard Model is shown as a dotted line. (B) Predictive accuracy of the Standard Model when applied separately to patients classified to each subgroup. Subgroup-1 has lower accuracy than Subgroup-3 and Subgroup-4. (C-statistics in A and B cannot be compared, as they are based on models from different populations).

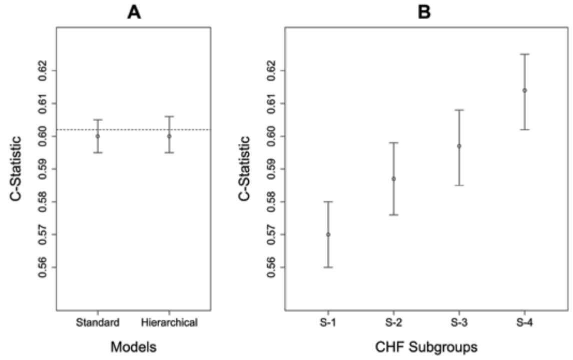


CHF Prediction Model

The inclusion and exclusion criteria (Multimedia Appendix 2) resulted in a cohort of 295,761 patients (51,573 cases and 244,188 controls). As shown in Figure 6A, the Standard Model had a C-statistic of 0.600 (95% CI 0.595-0.605), which was not significantly different (*P*=.29) from the Hierarchical Model, which also had a C-statistic of 0.600 (95% CI 0.595-0.606). The calibration plots revealed that all the models had a slope close to 1 and an intercept close to 0 (Multimedia Appendix 5).

As shown in Figure 6B, the Standard Model was used to measure the predictive accuracy of patients in each subgroup. The results showed that Subgroup-1 had a lower C-statistic than Subgroup-4. Although the C-statistics in Figures 6A and 6B cannot be compared as they are based on models developed from different populations, these results reveal that the current CMS readmission model for CHF might be underperforming for a CHF patient subgroup, pinpointing which one might benefit from a Subgroup-Specific model.

Figure 6. (A) Predictive accuracy of the Standard Model compared with the Hierarchical Model in congestive heart failure (CHF) as measured by the C-statistic. The C-statistic for the Centers for Medicare & Medicaid Services Standard Model is shown as a dotted line. (B) Predictive accuracy of the Standard Model when applied separately to patients classified to each subgroup. Subgroup-1 has lower accuracy than Subgroup-3 and Subgroup-4. (C-statistics in A and B cannot be compared, as they are based on models from different populations).

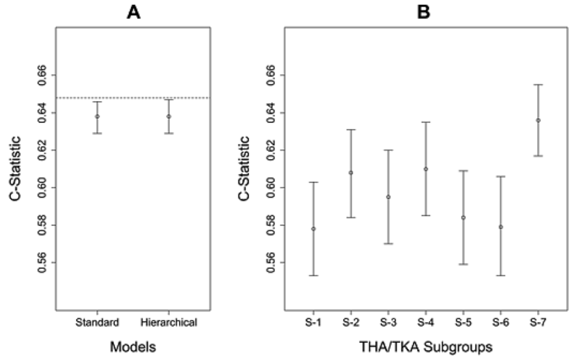


THA/TKA Prediction Model

The inclusion and exclusion criteria (Multimedia Appendix 2) resulted in a cohort of 356,772 patients (16,520 cases and 340,252 controls). As shown in Figure 7A, the Standard Model had a C-statistic of 0.638 (95% CI 0.629-0.646), which was not significantly different (*P*=.69) from the Hierarchical Model, which had a C-statistic of 0.638 (95% CI 0.629-0.647). The calibration plots (Multimedia Appendix 5) revealed that both the models had a slope close to 1 and an intercept close to 0 (Multimedia Appendix 5).

As shown in Figure 7B, the Standard Model was used to measure the predictive accuracy of patients in each subgroup. The results showed that Subgroup-1 had a lower C-statistic than Subgroup-4. Again, although the C-statistics in Figures 7A and 7B cannot be compared as they are based on models developed from different populations, similar to the results in COPD, these results reveal that the current CMS readmission model for THA/TKA might be underperforming for 4 patient subgroups, pinpointing which ones might benefit from a Subgroup-Specific Model.

Figure 7. (A) Predictive accuracy of the Standard Model compared with the Hierarchical Model in total hip arthroplasty/total knee arthroplasty (THA/TKA) as measured by the C-statistic. The C-statistic for the Centers for Medicare & Medicaid Services Standard Model is shown as a dotted line. (B) Predictive accuracy of the Standard Model when applied separately to patients classified to each subgroup. Subgroup-1 has lower accuracy than Subgroup-7. (C-statistics in A and B cannot be compared, as they are based on models developed from different populations).



CMS Standard Model Versus CMS Hierarchical Model

Unlike the CMS published models, the models we developed used only the comorbidities that survived the feature selection. Therefore, to perform a head-to-head comparison with the published CMS models, we also developed a CMS Standard Model (using the same variables from the published CMS model) and compared it to the corresponding CMS Hierarchical Model (with an additional variable for subgroup membership) in each condition. Similar to the models in Figures 5-7, there were no significant differences in the C-statistics between the 2 modeling approaches in any condition (Multimedia Appendix 5). However, as shown in Table 3, the CMS Hierarchical Model for COPD had significantly higher NRI but not significantly higher IDI than the CMS Standard Model, whereas the CMS Hierarchical Model for CHF had a significantly lower NRI and IDI than the CMS Standard Model, and the CMS Hierarchical Model for THA/TKA had a significantly higher NRI but not significantly higher IDI than the CMS Standard Model. Furthermore, similar to the results presented in 6B, 7B, and 8B, when the CMS Standard Model was used to predict readmission separately in subgroups within each index condition, it identified subgroups that underperformed, pinpointing which ones might benefit from a Subgroup-Specific Model (Multimedia Appendix 5). In summary, the comparisons between the CMS Standard Models and the respective CMS Hierarchical Models showed that in the 2 conditions (COPD and THA/TKA), there was a small but statistically significant improvement in discriminating between the readmitted and not readmitted patients as measured by NRI, but not as measured by the C-statistic or IDI, and that a subgroup in each index condition might be underperforming when using the CMS Standard Model.

Table 3. Comparison of the Centers for Medicare & Medicaid Services (CMS) Standard Model with the CMS Hierarchical Model across the three index conditions based on net reclassification improvement (NRI) and integrated discrimination improvement (IDI).

Model	NRI						IDI		
	Categorical (95% CI)	*z* value	*P* value	Continuous (95% CI)	*z* value	*P* value	IDI (95% CI)	*z* value	*P* value
COPD[a]	0.023 (0.012 to 0.034)	−4.10	<.001	0.059 (0.034 to 0.083)	−4.68	<.001	0.0002 (−0.0004 to 0.0008)	−0.65	.51
CHF[b]	−0.010 (−0.016 to −0.004)	3.27	.001	−0.038 (−0.057 to −0.019)	3.92	<.001	−0.0006 (−0.0009 to −0.0003)	3.92	<.001
THA/TKA[c]	0.022 (0.012 to 0.032)	−4.31	<.001	0.111 (0.080 to 0.142)	−7.01	<.001	−0.003 (−0.004 to −0.002)	5.88	<.001

[a] COPD: chronic obstructive pulmonary disease.

[b] CHF: congestive heart failure.

[c] THA/TKA: total hip arthroplasty/total knee arthroplasty.

Discussion

Overview

Our overall approach of using the MIPS framework to identify patient subgroups through visual analytics, and using those subgroups to build classification and prediction models revealed strengths and limitations for each modeling approach and for our data source. This examination provided insights for developing future clinical decision support systems and a methodological framework for improving the clinical interpretability of subgroup modeling results.

Strengths and Limitations of Modeling Methods and Data Source

Visual Analytical Modeling

The results revealed three strengths of the visual analytical modeling: (1) the use of bipartite networks to simultaneously model patients and comorbidities enabled the automatic identification of patient-comorbidity biclusters and the integrated analysis of co-occurrence and risk; (2) the use of a bipartite modularity maximization algorithm to identify the biclusters enabled the measurement of the strength of the biclustering, critical for gauging its significance; and (3) the use of a graph representation enabled the results to be visualized through a network. Furthermore, the clinician stakeholders' request to juxtapose the risk of each subgroup with their visualizations appeared to be driven by the need to reduce working memory loads (from having to remember that information when its spread over different outputs), which could have enhanced their ability to match bicluster patterns with chunks (previously learned patterns of information) stored in long-term memory. The resulting visualizations enabled them to recognize subtypes based on co-occurring comorbidities in each subgroup, reason about the processes that precipitate readmission based on the risk of each subtype relative to the other subtypes, and propose interventions that were targeted to those subtypes and their risks. Finally, the fact that the geriatrician could not fully interpret the THA/TKA network because it combined 2 fairly different conditions suggests that the clinical interpretations were not the result of a *confirmation bias* (interpretations leaning toward fitting the results).

However, the results also revealed two limitations: (1) although modularity is estimated using a closed-form equation (formula), no closed-form equation exists to estimate modularity variance, which is necessary to measure its significance. To estimate modularity variance, we used a permutation test by generating 1000 random permutations of the data and then compared the modularity generated from the real data, to the mean modularity generated from the permuted data. Given the size of our data sets (ranging from 7000 to 25,000 patients), this computationally expensive test took approximately 7 days to complete, despite the use of a dedicated server with multiple cores, and (2) although bicluster modularity was successful in identifying significant and meaningful patient-comorbidity biclusters, the visualizations themselves were extremely dense and therefore potentially concealed patterns within and between the subgroups. Future research should explore defining a closed-form equation to estimate modularity variance, with the goal of accelerating the estimation of modularity significance, and more powerful analytical and visualization methods to reveal intra- and intercluster associations in large and dense networks.

Classification Modeling

The results revealed two strengths of the classification modeling: (1) the use of a simple multinomial classifier was adequate to predict with high accuracy the subgroup to which a patient belonged; (2) because the model produced membership probabilities for each patient for each subgroup, the model captured the dense intercluster edges observed in the network visualization; and (3) the coefficients of the trained classifier could be inspected by an analyst, making it more transparent (relative to most deep learning classifiers that tend to be black boxes).

However, because we dichotomized the classification probabilities into a single subgroup membership, our approach did not fully leverage membership probabilities for modeling and visual interpretation. For example, some patients have high classification probabilities (representing strong membership) for a single subgroup (as shown by patients in the outer periphery of the biclusters with edges only within their bicluster), whereas others have equal probabilities for all subgroups (as shown in the inner periphery of the biclusters with edges going to multiple clusters). Future research should explore incorporating the probability of subgroup membership into the design of Hierarchical Models for improving predictive accuracy, and visualization methods for helping clinicians interpret patients with different profiles of membership strength, with the goal of designing patient-specific interventions.

Predictive Modeling

The results revealed two strengths of the predictive modeling: (1) the use of the Standard Model to measure predictive accuracy across the subgroups helped to pinpoint which subgroups tended to have lower predictive accuracy than the rest and therefore which of them could benefit from a more complex but accurate Subgroup-Specific Model and (2) despite the use of a simple Hierarchical Model with a dichotomized membership label for each patient, the predictive CMS models detected significant differences in the prediction accuracy as measured by NRI in 2 of the conditions, when compared with the CMS Standard Models. However, the results also revealed that the differences in predictive accuracy as measured by the C-statistic and NRI were small, suggesting that comorbidities alone were potentially insufficient for accurately predicting readmission. Future research should explore the use of electronic health records and multiple subgroup-specific models targeted to each subgroup (enabling each model to have different slopes and intercepts) to potentially improve the predictive accuracy of the prediction models.

Data Source

The Medicare claims data had four key strengths: (1) the scale of the data sets that enabled subgroup identification with sufficient statistical power; (2) spread of the data collected from across the United States, which enabled generalizability of the results; (3) data about older adults, which enabled examination of subgroups in an underrepresented segment of the US population; and (4) data used by CMS to build predictive readmission models, which enabled a head-to-head comparison with the Hierarchical Modeling approach.

However, these data had two critical limitations: (1) as we compared our models with the CMS models, we had to use the same definition for controls (90 days with no readmission) that had been used, which introduced a selection bias that exaggerated the separation between cases and controls. Similarly, by excluding patients who died, this exclusion criterion potentially biased the results toward healthier patients and (2) administrative data have known limitations, such as the lack of comorbidity severity and test results, which could strongly impact the accuracy of predictive models. Future research should consider the use of national-level electronic health record data, such as those assembled by the National COVID Cohort Collaborative [54] and the TriNetX [55] initiatives, which could overcome these limitations by providing laboratory values and comorbidity severity but could also introduce new as yet unknown limitations.

Implications for Clinical Decision Support That Leverage Patient Subgroups

Although the focus of this project was to develop and evaluate the MIPS framework, its application to 3 index conditions, coupled with extensive discussions with clinicians, led to insights for designing a future clinical decision support system. Such a system could integrate the outputs from all 3 models in MIPS. As we have shown, the visual analytical model automatically identified and visualized the patient subgroups, which enabled the clinicians to comprehend the co-occurrence and risk information in the visualization, reason about the processes that lead to readmission in each subgroup, and design targeted interventions. The classification model leveraged the observation that many patients have comorbidities in other biclusters (shown by a large number of edges between biclusters) and accordingly generated a membership probability (MP) of a patient belonging to each bicluster, from which the highest was chosen for bicluster membership. Finally, the predictive model calculated the risk of readmission for a patient by using the most accurate model designed for the bicluster to which the patient belonged.

The outputs from these models could be integrated into a clinical decision support system to provide recommendations for a specific patient using the following algorithm: (1) use the classifier to generate the MP of a new patient belonging to each subgroup; (2) use the predictive model to calculate the risk (R) of that patient in each subgroup; (3) generate an importance score (IS) for each subgroup, such as by calculating a *membership-weighted risk* [MP x R]; (4) rank the subgroups and their respective interventions using IS; and (5) use the ranking to display in descending order, the subgroup comorbidity profiles along with their respective potential mechanisms, recommended treatments, and the respective IS. Such model-based information, displayed through a user-friendly interface, could enable a clinician to rapidly scan the ranked list to (1) determine why a specific patient profile fits into one or more subgroups, (2) review the potential mechanisms and interventions ranked by their importance, and (3) use the combined information to design a treatment that is customized for the real-world context of the patient. Consequently, such a clinical decision-support system could not only provide a quantitative ranking of membership to different subgroups and the IS for the associated interventions, but also enable the clinician to understand the rationale underlying those recommendations, making the system interpretable and explainable. Our current work explores a framework called Model-based Subtype and Treatment Recommendations (MASTR) for developing such clinical decision-support systems, and evaluating them to determine their clinical efficacy in comparison to standard-of-care.

Implications for Analytical Granularity to Enhance the Interpretability of Patient Subgroups

Although the visual analytical model enabled clinicians to interpret the patient subgroups, they were unable to interpret the associations within and between the subgroups because of the large number of nodes in each bicluster and the dense edges between them. Several network filtering methods [56,57] have been developed to *thin out* such dense networks such as by dropping or bundling nodes and edges based on user-defined criteria, to improve visual interpretation. However, such filtering could bias the results or modify the clusters resulting from reduced data.

An alternate approach that preserves the full data set leverages the notion of analytic granularity, in which the data are progressively analyzed at different levels. For example, we have analyzed patients with COVID-19 [11] at the cohort, subgroup, and patient levels, and we are currently using the same approach to examine symptom co-occurrence and risk at each level in patients with Long COVID. Our preliminary results suggest that analyzing data at different levels of granularity enables clinicians to progressively interpret patterns, such as within and between subgroups, in addition to guiding the systematic development of new algorithms. For example, at the subgroup level, we have designed an algorithm that identifies which patient subgroups have a significantly higher probability of having characteristics that are clustered in another subgroup, providing critical information to clinicians about how to design interventions for such overlapping subgroups. Furthermore, at the patient level, we have identified patients that are very dissimilar to their subgroups based on their pattern of characteristics inside and outside their subgroup. Such dissimilar patients could be flagged to examine whether they need individualized interventions compared with those recommended for the rest of their subgroups. Such analytical granularity could therefore inform the design of interventions by clinicians in addition to the design of decision support systems that provide targeted and interpretable recommendations to physicians, who can then customize them to fit the real-world context of a patient.

Implications of the MIPS Framework for Precision Medicine

Although we have demonstrated the application of the MIPS framework across multiple readmission conditions, its architecture has 3 properties that should enable its generalizability across other medical conditions. First, as shown in Figure 1, the framework is *modular* with explicit inputs and outputs, enabling the use of other methods in each of the 3 modeling steps. For example, the framework can use other biclustering (eg, nonnegative matrix factorization) [58], classification (eg, deep learning) [59], and prediction methods (eg, subgroup-specific modeling) [16]. Second, the framework is *extensible*, enabling elaboration of the methods at each modeling step to improve the analysis and interpretation of subgroups. For example, as discussed earlier, analytical granularity at the cohort, subgroup, and patient levels could improve the interpretability of subgroups in large and dense data sets. Third, the framework is *integrative* as it systematically combines the strengths of machine learning and statistical and precision medicine approaches. For example, visual analytical modeling leverages search algorithms to discover co-occurrence in large data sets, classification and prediction modeling leverages probability theory to measure the risk of co-occurrence patterns, and clinicians leverage medical knowledge and human cognition to interpret patterns of co-occurrence and risk for designing precision medicine interventions. Therefore, the integration of these different models with a focus on their clinical interpretation operationalizes *team-centered informatics* [60] designed to facilitate data scientists, biostatisticians, and clinicians in multidisciplinary translational teams [61] to work more effectively across disciplinary boundaries with the goal of designing precision medicine interventions. Our current research tests the generality of the MIPS framework in other conditions, such as in Long COVID and poststroke depression, with the goal of designing and evaluating precision medicine interventions targeted to patient subgroups.

Conclusions

Although several studies have identified patient subgroups in different health conditions, there is a considerable gap between the identification of subgroups and their modeling and interpretation for clinical applications. Here, we developed MIPS, a novel analytical framework to bridge this gap, using a 3-step modeling approach. A visual analytical method automatically identified statistically significant and replicated patient subgroups and their frequently co-occurring comorbidities, which were clinically significant. Next, a multinomial logistic regression classifier was highly accurate in correctly classifying patients into subgroups identified by the visual analytical model. Finally, despite using a simple hierarchical logistic regression model to incorporate subgroup information, the predictive models showed a statistically significant improvement in discriminating between readmitted and not readmitted patients in 2 of the 3 readmission conditions, and additional analysis pinpointed for which patient subgroups the current CMS model might be underperforming. Furthermore, the integration of the 3 models helped to (1) elucidate the data input and output dependencies among the models, enabling clinicians to interpret the patient subgroups, reason about mechanisms precipitating hospital readmission, and design targeted interventions and (2) provide a generalizable framework for the development of future clinical decision support systems that integrate outputs from each of the 3 modeling approaches.

However, the evaluation of MIPS across the 3 readmission index conditions also helped to identify the limitations of each modeling method, and of the data. The visual analytical model was too dense to enable clinicians to interpret the associations within and between subgroups, and the absence of a closed-form equation to measure modularity variance required a computationally expensive process to measure the significance of the biclustering. Furthermore, the small improvement in predictive accuracy suggested that comorbidities alone were insufficient for accurately predicting hospital readmission.

By leveraging the modular and extensible nature of the MIPS framework, future research should address these limitations by developing more powerful algorithms that analyze subgroups at different levels of granularity to improve the interpretability of intra- and intercluster associations and the evaluation of subgroup-specific models to predict outcomes. Furthermore, data from electronic health records made available through national-level data initiatives, such as National COVID Cohort Collaborative and TriNetX, now provide access to critical variables, including laboratory results and comorbidity severity, which should lead to higher accuracy in predicting adverse outcomes. Finally, extensive discussions with clinicians have confirmed the need for decision support systems that integrate outputs from the 3 models to provide for a specific patient, predicted subgroup memberships, and ranked interventions, along with associated subgroup profiles and mechanisms. Such interpretable and explainable systems could enable clinicians to use patient subgroup information for informing the design of precision medicine interventions, with the goal of reducing adverse outcomes such as unplanned hospital readmissions and beyond.

Acknowledgments

The authors thank Tianlong Chen, Clark Andersen, Yu-Li Lin, Gautam Vallabha, Erich Kummerfeld, and Emmanuel Santillana for their assistance in conducting the analyses. This study was supported in part by the Patient-Centered Outcomes Research Institute (ME-1511-33194) and the Clinical and Translational Science Award (UL1 TR001439) from the National Center for Advancing Translational Sciences at the National Institutes of Health, the University of Texas Medical Branch Claude D Pepper Older Americans Independence Center funded by the National Institute of Aging at the National Institutes of Health (P30 AG024832), MD Anderson Cancer Center, and the National Library of Medicine (R01 LM012095) at the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Patient-Centered Outcomes Research Institute or National Institutes of Health. Data from Medicare were analyzed using a Centers for Medicare & Medicaid Services data-use agreement (CMS DUA RSCH-2017-51404).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Analytical methods for modeling and interpreting patient subgroups.
[DOCX File , 1257 KB - medinform_v10i12e37239_app1.docx]

Multimedia Appendix 2
Patient inclusion and exclusion criteria.
[DOCX File , 184 KB - medinform_v10i12e37239_app2.docx]

Multimedia Appendix 3
Variable and feature selection.
[DOCX File , 45 KB - medinform_v10i12e37239_app3.docx]

Multimedia Appendix 4
Classification modeling.
[DOCX File , 45 KB - medinform_v10i12e37239_app4.docx]

Multimedia Appendix 5
Predictive modeling.
[DOCX File , 6683 KB - medinform_v10i12e37239_app5.docx]

References

1. McClellan J, King MC. Genetic heterogeneity in human disease. Cell 2010 Apr 16;141(2):210-217 [FREE Full text] [doi: 10.1016/j.cell.2010.03.032] [Medline: 20403315]
2. Waldman SA, Terzic A. Therapeutic targeting: a crucible for individualized medicine. Clin Pharmacol Ther 2008 May;83(5):651-654. [doi: 10.1038/clpt.2008.65] [Medline: 18425084]
3. Rouzier R, Perou CM, Symmans WF, Ibrahim N, Cristofanilli M, Anderson K, et al. Breast cancer molecular subtypes respond differently to preoperative chemotherapy. Clin Cancer Res 2005 Aug 15;11(16):5678-5685. [doi: 10.1158/1078-0432.CCR-04-2421] [Medline: 16115903]
4. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 2001 Sep 11;98(19):10869-10874 [FREE Full text] [doi: 10.1073/pnas.191367098] [Medline: 11553815]
5. Fitzpatrick AM, Teague WG, Meyers DA, Peters SP, Li X, Li H, National Institutes of Health/National Heart, Lung, and Blood Institute Severe Asthma Research Program. Heterogeneity of severe asthma in childhood: confirmation by cluster analysis of children in the National Institutes of Health/National Heart, Lung, and Blood Institute Severe Asthma Research Program. J Allergy Clin Immunol 2011 Feb;127(2):382-9.e1. [doi: 10.1016/j.jaci.2010.11.015] [Medline: 21195471]

6. Haldar P, Pavord ID, Shaw DE, Berry MA, Thomas M, Brightling CE, et al. Cluster analysis and clinical asthma phenotypes. Am J Respir Crit Care Med 2008 Aug 01;178(3):218-224 [FREE Full text] [doi: 10.1164/rccm.200711-1754OC] [Medline: 18480428]
7. Lötvall J, Akdis CA, Bacharier LB, Bjermer L, Casale TB, Custovic A, et al. Asthma endotypes: a new approach to classification of disease entities within the asthma syndrome. J Allergy Clin Immunol 2011 Feb;127(2):355-360. [doi: 10.1016/j.jaci.2010.11.037] [Medline: 21281866]
8. Nair P, Pizzichini MM, Kjarsgaard M, Inman MD, Efthimiadis A, Pizzichini E, et al. Mepolizumab for prednisone-dependent asthma with sputum eosinophilia. N Engl J Med 2009 Mar 05;360(10):985-993. [doi: 10.1056/NEJMoa0805435] [Medline: 19264687]
9. Ortega HG, Liu MC, Pavord ID, Brusselle GG, FitzGerald JM, Chetta A, MENSA Investigators. Mepolizumab treatment in patients with severe eosinophilic asthma. N Engl J Med 2014 Sep 25;371(13):1198-1207. [doi: 10.1056/NEJMoa1403290] [Medline: 25199059]
10. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med 2015 Feb 26;372(9):793-795 [FREE Full text] [doi: 10.1056/NEJMp1500523] [Medline: 25635347]
11. Bhavnani SK, Kummerfeld E, Zhang W, Kuo YF, Garg N, Visweswaran S, et al. Heterogeneity in COVID-19 patients at multiple levels of granularity: from biclusters to clinical interventions. AMIA Jt Summits Transl Sci Proc 2021 May 17;2021:112-121 [FREE Full text] [Medline: 34457125]
12. Lacy ME, Wellenius GA, Carnethon MR, Loucks EB, Carson AP, Luo X, et al. Racial differences in the performance of existing risk prediction models for incident type 2 diabetes: the CARDIA study. Diabetes Care 2016 Feb;39(2):285-291 [FREE Full text] [doi: 10.2337/dc15-0509] [Medline: 26628420]
13. Baker JJ. Medicare payment system for hospital inpatients: diagnosis-related groups. J Health Care Finance 2002;28(3):1-13. [Medline: 12079147]
14. Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search--a recursive partitioning method for establishing response to treatment in patient subpopulations. Stat Med 2011 Sep 20;30(21):2601-2621. [doi: 10.1002/sim.4289] [Medline: 21786278]
15. Kehl V, Ulm K. Responder identification in clinical trials with censored data. Comput Stat Data Anal 2006 Mar;50(5):1338-1355. [doi: 10.1016/j.csda.2004.11.015]
16. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY, USA: Springer; 2001.
17. Abu-jamous B, Fa R, Nandi A. Integrative Cluster Analysis in Bioinformatics. West Sussex, UK: John Wiley & Sons; 2015.
18. Lochner KA, Cox CS. Prevalence of multiple chronic conditions among Medicare beneficiaries, United States, 2010. Prev Chronic Dis 2013 Apr 25;10:E61 [FREE Full text] [doi: 10.5888/pcd10.120137] [Medline: 23618541]
19. Aryal S, Diaz-Guzman E, Mannino DM. Prevalence of COPD and comorbidity. In: Rabe KF, Wedzicha JA, Wouters EF, editors. European Respiratory Monograph: COPD and Comorbidity. Lausanne, Switzerland: European Respiratory Society; 2013:1-12.
20. Baty F, Putora PM, Isenring B, Blum T, Brutsche M. Comorbidities and burden of COPD: a population based case-control study. PLoS One 2013 May 17;8(5):e63285 [FREE Full text] [doi: 10.1371/journal.pone.0063285] [Medline: 23691009]
21. Moni MA, Liò P. Network-based analysis of comorbidities risk during an infection: SARS and HIV case studies. BMC Bioinformatics 2014 Oct 24;15(1):333 [FREE Full text] [doi: 10.1186/1471-2105-15-333] [Medline: 25344230]
22. Cramer AO, Waldorp LJ, van der Maas HL, Borsboom D. Comorbidity: a network perspective. Behav Brain Sci 2010 Jun;33(2-3):137-193. [doi: 10.1017/S0140525X09991567] [Medline: 20584369]
23. Islam MM, Valderas JM, Yen L, Dawda P, Jowsey T, McRae IS. Multimorbidity and comorbidity of chronic diseases among the senior Australians: prevalence and patterns. PLoS One 2014 Jan 8;9(1):e83783 [FREE Full text] [doi: 10.1371/journal.pone.0083783] [Medline: 24421905]
24. Folino F, Pizzuti C, Ventura M. A comorbidity network approach to predict disease risk. In: Proceedings of the 1st International Conference on Information, Technology in Bio- and Medical Informatics. 2010 Presented at: ITBAM '10; September 1-2, 2010; Bilbao, Spain p. 102-109. [doi: 10.1007/978-3-642-15020-3_10]
25. Newman ME. Networks: An Introduction. Oxford, UK: Oxford University Press; 2010.
26. Treviño III S, Nyberg A, Del Genio CI, Bassler KE. Fast and accurate determination of modularity and its effect size. J Stat Mech 2015 Feb 03;2015(2):P02003. [doi: 10.1088/1742-5468/2015/02/p02003]
27. Chauhan R, Ravi J, Datta P, Chen T, Schnappinger D, Bassler KE, et al. Reconstruction and topological characterization of the sigma factor regulatory network of Mycobacterium tuberculosis. Nat Commun 2016 Mar 31;7:11062 [FREE Full text] [doi: 10.1038/ncomms11062] [Medline: 27029515]
28. Bhavnani SK, Dang B, Penton R, Visweswaran S, Bassler KE, Chen T, et al. How high-risk comorbidities co-occur in readmitted patients with hip fracture: big data visual analytical approach. JMIR Med Inform 2020 Oct 26;8(10):e13567 [FREE Full text] [doi: 10.2196/13567] [Medline: 33103657]
29. Fruchterman TM, Reingold EM. Graph drawing by force-directed placement. Softw Pract Exper 1991 Nov;21(11):1129-1164. [doi: 10.1002/spe.4380211102]

30. Dang B, Chen T, Bassler KE, Bhavnani SK. ExplodeLayout: enhancing the comprehension of large and dense networks. AMIA Jt Summits Transl Sci Proc 2016.
31. Bhavnani SK, Chen T, Ayyaswamy A, Visweswaran S, Bellala G, Rohit D, et al. Enabling comprehension of patient subgroups and characteristics in large bipartite networks: implications for precision medicine. AMIA Jt Summits Transl Sci Proc 2017 Jul 26;2017:21-29 [FREE Full text] [Medline: 28815099]
32. Bhavnani SK, Eichinger F, Martini S, Saxman P, Jagadish HV, Kretzler M. Network analysis of genes regulated in renal diseases: implications for a molecular-based classification. BMC Bioinformatics 2009 Sep 17;10 Suppl 9:S3 [FREE Full text] [doi: 10.1186/1471-2105-10-S9-S3] [Medline: 19761573]
33. Bhavnani SK, Bellala G, Ganesan A, Krishna R, Saxman P, Scott C, et al. The nested structure of cancer symptoms. Implications for analyzing co-occurrence and managing symptoms. Methods Inf Med 2010;49(6):581-591 [FREE Full text] [doi: 10.3414/ME09-01-0083] [Medline: 21085743]
34. Bhavnani SK, Ganesan A, Hall T, Maslowski E, Eichinger F, Martini S, et al. Discovering hidden relationships between renal diseases and regulated genes through 3D network visualizations. BMC Res Notes 2010 Nov 11;3:296 [FREE Full text] [doi: 10.1186/1756-0500-3-296] [Medline: 21070623]
35. Bhavnani SK, Victor S, Calhoun WJ, Busse WW, Bleecker E, Castro M, et al. How cytokines co-occur across asthma patients: from bipartite network analysis to a molecular-based classification. J Biomed Inform 2011 Dec;44 Suppl 1:S24-S30 [FREE Full text] [doi: 10.1016/j.jbi.2011.09.006] [Medline: 21986291]
36. Bhavnani SK, Bellala G, Victor S, Bassler KE, Visweswaran S. The role of complementary bipartite visual analytical representations in the analysis of SNPs: a case study in ancestral informative markers. J Am Med Inform Assoc 2012 Jun;19(e1):e5-12 [FREE Full text] [doi: 10.1136/amiajnl-2011-000745] [Medline: 22718038]
37. Bhavnani SK, Dang B, Bellala G, Divekar R, Visweswaran S, Brasier AR, et al. Unlocking proteomic heterogeneity in complex diseases through visual analytics. Proteomics 2015 Apr;15(8):1405-1418 [FREE Full text] [doi: 10.1002/pmic.201400451] [Medline: 25684269]
38. Bhavnani SK, Dang B, Kilaru V, Caro M, Visweswaran S, Saade G, et al. Methylation differences reveal heterogeneity in preterm pathophysiology: results from bipartite network analyses. J Perinat Med 2018 Jul 26;46(5):509-521 [FREE Full text] [doi: 10.1515/jpm-2017-0126] [Medline: 28665803]
39. Jencks SF, Williams MV, Coleman EA. Rehospitalizations among patients in the Medicare fee-for-service program. N Engl J Med 2009 Apr 02;360(14):1418-1428. [doi: 10.1056/NEJMsa0803563] [Medline: 19339721]
40. Report to Congress: Promoting Greater Efficiency in Medicare. MedPac (Medicare Payment Advisory Commission). 2007 Jun. URL: https://www.ruralhealthinfo.org/assets/4544-19961/Jun07_EntireReport.pdf [accessed 2021-11-30]
41. Ashton CM, Del Junco DJ, Souchek J, Wray NP, Mansyur CL. The association between the quality of inpatient care and early readmission: a meta-analysis of the evidence. Med Care 1997 Oct;35(10):1044-1059. [doi: 10.1097/00005650-199710000-00006] [Medline: 9338530]
42. 2015 Condition-specific measures updates and specifications report: Hospital-level 30-day risk-standardized readmission measures on acute myocardial infarction, heart failure, pneumonia, chronic obstructive pulmonary disease, and stoke. Centers for Medicare & Medicaid Services. 2015 Apr 20. URL: http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Measure-Methodology.html [accessed 2021-11-30]
43. Keenan PS, Normand SL, Lin Z, Drye EE, Bhat KR, Ross JS, et al. An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. Circ Cardiovasc Qual Outcomes 2008 Sep;1(1):29-37. [doi: 10.1161/CIRCOUTCOMES.108.802686] [Medline: 20031785]
44. 2015 Procedure-specific readmission measures updates and specifications report: elective primary total hip arthroplasty and/or total knee arthroplasty, and isolated coronary artery bypass graft surgery. Centers for Medicare & Medicaid Services. 2015 Apr 20. URL: http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Measure-Methodology.html [accessed 2021-11-30]
45. Using SAS® to Perform Individual Matching in Design of Case-Control Studies. SAS. 2010. URL: https://support.sas.com/resources/papers/proceedings10/061-2010.pdf [accessed 2020-05-05]
46. Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc 1971 Dec;66(336):846-850. [doi: 10.2307/2284239]
47. Sharif R, Parekh TM, Pierson KS, Kuo YF, Sharma G. Predictors of early readmission among patients 40 to 64 years of age hospitalized for chronic obstructive pulmonary disease. Ann Am Thorac Soc 2014 Jun;11(5):685-694 [FREE Full text] [doi: 10.1513/AnnalsATS.201310-358OC] [Medline: 24784958]
48. Grosso LM, Curtis JP, Lin Z, Geary LL, Vellanky S, Oladele C. Hospital-level 30-day all-cause risk-standardized readmission rate following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA). Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation. 2012. URL: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Measure-Methodology [accessed 2022-06-08]
49. Medicare Enrollment Charts. Chronic Conditions Data Warehouse. 2022 Mar 29. URL: https://www2.ccwdata.org/web/guest/medicare-charts/medicare-enrollment-charts#a1_coverage_trend [accessed 2022-06-08]
50. 2020 profile of older Americans. Administration for Community Living. 2021 May. URL: https://acl.gov/sites/default/files/Aging%20and%20Disability%20in%20America/2020ProfileOlderAmericans.Final_.pdf [accessed 2022-03-29]

51. 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. [doi: 10.1016/0021-9681(87)90171-8] [Medline: 3558716]
52. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care 1998 Jan;36(1):8-27. [doi: 10.1097/00005650-199801000-00004] [Medline: 9431328]
53. 2017 Condition-specific measures updates and specifications report: Hospital-level 30-day risk-standardized readmission measures on acute myocardial infarction, heart failure, pneumonia, chronic obstructive pulmonary disease, and stroke. Centers for Medicare & Medicaid Services. 2017. URL: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Measure-Methodology.html [accessed 2017-05-31]
54. Bennett TD, Moffitt RA, Hajagos JG, Amor B, Anand A, Bissell MM, National COVID Cohort Collaborative (N3C) Consortium. Clinical characterization and prediction of clinical severity of SARS-CoV-2 infection among US adults using data from the US National COVID Cohort Collaborative. JAMA Netw Open 2021 Jul 01;4(7):e2116901 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.16901] [Medline: 34255046]
55. Topaloglu U, Palchuk MB. Using a federated network of real-world data to optimize clinical trials operations. JCO Clin Cancer Inform 2018 Dec;2:1-10 [FREE Full text] [doi: 10.1200/CCI.17.00067] [Medline: 30652541]
56. Dogrusoz U, Karacelik A, Safarli I, Balci H, Dervishi L, Siper MC. Efficient methods and readily customizable libraries for managing complexity of large networks. PLoS One 2018 May 29;13(5):e0197238 [FREE Full text] [doi: 10.1371/journal.pone.0197238] [Medline: 29813080]
57. Wu J, Zhu F, Liu X, Yu H. An information-theoretic framework for evaluating edge bundling visualization. Entropy (Basel) 2018 Aug 21;20(9):625 [FREE Full text] [doi: 10.3390/e20090625] [Medline: 33265714]
58. Dhillon IS, Sra S. Generalized nonnegative matrix approximations with Bregman divergences. In: Proceedings of the 18th International Conference on Neural Information Processing Systems. 2005 Presented at: NIPS '05; December 5-8, 2005; Vancouver, Canada p. 283-290.
59. Dilsizian ME, Siegel EL. Machine meets biology: a primer on artificial intelligence in cardiology and cardiac imaging. Curr Cardiol Rep 2018 Oct 18;20(12):139. [doi: 10.1007/s11886-018-1074-8] [Medline: 30334108]
60. Bhavnani SK, Visweswaran S, Divekar R, Brasier AR. Towards team-centered informatics: accelerating innovation in multidisciplinary scientific teams through visual analytics. J Appl Behav Sci 2018 Nov 05;55(1):50-72. [doi: 10.1177/0021886318794606]
61. Wooten KC, Calhoun WJ, Bhavnani S, Rose RM, Ameredes B, Brasier AR. Evolution of Multidisciplinary Translational Teams (MTTs): insights for accelerating translational innovations. Clin Transl Sci 2015 Oct;8(5):542-552 [FREE Full text] [doi: 10.1111/cts.12266] [Medline: 25801998]

Abbreviations

CHF: congestive heart failure
CMS: Centers for Medicare & Medicaid Services
COPD: chronic obstructive pulmonary disease
IDI: integrated discrimination improvement
IS: importance score
MIPS: modeling and interpreting patient subgroups
MP: membership probability
NRI: net reclassification improvement
OR: odds ratio
RI: Rand Index
THA/TKA: total hip arthroplasty/total knee arthroplasty

Edited by C Lovis; submitted 11.02.22; peer-reviewed by E Massou; comments to author 10.03.22; revised version received 06.04.22; accepted 02.05.22; published 07.12.22.

Please cite as:
Bhavnani SK, Zhang W, Visweswaran S, Raji M, Kuo YF
A Framework for Modeling and Interpreting Patient Subgroups Applied to Hospital Readmission: Visual Analytical Approach
JMIR Med Inform 2022;10(12):e37239
URL: https://medinform.jmir.org/2022/12/e37239
doi:10.2196/37239
PMID:35537203

Original Paper

Identifying Patterns of Clinical Interest in Clinicians' Treatment Preferences: Hypothesis-free Data Science Approach to Prioritizing Prescribing Outliers for Clinical Review

Brian MacKenna[1*], MPharm; Helen J Curtis[1*], DPhil; Lisa E M Hopcroft[1], PhD; Alex J Walker[1], PhD; Richard Croker[1], MSc; Orla Macdonald[1], MPharm; Stephen J W Evans[2], MSc; Peter Inglesby[1], MPhil; David Evans[1], MPhil; Jessica Morley[1], MSc; Sebastian C J Bacon[1], BA; Ben Goldacre[1], MRCPsych

[1]Bennett Institute for Applied Data Science, Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom

[2]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Ben Goldacre, MRCPsych
Bennett Institute for Applied Data Science
Nuffield Department of Primary Care Health Sciences
University of Oxford
Radcliffe Primary Care Building, Radcliffe Observatory Quarter
32 Woodstock Road
Oxford, OX2 6GG
United Kingdom
Phone: 44 01865 617855
Email: ben.goldacre@phc.ox.ac.uk

Abstract

Background: Data analysis is used to identify signals suggestive of variation in treatment choice or clinical outcome. Analyses to date have generally focused on a hypothesis-driven approach.

Objective: This study aimed to develop a hypothesis-free approach to identify unusual prescribing behavior in primary care data. We aimed to apply this methodology to a national data set in a cross-sectional study to identify chemicals with significant variation in use across Clinical Commissioning Groups (CCGs) for further clinical review, thereby demonstrating proof of concept for prioritization approaches.

Methods: Here we report a new data-driven approach to identify unusual prescribing behaviour in primary care data. This approach first applies a set of filtering steps to identify chemicals with prescribing rate distributions likely to contain outliers, then applies two ranking approaches to identify the most extreme outliers amongst those candidates. This methodology has been applied to three months of national prescribing data (June-August 2017).

Results: Our methodology provides rankings for all chemicals by administrative region. We provide illustrative results for 2 antipsychotic drugs of particular clinical interest: promazine hydrochloride and pericyazine, which rank highly by outlier metrics. Specifically, our method identifies that, while promazine hydrochloride and pericyazine are barely used by most clinicians (with national prescribing rates of 11.1 and 6.2 per 1000 antipsychotic prescriptions, respectively), they make up a substantial proportion of antipsychotic prescribing in 2 small geographic regions in England during the study period (with maximum regional prescribing rates of 298.7 and 241.1 per 1000 antipsychotic prescriptions, respectively).

Conclusions: Our hypothesis-free approach is able to identify candidates for audit and review in clinical practice. To illustrate this, we provide 2 examples of 2 very unusual antipsychotics used disproportionately in 2 small geographic areas of England.

(JMIR Med Inform 2022;10(12):e41200) doi:10.2196/41200

KEYWORDS

prescribing; NHS England; antipsychotics; promazine hydrochloride; pericyazine; clinical audit; data science

Introduction

Since 2011, the National Health Service (NHS) in England has openly shared detailed monthly general practice prescribing data to the level of individual doses, chemicals, and brands, aggregated at the individual general practice level. These data have supported original research on a broad range of topics as well as supporting systematic audit and review programs to realize improvements in primary care prescribing [1].

Our group produces OpenPrescribing.net [2], a free and widely used tool where anyone can explore the prescriptions dispensed at any practice in England and monitor prescribing patterns down to the level of individual brands, formulations, and doses. OpenPrescribing offers data-driven feedback to assist regional- and practice-level medicines optimization teams and identifies areas for review of which they may not otherwise have been aware. For example, we identify whether each NHS organization is an outlier on more than 80 predefined measures covering a range of prescribing safety, cost-effectiveness, and efficacy issues. Unique savings opportunities for each practice by making comparisons between brands or generics and formulations are also calculated [3], and there is evidence that these savings are realized [4].

Typically, data science for service audit and quality improvement is hypothesis driven: identifying a targeted behavior and using data to measure the achievement of that goal [5,6]. Given the vast scale of openly available NHS prescribing data (more than 2 billion rows of data covering 8000 organizations during the past decade) and the vast range of clinical behaviors and potential signals for variation in care that may lie within this data set, we set out to develop new hypothesis-free data science techniques to identify new opportunities for service improvement driven by variation in care.

Our overall analytic aim was to prototype and describe methods to identify previously unknown signals of clinical interest in prescribing data (existing methodology to identify outliers often focuses on financial aspects of prescribing [7-9] or is focused on a specific clinical question [10,11]). We ran a series of internal workshops to develop a short list of data science methods that might be used to identify prescribing behaviors that are unusually distributed across NHS organizations or regions. Here, we briefly report the successful deployment of one such method (ranking chemicals by kurtosis and a ratio between intercentile differences across all chemical-class pairs) and demonstrate how this identified high prescribing of unusual antipsychotics in 2 small regions of England.

Methods

Study Design and Data Sources

We conducted a cross-sectional study using open NHS prescribing data on all dispensed products prescribed by general practices in England, June-August 2017, extracted from the OpenPrescribing database. A relatively short 3-month window was chosen, owing to the fact that this work represents a proof of concept. The data set includes, for each practice, product and month of prescribing, the number of items prescribed (equivalent to the number of prescription forms on which each product appeared), and the total quantity (eg, tablets and mL). Practices were grouped by their parent Clinical Commissioning Group (CCG), an NHS administrative region. In England, approximately 7000 NHS general practices were arranged into 207 CCGs in 2017.

Data Processing

All chemicals prescribed in England were assigned to a “class” of chemicals, using their British National Formulary (BNF) legacy code to identify the chemical’s relevant BNF subparagraph. We limited our search to chemicals in chapters 1-15 of the BNF (1511 prescribed chemicals) to exclude chapters not following a chemical/subparagraph structure, which largely cover nonmedicinal products such as dressings. For each chemical-class pair, the number of items (similar to a prescription in prescribing data) that were prescribed for each chemical was expressed as a proportion of the total items prescribed of all chemicals in its class. These chemical-class proportion values were calculated for each CCG. To avoid including rarely prescribed classes of chemicals that would generate spurious findings, we excluded 116 chemicals with the lowest total items prescribed (specifically, the lowest two centiles) and 4 chemicals that were used by less than 50 CCGs. In total, then, 1395 chemicals were subject to analysis.

Ranking Chemical-Class Pairs by Outlier Metrics

We first sought to focus our analysis on those chemicals with the distribution characteristics indicative of (1) reasonable variability and (2) positive outliers among CCGs (ie, outliers with higher prescribing rates rather than outliers at lower prescribing rates): chemical-class pairs were filtered where range>10% and skew>0. This identified 412 candidate chemicals of interest. To further refine this group of chemicals, we retained only those candidates for which (1) the median proportion was <0.1, that is, those prescribed at a very low rate, or not at all, by most CCGs and (2) the number of prescriptions nationally was not small (at least 1000 prescriptions), so as to limit the impact of random fluctuations in small numbers of prescriptions. These further filtering steps reduced our candidate list to 204 chemicals.

We then implemented 2 alternative ranking approaches to identify outliers among our candidate chemicals. The first was kurtosis, which can be described as a numerical measure of the extent to which the tails of a given distribution are heavier or lighter than a normal distribution; overall, data sets with high kurtosis will tend to have more extreme outliers than data sets with low kurtosis. Kurtosis is a good method for detecting an unknown number of outliers in a data set [12,13]. We calculated the kurtosis for each candidate chemical-class pair across all CCGs and ranked the chemicals by this kurtosis value (highest to lowest). We then generated an alternative ranking of chemicals using a ratio calculated as the intercentile range of the chemical-class proportion between the 95th and 97th centiles (the top prescribing CCGs) to the intercentile range between 50th and 95th centiles (those CCGs prescribing at more moderate rates); this ratio will hereafter be referred to as the “high:mid centile ratio”.

Both approaches sort all chemicals into an order where, for the most highly ranked chemicals, there are very substantial differences between CCGs in the extent to which that chemical is used in the context of all prescribing of all chemicals in its class. This ranking was used to prioritize the chemical-class pairs for manual evaluation by clinical staff (BMK, RC, OM, and BG) for signals of clinical interest.

Visualizing Prescribing Rates Using Choropleth Maps

For selected chemical-class pairs of clinical interest, we generated a choropleth map using OpenPrescribing.net to visualize the geographical distribution of prescribing of each chemical as a proportion of its class. Data management was performed using Python and Google BigQuery, with the analysis carried out using Python (authors HJC and LEMH). Data and charts, as well as all code for data management and analysis are openly available for inspection and reuse on GitHub [14].

Ethical Considerations

This study uses exclusively open publicly available data; therefore, no ethical approval was required.

Results

A total of 204 chemicals were found to have prescribing rate distributions indicative of positive outliers among administrative regions in the NHS in England. Figure 1 summarizes the high:mid centile ratio and kurtosis value for these chemicals. The top 5 ranked chemicals by either outlier measurement are highlighted.

Clinical review of these results identified 2 chemical substances to illustrate the methodology: promazine hydrochloride (high:mid centile ratio: 1.804, kurtosis: 43.61) and pericyazine (high:mid centile ratio: 0.880, kurtosis: 49.60). These 2 antipsychotic drugs are the top 2 ranking chemicals by high:mid centile ratio and also rank in the top 10 (ninth and seventh, respectively) by kurtosis.

Exploring these chemicals in more detail, pericyazine is shown to be prescribed at a much higher rate in the East of England (Table 1), with 13,119 in 277,470 (4.7%) antipsychotic prescriptions being for pericyazine, compared to 15,344 in 2,489,069 (0.6%) nationally. OpenPrescribing choropleth maps demonstrate that this high level of prescribing was concentrated particularly in Norwich and the Norfolk area more widely (Figure 2A; Multimedia Appendix 1). Promazine hydrochloride is shown to be prescribed at higher levels in the North West of England (Table 1), accounting for 20,060 in 412,624 (4.9%) antipsychotic prescriptions compared to 27,724 in 2,489,069 (1.1%) nationally. Again, these outlier prescribing behaviors were concentrated in specific CCGs: Bolton and the wider Greater Manchester area (Figure 2B; Multimedia Appendix 2).

Figure 1. Prioritizing 204 candidate chemicals using ranking by 2 outlier metrics. The top 5 chemicals by either the high:med centile ratio or kurtosis are highlighted in orange; all other chemicals are shown in gray. Each metric is summarized as a histogram of chemical counts along the corresponding axis. Pericyazine and promazine hydrochloride (our chemicals of interest) are highlighted in bold.

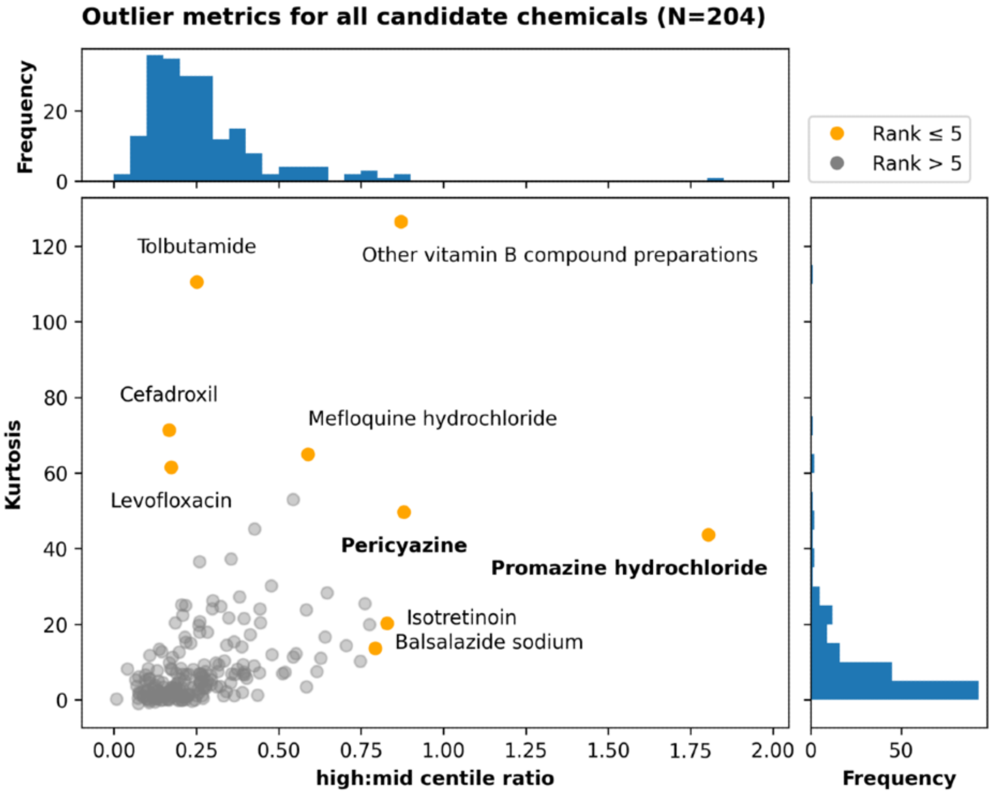
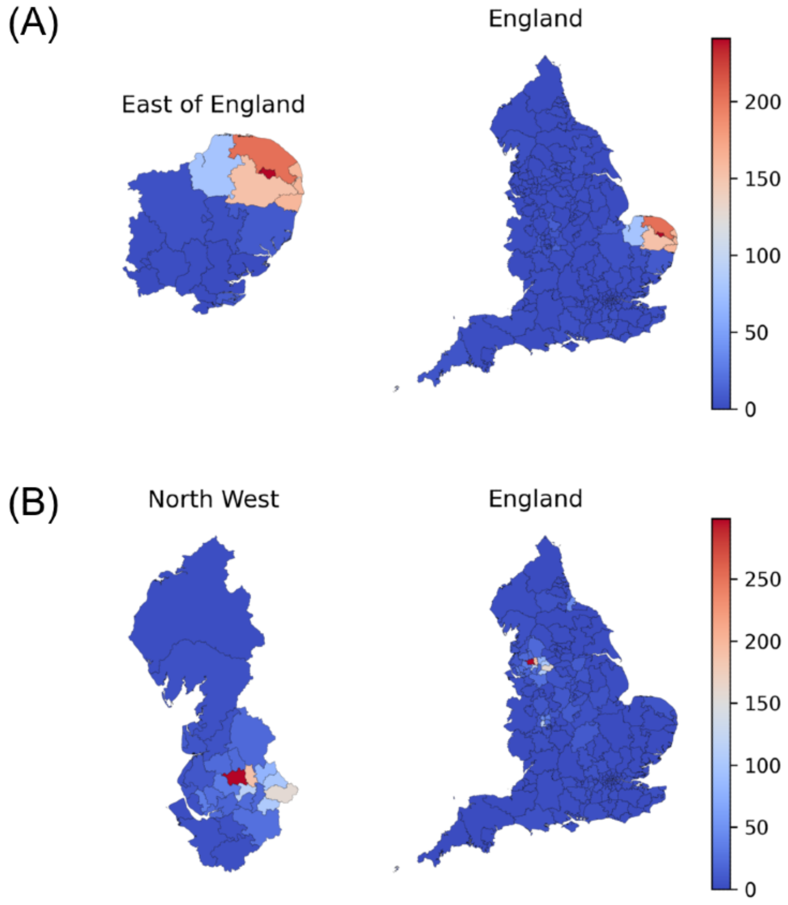


Table 1. National and regional prescribing counts and rates (per 1000) of pericyazine (N=15,344) and promazine hydrochloride (N=27,724) in all regions in England (June-August 2017). Prescribing rates for both chemicals are expressed per 1000 antipsychotic prescriptions (N=2,489,069).

Region	Antipsychotics, n	Pericyazine		Promazine hydrochloride	
		Prescribed, n	Per 1000	Prescribed, n	Per 1000
East Midlands	188,593	155	0.8	1088	5.8
East of England	277,470	13,119	47.3	381	1.4
Kent, Surrey, and Sussex	189,490	192	1.0	334	1.8
North Central and East London	151,108	84	0.6	122	0.8
North East	162,212	96	0.6	620	3.8
North West	412,624	315	0.8	20,060	48.6
North West London	89,949	5	0.1	100	1.1
South London	124,296	103	0.8	127	1.0
South West	209,838	306	1.5	131	0.6
Thames Valley	75,489	10	0.1	163	2.2
Wessex	121,442	104	0.9	276	2.3
West Midlands	246,907	580	2.3	3951	16.0
Yorkshire and the Humber	239,651	275	1.1	371	1.5
All	2,489,069	15,344	6.2	27,724	11.1

Figure 2. Total number of prescriptions for (A) pericyazine (B) promazine hydrochloride per 1000 antipsychotic prescriptions for all CCGs in England in June-August 2017. The colour scale in each plot indicates the number of prescriptions per 1000 antipsychotic prescriptions in the corresponding geographic region.



Discussion

Summary

Using a hypothesis-free approach, we have applied data science techniques to a national data set to identify outliers. Following subsequent clinical review, 2 unusual antipsychotic medications, in very limited use nationally, that are very commonly prescribed in 2 small geographic regions of England were identified. Specifically, pericyazine makes up only 0.6% (15,344 of 2,489,069) of all antipsychotic prescriptions nationally, but in Norwich it represents 24.1% (5197 of 21,553) of all antipsychotic prescriptions; promazine hydrochloride makes up 1.1% (27,724 of 2,489,069) of all antipsychotic prescriptions nationally; however, in Bolton, it represents 29.9% (5549 of 18,577) of all antipsychotic prescriptions.

Strengths and Weaknesses

This study is a proof of concept with a pragmatic and exploratory approach to the methodology and is still under iterative development with regard to optimizing metrics and parameters. As such, we recognize that there are limitations in the metrics that we have used to rank the chemical-class pairs as described here. For example, it is possible that the rankings being generated could be misleading where a small number of CCGs are under prescribers for particular chemicals, thereby inflating the outlier status of the same chemical in other areas. Furthermore, the effect of variability where the number of prescriptions is small is not yet known, although we do seek to mitigate against this by removing chemicals that are prescribed at particularly low volumes. However, we do not present this work as a stand-alone method for outlier detection; rather, we present it as an approach to prioritize and focus on manual clinical audit and review.

Our study does cover a reasonably short period of time (June-August 2017), again owing to it being a proof of concept. However, the OpenPrescribing data set used does include all prescribing in all typical practices in England, thereby minimizing the potential for obtaining a biased sample. Furthermore, the chemicals identified using our approach do represent legitimate targets for further investigation; unfortunately, our reporting of this work and subsequent investigations into the reasons for these prescribing outliers were disrupted by the COVID-19 pandemic.

Findings in Context

Pericyazine has been used infrequently for schizophrenia and for short-term adjunctive management of severe anxiety, psychomotor agitation, and violent or dangerously impulsive behavior [15]. There is no mention of pericyazine in any guideline on the National Institute for Health and Care Excellence (NICE) website, the main source of clinical guidelines in England. A 2014 Cochrane review on pericyazine identified only 5 studies suitable for inclusion, could not determine the effect of pericyazine in schizophrenia given the low quality of evidence, and found a higher incidence of side effects compared to atypical antipsychotics [16]. A PubMed search identified only 73 publications that contain the word “pericyazine” [17] in any way since 1965 compared with over 22,000 results for haloperidol and 11,000 for risperidone. Promazine hydrochloride is licensed in psychomotor agitation and agitation or restlessness in the older adults [18]. It is not mentioned in any NICE guideline and appears in only 1355 PubMed records [19] (peaking in 1964). We are aware of no prior work using data science techniques hypothesis-free to systematically identify outliers for any given treatment choice or clinical outcome in the manner outlined here.

Policy Implications and Interpretation

We report only the fact of a substantial deviation from national prescribing norms in these 2 small regions and make no direct comment on the appropriateness of using these medications in any single patient or in general. It was outside the scope of this work to engage in a detailed qualitative or other study to understand the reasons for the high usage of these 2 unusual antipsychotics in these 2 regions; however, we note that promazine hydrochloride and pericyazine have previously appeared in treatment formularies for Greater Manchester and Norfolk, respectively. In addition, it is noted that antipsychotic medication is typically initiated in secondary care, with prescribing taken over in general practice.

The Department of Health and Social Care recently consulted on an ambitious plan to harness data to improve health delivery and outcomes [20]. The use of data to identify variation in clinical activity and outcomes is long established [21,22], and recent flagship projects in the NHS such as RightCare and Getting it Right First Time are focused on identifying and addressing variation in care. However, these approaches typically rely on a traditional approach, whereby desirable clinical activities or outcomes are prospectively defined by clinicians or commissioners, and adherence is then measured by analysing relevant data. It is highly unlikely that these conventional methods would ever have identified the unusual prescribing behaviors reported in this paper. Similarly, it is likely that there are many further clinically interesting signals that could be identified by taking a variety of data-driven approaches to detecting unusual clinical activity or outcomes across the full universe of NHS data.

In our experience of running OpenPrescribing.net, the key barrier to better use of data for service improvement is an unhelpful cultural and practical divide between purely academic work on health data, and practical use of data in service analytics. This is exemplified by, in general, the use of different teams, different funding mechanisms, different institutions, and different data infrastructures. As the methods, data, and overarching objectives of both domains overlap substantially, we hope that funders and commissioners can help to bring these strands of work together.

Future Research

These findings will contribute to a wider program of work, which aims to develop a range of interactive tools on OpenPrescribing.net to present candidate signals of interest for substantial divergence from national prescribing norms at the level of individual practices, CCGs, and other key NHS organizational groupings such as primary care networks and integrated care systems. For this web-based service, we expect to present a wide variety of signals at scale, without further context on evidence or guidelines, as a trigger for positive local discussion and further exploration by clinical or commissioning teams, and inviting feedback on whether they found the signals to be helpful in identifying any previously unrecognized opportunities to change local prescribing practices or understanding the reasons for any divergences.

Conclusions

We describe a hypothesis-free approach to identify candidates for audit and review in clinical practice, with examples highlighted of 2 very unusual antipsychotics used disproportionately in 2 small geographic areas of England.

Acknowledgments

This work was supported by the National Institute for Health Research (NIHR) Biomedical Research Centre, Oxford, a Health Foundation grant (award reference number 7599); a NIHR School of Primary Care Research grant (award reference number 327); the NIHR under its Research for Patient Benefit Programme (grant reference number PB-PG-0418-20036) and by the NIHR Applied Research Collaboration Oxford and Thames Valley. The OpenPrescribing project is currently funded by NHS England, with additional funding support for specific topics from the National Institute for Health Research, and the Health Foundation. Phase 1 of this project leading to beta launch in December 2015 was funded by the West of England Academic Health Science Network. The views expressed in this publication are those of the authors and not necessarily those of the NIHR, NHS England, or the Department of Health and Social Care. The funders had no role in the study design, collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Data Availability

This study uses exclusively open, publicly available data. Processed data is additionally available within the study repository along with all analysis code [14].

Authors' Contributions

BG conceived the study with input from SJWE. HJC and AJW designed the methods. HJC and LEMH collected and analyzed the data with methodological and interpretation input from AJW, BMK, RC, and BG. HJC, AJW, and SCJB all directly accessed and verified the underlying data, as published by NHS Business Services Authority. All authors confirm that they had full access to all the data in the study and accept responsibility to submit for publication. BMK drafted the manuscript with input from RC, HJC, SJWE, BG, OM, and JM. LEMH prepared the manuscript for final submission. All authors contributed and approved the final manuscript. SCJB was the lead engineer on the associated website resource with input from DE and PI. BG supervised the project and is the guarantor.

Conflicts of Interest

All authors have completed the ICMJE (International Committee of Medical Journal Editors) uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare the following: BG has received research funding from the Laura and John Arnold Foundation, the NHS National Institute for Health Research (NIHR), the NIHR School of Primary Care Research, the NIHR Oxford Biomedical Research Centre, the Mohn-Westlake Foundation, NIHR Applied Research Collaboration Oxford and Thames Valley, Wellcome Trust, the Good Thinking Foundation, Health Data Research UK, the Health Foundation, the World Health Organization, UKRI, Asthma UK, the British Lung Foundation, and the Longitudinal Health and Wellbeing strand of the National Core Studies program; he also receives personal income from speaking and writing for lay audiences on the misuse of science. BMK and OM work for the NHS and are seconded to the Bennett Institute for Applied Data Science. All other University of Oxford authors are employed on BG’s grants.

Multimedia Appendix 1
Pericyazine prescriptions in the East of England (June-August 2017).
[XLSX File (Microsoft Excel File), 18 KB - medinform_v10i12e41200_app1.xlsx]

Multimedia Appendix 2
Promazine hydrochloride prescriptions in the North West of England (June-August 2017).
[XLSX File (Microsoft Excel File), 18 KB - medinform_v10i12e41200_app2.xlsx]

References

1. Goldacre B, MacKenna B. The NHS deserves better use of hospital medicines data. BMJ 2020 Jul 17;370:m2607. [doi: 10.1136/bmj.m2607] [Medline: 32680848]
2. Explore England's prescribing data. OpenPrescribing.net. URL: https://openprescribing.net/ [accessed 2022-10-18]

3. Croker R, Walker AJ, Bacon S, Curtis HJ, French L, Goldacre B. New mechanism to identify cost savings in English NHS prescribing: minimising 'price per unit', a cross-sectional study. BMJ Open 2018 Feb 08;8(2):e019643 [FREE Full text] [doi: 10.1136/bmjopen-2017-019643] [Medline: 29439078]
4. Walker AJ, Curtis HJ, Croker R, Bacon S, Goldacre B. Measuring the impact of an open web-based prescribing data analysis service on clinical practice: cohort study on NHS England data. J Med Internet Res 2019 Jan 16;21(1):e10929. [doi: 10.2196/10929]
5. Croker R, Walker A, Goldacre B. Why did some practices not implement new antibiotic prescribing guidelines on urinary tract infection? A cohort study and survey in NHS England primary care. J Antimicrob Chemother 2019 Apr 01;74(4):1125-1132. [doi: 10.1093/jac/dky509] [Medline: 30590552]
6. Curtis HJ, Walker AJ, MacKenna B, Croker R, Goldacre B. Prescription of suboptimal statin treatment regimens: a retrospective cohort study of trends and variation in English primary care. Br J Gen Pract 2020 Jun 29;70(697):e525-e533. [doi: 10.3399/bjgp20x710873]
7. Aral KD, Güvenir HA, Sabuncuoğlu I, Akar AR. A prescription fraud detection model. Comput Methods Programs Biomed 2012 Apr;106(1):37-46. [doi: 10.1016/j.cmpb.2011.09.003] [Medline: 22088866]
8. Bucholc M, O'Kane M, Ashe S, Wong-Lin K. Prescriptive variability of drugs by general practitioners. PLoS One 2018 Feb 20;13(2):e0189599 [FREE Full text] [doi: 10.1371/journal.pone.0189599] [Medline: 29462143]
9. Hirsch O, Donner-Banzhoff N, Schulz M, Erhart M. Detecting and visualizing outliers in provider profiling using funnel plots and mixed effects models—an example from prescription claims data. Int J Environ Res Public Health 2018 Sep 15;15(9):2015 [FREE Full text] [doi: 10.3390/ijerph15092015] [Medline: 30223551]
10. Mordecai L, Reynolds C, Donaldson LJ, de C Williams AC. Patterns of regional variation of opioid prescribing in primary care in England: a retrospective observational study. Br J Gen Pract 2018 Feb 12;68(668):e225-e233. [doi: 10.3399/bjgp18x695057]
11. MacKenna B, Curtis HJ, Walker AJ, Croker R, Bacon S, Goldacre B. Trends and variation in unsafe prescribing of methotrexate: a cohort study in English NHS primary care. Br J Gen Pract 2020 Jun 22;70(696):e481-e488. [doi: 10.3399/bjgp20x710993]
12. Livesey JH. Kurtosis provides a good omnibus test for outliers in small samples. Clin Biochem 2007 Sep;40(13-14):1032-1036. [doi: 10.1016/j.clinbiochem.2007.04.003] [Medline: 17499683]
13. Verma SP, Díaz-González L, Rosales-Rivera M, Quiroz-Ruiz A. Comparative performance of four single extreme outlier discordancy tests from Monte Carlo simulations. Sci World J 2014;2014:746451 [FREE Full text] [doi: 10.1155/2014/746451] [Medline: 24737992]
14. Outliers by kurtosis: repository. GitHub. URL: https://github.com/ebmdatalab/kurtosis-pericyazine [accessed 2022-10-18]
15. Summary of product characteristics—pericyazine 10mg tablets. Zentiva. URL: https://www.medicines.org.uk/emc/product/3969/smpc [accessed 2022-10-18]
16. Matar H, Almerie M, Makhoul S, Xia J, Humphreys P. Pericyazine for schizophrenia. Cochrane Database Syst Rev 2014;5:CD007479. [doi: 10.1002/14651858.cd007479.pub2]
17. Keyword search for “pericyazine.”. PubMed. URL: https://web.archive.org/web/20210823105333/https://pubmed.ncbi.nlm.nih.gov/?term=pericyazine&sort=date [accessed 2022-10-18]
18. Summary of Product Characteristics - Promazine 25mg film-coated tablets. Teva UK Limited. 2022 Feb 16. URL: https://products.tevauk.com/mediafile/id/52087.pdf [accessed 2022-10-31]
19. Keyword search for “promazine.”. PubMed. URL: https://web.archive.org/web/20210823110138/https://pubmed.ncbi.nlm.nih.gov/?term=promazine&sort=date [accessed 2022-10-18]
20. Data saves lives: reshaping health and social care with data. Department of Health & Social Care, UK. URL: https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data [accessed 2022-10-18]
21. Glover JA. The incidence of tonsillectomy in school children. Proc R Soc Med 2016 Dec 06;31(10):1219-1236. [doi: 10.1177/003591573803101027]
22. Fisher ES, Wennberg DE, Stukel TA, Gottlieb DJ, Lucas FL, Pinder EL. The implications of regional variations in Medicare spending. Part 1: the content, quality, and accessibility of care. Ann Intern Med 2003 Feb 18;138(4):273-287 [FREE Full text] [doi: 10.7326/0003-4819-138-4-200302180-00006] [Medline: 12585825]

Abbreviations

BNF: British National Formulary
CCG: Clinical Commissioning Group
NICE: National Institute for Health and Care Excellence
NHS: National Health Service

Edited by C Lovis; submitted 19.07.22; peer-reviewed by F Moriarty; comments to author 11.09.22; revised version received 14.10.22; accepted 16.10.22; published 20.12.22.

Please cite as:
MacKenna B, Curtis HJ, Hopcroft LEM, Walker AJ, Croker R, Macdonald O, Evans SJW, Inglesby P, Evans D, Morley J, Bacon SCJ, Goldacre B
Identifying Patterns of Clinical Interest in Clinicians' Treatment Preferences: Hypothesis-free Data Science Approach to Prioritizing Prescribing Outliers for Clinical Review
JMIR Med Inform 2022;10(12):e41200
URL: https://medinform.jmir.org/2022/12/e41200
doi:10.2196/41200
PMID:36538350

Corrigenda and Addenda

Correction: The Application of Graph Theoretical Analysis to Complex Networks in Medical Malpractice in China: Qualitative Study

Shengjie Dong[1,2], MPH; Chenshu Shi[3], MSc; Wu Zeng[4], PhD; Zhiying Jia[1,5], MPH; Minye Dong[1], PhD; Yuyin Xiao[1], MPH; Guohong Li[1,6], PhD

[1]School of Public Health, Shanghai Jiao Tong University, Shanghai, China
[2]Shuguang Hospital Affiliated to Shanghai University of Traditional Chinese Medicine, Shanghai, China
[3]Center for Health Technology Assessment, China Hospital Development Institute, Shanghai Jiao Tong University, Shanghai, China
[4]Department of International Health, Georgetown University, Washington, DC, United States
[5]Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China
[6]China Hospital Development Institute, Shanghai Jiao Tong University, Shanghai, China

Corresponding Author:
Guohong Li, PhD
School of Public Health, Shanghai Jiao Tong University
No. 227 South Chongqing Road, Huangpu District
Shanghai, 200025
China
Phone: 86 21 63846590
Email: guohongli@sjtu.edu.cn

Related Article:

Correction of: https://medinform.jmir.org/2022/11/e35709

(JMIR Med Inform 2022;10(12):e44374) doi:10.2196/44374

In “The Application of Graph Theoretical Analysis to Complex Networks in Medical Malpractice in China: Qualitative Study” (JMIR Med Inform 2022;10(11):e35709) the authors noted one error:

The phone number of the corresponding author Guohong Li was corrected to 86 21 63846590.

The correction will appear in the online version of the paper on the JMIR Publications website on December 7, 2022 together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 24.11.22; this is a non–peer-reviewed article;accepted 01.12.22; published 07.12.22.

Please cite as:
Dong S, Shi C, Zeng W, Jia Z, Dong M, Xiao Y, Li G
Correction: The Application of Graph Theoretical Analysis to Complex Networks in Medical Malpractice in China: Qualitative Study
JMIR Med Inform 2022;10(12):e44374
URL: https://medinform.jmir.org/2022/12/e44374
doi:10.2196/44374
PMID:36476496

Original Paper

A Privacy-Preserving Distributed Medical Data Integration Security System for Accuracy Assessment of Cancer Screening: Development Study of Novel Data Integration System

Atsuko Miyaji[1,2*], PhD; Kaname Watanabe[3,4*], MD, PhD; Yuuki Takano[1], PhD; Kazuhisa Nakasho[5], PhD; Sho Nakamura[3,6], MD, PhD; Yuntao Wang[1], PhD; Hiroto Narimatsu[3,4,6], MD, PhD

[1]Graduate School of Engineering, Osaka University, Suita, Japan

[2]Japan Advanced Institute of Science and Technology, Nomi, Japan

[3]Cancer Prevention and Control Division, Kanagawa Cancer Center Research Institute, Yokohama, Japan

[4]Department of Genetic Medicine, Kanagawa Cancer Center, Yokohama, Japan

[5]Graduate School of Science and Technology for Innovation, Yamaguchi University, Ube, Japan

[6]Graduate School of Health Innovation, Kanagawa University of Human Services, Kawasaki, Japan

[*]these authors contributed equally

Corresponding Author:
Kaname Watanabe, MD, PhD
Cancer Prevention and Control Division
Kanagawa Cancer Center Research Institute
2-3-2 Nakao, Asahi-ku
Yokohama, 241-8515
Japan
Phone: 81 45 520 2222 ext 4020
Fax: 81 45 520 2216
Email: ka-watanabe@gancen.asahi.yokohama.jp

Abstract

Background: Big data useful for epidemiological research can be obtained by integrating data corresponding to individuals between databases managed by different institutions. Privacy information must be protected while performing efficient, high-level data matching.

Objective: Privacy-preserving distributed data integration (PDDI) enables data matching between multiple databases without moving privacy information; however, its actual implementation requires matching security, accuracy, and performance. Moreover, identifying the optimal data item in the absence of a unique matching key is necessary. We aimed to conduct a basic matching experiment using a model to assess the accuracy of cancer screening.

Methods: To experiment with actual data, we created a data set mimicking the cancer screening and registration data in Japan and conducted a matching experiment using a PDDI system between geographically distant institutions. Errors similar to those found empirically in data sets recorded in Japanese were artificially introduced into the data set. The matching-key error rate of the data common to both data sets was set sufficiently higher than expected in the actual database: 85.0% and 59.0% for the data simulating colorectal and breast cancers, respectively. Various combinations of name, gender, date of birth, and address were used for the matching key. To evaluate the matching accuracy, the matching sensitivity and specificity were calculated based on the number of cancer-screening data points, and the effect of matching accuracy on the sensitivity and specificity of cancer screening was estimated based on the obtained values. To evaluate the performance, we measured central processing unit use, memory use, and network traffic.

Results: For combinations with a specificity ≥99% and high sensitivity, the date of birth and first name were used in the data simulating colorectal cancer, and the matching sensitivity and specificity were 55.00% and 99.85%, respectively. In the data simulating breast cancer, the date of birth and family name were used, and the matching sensitivity and specificity were 88.71% and 99.98%, respectively. Assuming the sensitivity and specificity of cancer screening at 90%, the apparent values decreased to 74.90% and 89.93%, respectively. A trial calculation was performed using a combination with the same data set and 100% specificity. When the matching sensitivity was 82.26%, the apparent screening sensitivity was maintained at 90%, and the screening specificity decreased to 89.89%. For 214 data points, the execution time was 82 minutes and 26 seconds without parallelization and 11 minutes and 38 seconds with parallelization; 19.33% of the calculation time was for the data-holding institutions. Memory use was 3.4 GB for the PDDI server and 2.7 GB for the data-holding institutions.

Conclusions: We demonstrated the rudimentary feasibility of introducing a PDDI system for cancer-screening accuracy assessment. We plan to conduct matching experiments based on actual data and compare them with the existing methods.

(JMIR Med Inform 2022;10(12):e38922) doi:10.2196/38922

KEYWORDS

data linkage; data security; secure data integration; privacy-preserving linkage; secure matching privacy-preserving linkage; private set intersection; PSI; privacy-preserving distributed data integration; PDDI; big data; medical informatics; cancer prevention; cancer epidemiology; epidemiological survey

Introduction

Distributed Data Integration in Epidemiological Studies

With advances in information technology and enhanced data-collection systems, health databases are becoming increasingly abundant. Similar to other countries, the government and academic societies in Japan collect and manage a disease database. In addition, there are patient-based disease databases and population-based cohort study databases that are collected and managed mainly by research institutes [1-5]. Integrating health information held in these independent databases benefits epidemiological studies and public health practices; for example, it is possible to determine important correlations and causal relationships, such as between the onset of disease and the health status of an individual, which cannot be determined using a single database. Therefore, it is important to link databases managed by different institutions [6-8].

There are challenges associated with linking independent databases. The first is the guarantee of information privacy, including the handling of personally identifiable information. Concerns and considerations regarding privacy and data security are paramount; policies and regulations on the collection, use, and movement of personally identifiable information are becoming more stringent [9]. Therefore, in data linkage, sufficient measures to prevent the leakage of personal information are required, which have led to an increase in attendant costs, including labor. The second challenge is the construction of an efficient data linkage system. In countries where a unique identification key, such as the national identification number, is given to each individual and multiple medical or welfare-related data systems are linked, more efficient matching is possible compared with countries where such unique identifiers are not provided to every citizen. Nordic countries are representative of those using such unique identifiers. However, owing to privacy concerns, many issues need to be resolved before linking the databases; therefore, only a few countries have introduced such identifiers so far [10,11]. In countries where the unique identification key system has not been put into practical use, it is even more difficult to build a system that meets information privacy requirements and linkage efficiency. Consequently, it has been impossible to link databases managed by different institutions at a practical level in Japan.

Secure Data Integration

To safely and effectively collate the data held by each institution in a decentralized state and use them, it is desirable to exchange only necessary information as much as possible without leaking personal information to the outside. However, without a unique identification key, it is common to use personal information, such as name and date of birth, as the key to perform matching [9,12]. The methods that are widely practiced today include one in which a data provider or user performs a matching operation or the method in which a data set containing personal information is passed to a third party (data depository) to perform the matching. Both methods require the movement of personal information that serves as the key to carry out the match. Although some studies [13,14] related to the linkage between 2 databases have been conducted, they are still vulnerable in terms of security and privacy. In fact, in a report by Kho et al [13], a hash value of names was used to match names so that a dictionary attack can determine which hospital a patient is in. A dictionary attack is a method in which the hash values of a precreated patient list are matched with the hash values stored in a system database. As the hash values of a limited range of data, such as patient lists, are vulnerable to a dictionary attack, the use of simple hash tables should be avoided. Furthermore, the proposal by Kho et al assumes that the database is owned by a single institution. In a report by Godlove et al [14], the system and other details were not described; therefore, the method of matching is a black box.

Therefore, strict countermeasures against information leakage and the costs involved are obstacles to conducting large-scale epidemiological studies. There are technical efforts to more securely approach a solution to this issue. Under the private set intersection protocol, which has been attracting attention in recent years, data other than those commonly included in data sets, distributed and managed by multiple data-holding institutions, are kept secret from other institutions; hence, only commonly included data are accessible [15-18]. The technology discussed in a previous report [18], which is an extension of private set intersection, focuses on the fact that a data set of medical-related information is generally composed of multiple attributes. After specifying an attribute as the matching key, the data associated with the same key attribute commonly included in each institution are integrated. It is called privacy-preserving distributed data integration (PDDI) because it integrates distributed data while ensuring privacy. Notably, unlike the proposal by Kho et al [13], PDDI does not simply match in the hash values of matching keys; therefore, information on whether a given patient is included in an institution is not available, and unlike Godlove et al [14], the specification is not a black box but is obvious. Studies on the application of newly developed PDDI systems to medical data are ongoing [19]. The PDDI system is expected to enable the secure integration of health information held in databases managed by different institutions and to enable epidemiological studies to be conducted with high security.

Challenges in Implementing the Technology

PDDI is an established technology, but several additional steps must be taken before its implementation. The most important aspect is to show that the system can maintain sufficient matching accuracy and performance for operational purposes while keeping personal information secure, even when using actual data. The matching keys that are commonly used when a national identification number or similar identifier is not available, such as name and date of birth, include various errors, such as typing errors, at the time of input and orthographic variants owing to differences in the input format. Especially, in Japan, the lack of a standardized identification format also contributes to this effect. Therefore, the identification of identical persons tends to be associated with a certain rate of failure, lowering the matching accuracy [20]. Low matching accuracy affects outcome detection and narrows the research design and research themes to which the system can be applied. Matching accuracy is determined by the quantity and nature of such errors and the matching method [21,22]. The errors that can be found in data types used as matching keys are also affected by the language and characters used in the description. The optimal method for addressing these errors must be considered separately for different countries, regions, and databases. Various strategies have been developed to increase the reliability of matching. These include prior data cleaning, standardizing formats, combining personal information that serves as matching keys, and taking various measures such as probabilistic approaches [9,12,23,24]. However, it is unclear, especially in Japan, which data items can be used as matching keys to maximize the matching accuracy where a unique matching key cannot be used. The other aspect is the system performance. PDDI systems do not consolidate the data of each institution to 1 depository institution. The information held by each institution is encrypted within that institution, and the data are collected and distributed. However, the specifications of computer terminals of data-holding institutions and users vary considerably. Therefore, it is necessary to evaluate the performance of a linkage system for its stable use in a general-purpose environment.

The purpose of this project was to demonstrate that the security of personal information can be maintained in matching using actual data and that it is operationally accurate and performs significantly well for PDDI implementation and to identify which data items can be effective matching keys to perform data matching with high accuracy in situations where there is no unique matching key. However, because the use of personal information as a matching key is strictly controlled in Japan, a preliminary experiment was required using dummy data to experiment using actual data. In this study, we evaluated the protection of personal information, matching accuracy in cancer-screening accuracy assessment assuming a large-scale epidemiological study using artificially created data that simulate cancer screening and cancer registration data. If feasibility is confirmed in this study, we plan to carry out a verification study using actual data. The results of these studies are expected to be applied to large-scale population-based genomic cohort studies and large-scale studies using patient databases, thus contributing to further activation and development of database-based epidemiological research.

Methods

PDDI System

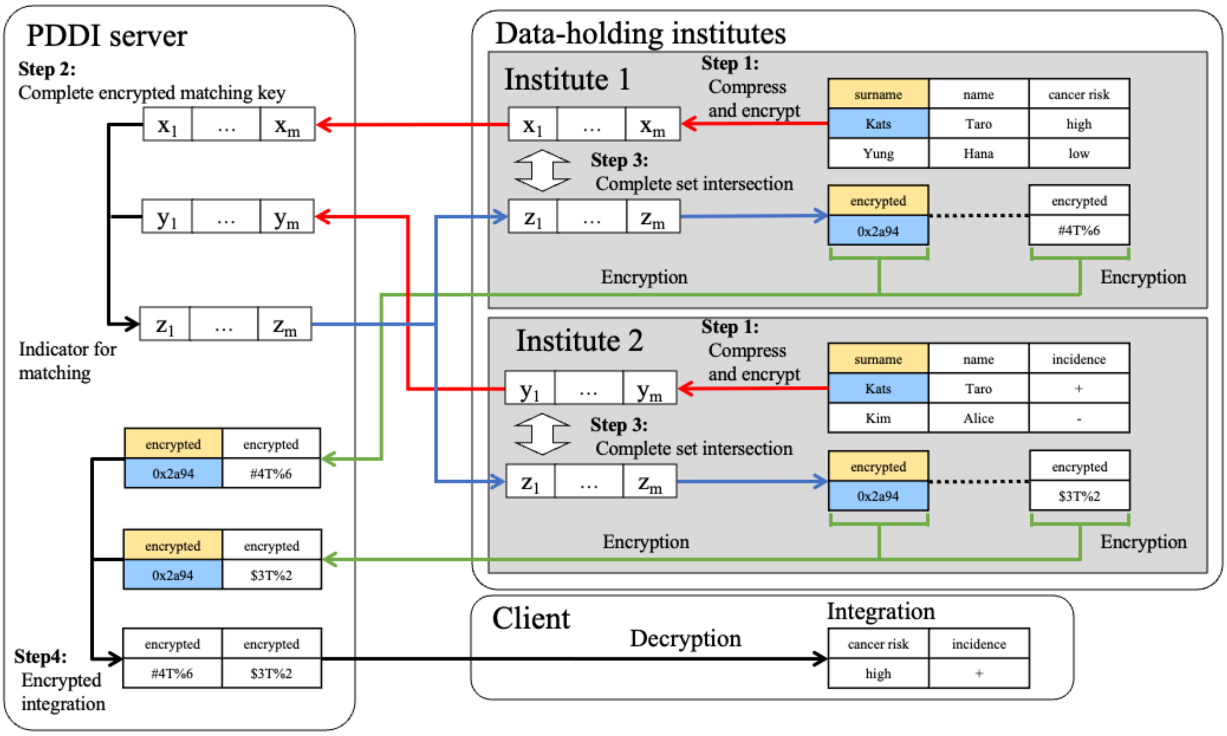
Overview

The features of PDDI used in this study are presented in our previous study [19], in which it is shown that PDDI consists of a secure computation server, data-holding institutions, and client. In PDDI systems, when there are multiple attributes per data sample, the database is divided into 3 types: key information, analysis target data, and *others*. The data to be analyzed, which are linked to the key commonly included in the database of each institution, are concealed and integrated. The key information and data to be analyzed may match. Important characteristics of PDDI systems are as follows:

1. No institution that uses the system, including those that own databases and those that receive data, can obtain any information other than the key information that is commonly shared between databases. Unlike the query-based method, the fact that 1 institution holds some information about the individual is not divulged to any other institution.
2. Key information used to match the data will not be divulged to any institution, including the PDDI secure computation server. In this paper, the PDDI secure computation server is denoted as PDDI server.
3. The processing time of each institution does not depend on the number of institutions involved in the system. There is no limit to the data available to each institution through the system.
4. No third-party institution collects or aggregates data to carry out matching.

We have described the PDDI algorithm in subsequent sections. Figure 1 shows the entire algorithmic process.

Figure 1. Schematic of the privacy-preserving distributed data integration (PDDI) system algorithm. Steps 1 to 4 represent each step of the merging process using the PDDI system described in the main text. The data held by each institution are encrypted and matched by the PDDI server using the data as the matching key. The analysis target data, which are related to the matching key without distinction between institutions, are decrypted only when they are provided to the client, and the matching-key information is never provided to the client.



Step 1: Irreversible Compression and Encryption

Each institution compresses the key used for collating the data set with a hash function, converts it into unique and irreversible information, and sends the data encrypted by homomorphic and probabilistic encryption to the PDDI server.

Step 2: Creation of Matching Keys

The PDDI server calculates the sum of the encrypted data obtained from each institution (called an encrypted matching key) and sends these to each institution. Note that the PDDI server does not have the decryption key; therefore, it cannot decrypt the encrypted matching key.

Step 3: Analysis of Target Data for Set Intersection Computation

Each institution decrypts the received encrypted matching key and obtains the matching key used for extracting the key that is commonly included in all institutions. Next, the analysis target data related to the commonly included key are encrypted and sent to the PDDI server.

Step 4: Integration of Encrypted Analysis Target Data

The server integrates the encrypted analysis target data sent from each institution and sends it to the client; the matching-key information is not sent to the client. In this study, 1 data-holding institution evaluates whether the matching was performed correctly; therefore, the data-holding institution acts as a client.

These matching keys are transformed into Bloom filters and then encrypted in each institution. The encryption is probabilistic, and thus, the same plaintext is encrypted into different values. Furthermore, it cannot be decrypted without the collaboration of all institutions. Then, they are sent to the PDDI server. Note that the encryption of the compressed matching key is probabilistic, which implies that the ciphertexts of the compressed matching keys are not equal even if the compressed matching keys are equal. Therefore, by using the ciphertext, anyone cannot guess whether a patient with the matching key is included in the institute, unlike the proposal by Kho et al [13]. For the same reason, the PDDI server neither reveals any information of the matching key in each institution nor guesses whether a patient with the matching key is included in the institute. This is a completely different privacy policy from that proposed by Kho et al [13].

The PDDI implementation environment, environment construction, and usability are described in Multimedia Appendix 1. The basic part of this system (code, encryption, and others) is currently being prepared for publication.

Experiment Model: Accuracy Assessment of Cancer Screening

Overview

In this study, we adopted accuracy assessment of cancer screening as a model for the matching experiment. Cancer screening is a general term for cancer-screening programs for the general population, which are conducted to reduce the mortality rate owing to early detection of cancer (secondary prevention). It is implemented around the world, centered on programs that have been scientifically recognized to reduce mortality, such as breast, cervical, and colorectal cancers [25-27]. The examinee is evaluated for the risk of having cancer based on the test results of each program. Patients who are determined to be at high risk, that is, those who are highly suspected of having cancer, are encouraged to visit a medical institution. Assessing the accuracy of cancer risk detection and controlling the quality of screening, so that the number of overlooked cancers and useless tests is kept to a minimum, constitute the major roles of cancer-screening accuracy control. Data on whether a patient who was judged to be at high risk in a program had cancer within a certain period (often 1-2 years) are required to assess the accuracy of cancer screening.

The biggest challenge in assessing cancer-screening accuracy is the collection and matching of distributed data. In many cases, cancer incidence, which represents the outcome of screening, needs to be obtained by matching with another source independent of the cancer-screening database; for example, a cancer registration database. In Japan, cancer-screening data are managed in a distributed state by the municipalities that are the implementing bodies. Moreover, cancer registration data are managed in a distributed manner by prefectures. Therefore, to collect and collate these data on a large-scale national or regional basis is difficult. The data size to be handled are large, and when there are many target municipalities, a lot of cumbersome procedures, which are not always standardized by the municipalities, are required to obtain the data. The greater the number of municipalities involved, the greater the movement of privacy information and the higher the risk of leakage. Therefore, in Japan, such studies are only conducted sporadically, using limited data from a small number of municipalities [28,29].

This system is characterized by no restrictions on the number of participating institutions or the amount of data held by the institutions and is considered an effective means for solving this problem. This system makes it easy to match the risk assessment information of distributed cancer screening with the cancer incidence information of cancer registration, which is expected to enable large-scale cancer-screening accuracy assessment, which has not yet been possible. Therefore, we surmised that applying a PDDI system for the assessment of cancer-screening accuracy is possible and devised an experimental plan using this model.

In cancer-screening accuracy assessment, indicators such as sensitivity, specificity, and positive predictive value are mainly used. If cancer screening indicates that there is a strong suspicion of having cancer (high risk), it is considered positive. In Japan, it is recommended to visit a medical institution, so this result is often called a “requiring detailed examination.” The other judgments are negative. Whether the patient has cancer is evaluated by comparing cancer incidence information in cancer registration data for 1 to 2 years from the date of consultation with the screening result. In other words, if the cancer screen is positive (there is a strong suspicion that the patient has cancer) and the cancer is subsequently diagnosed, the sensitivity, specificity, and positive predictive value in the context of assessment of the accuracy of cancer screening are defined as Textbox 1.

Textbox 1. Definition of items related to the accuracy of cancer screening

- Screening sensitivity=Proportion of patients with cancer who screen positive
- Screening specificity=Proportion of patients without cancer who screen negative
- Positive predictive value for screening=Proportion of cases giving positive screen results who are already patients

The accuracy of cancer screening is indicated by adding “screening” to distinguish it from the accuracy of matching, which will be described in the “study design” section.

Background of Practical Data-Matching Failures

In countries that do not have a national identification number, such as Japan, data are generally collated using personal information. In such an environment, the accuracy of matching is reduced owing to various errors that may appear in the data points used as matching keys. The sources of errors when using matching keys are careless mistakes, orthographic variance owing to changes in culture and institutions, and differences in notation. The matching-key information may also change: change of address because of moving and renaming because of marriage. The prevalence of errors varies depending on the format adopted by the data holder and ability of the input person. They are also heavily influenced by the language in which the data are written. Japanese is the de facto official language in Japan, where we live, and it is adopted as the default language in most systems and services in Japan. Many errors in Japanese registry data are due to language-specific problems. Details of the errors originating from Japanese language features are described in Multimedia Appendix 2.

Study Design

As mentioned in the Introduction section, the purpose of this project is to demonstrate the safety, accuracy, and performance of data matching using the PDDI system and to identify effective data items as matching keys. This study is the first step of the project. We used the PDDI system to perform a data set matching experiment between simulated cancer-screening and cancer registration data sets, in which the PDDI system was tasked with matching data belonging to the same individuals between the sets. Feasibility was evaluated based on data security, matching accuracy (sensitivity and specificity), and system performance.

In this experiment, we performed matching under multiple conditions using personal information, such as first and last names, phonetic spelling, date of birth, and address, and evaluated how much matching accuracy could be obtained by combining matching keys. Various matching algorithms were devised to prevent a decrease in sensitivity while maintaining specificity [9,12,23]. However, the purpose of this study was to evaluate the PDDI system, not the novel matching method, to improve the matching accuracy; therefore, these advanced matching algorithms were not considered. Methods for more accurate and practical matching will be considered in the next steps of this project. Instead, we estimated how much the matching accuracy would affect the estimation of cancer-screening accuracy. The feasibility of applying the model in this study was evaluated.

Unlike conventional systems that use a simple hash function to compress privacy information or that require a single server to collect and process all data, our system uses the latest security techniques. For example, all data through the network are encrypted, and decryption cannot be performed by a single institution but only by the cooperation of all distributed institutions, without centralizing the data. Therefore, it is important to verify that it can be implemented on a general-purpose computer rather than on a special server. We evaluated the performance of the system, the total data processing time, memory use, and network traffic required by PDDI. The PDDI server was introduced to reduce the processing time and amount of communication between data-holding institutions. In practice, the data processing time of data-holding institutions and the total data processing time required to collect the information contained in common is of critical importance.

Setting of the Matching Experiment

Four data sets were created to simulate cancer-screening and cancer registration data for 2 types of cancers: colorectal and breast cancers. First, using the web-based test-data generation service that is open to the public in Japan, we created pseudodata that included name, gender, date of birth, and address to serve as matching-key information [30-32]. This service automatically creates personal information, such as name, date of birth, address, and telephone number, from random combinations, which is common in Japan. By selecting the required information items and the desired amount of generated data, the user can obtain data that simulate nonexistent personal information. To account for the possibility that data generated by any particular service may contain certain tendencies or biases, we generated one-third of all the data points from each of the 3 separate services. Next, from the created pseudodata, 60 cases of colorectal cancer and 62 cases of breast cancer were selected as common data that can be matched. These were commonly included in both cancer-screening and cancer registration data sets. To make the simulated data resemble the actual data, we consulted the staff who had abundant experience in registry management and a physician who is an expert in epidemiological research, and the data were modified to include errors and orthographic variants that are often empirically recognized. Experience shows that the number of errors in the data set is expected to be <10%. Previous studies have reported that the number of errors and omissions in the data available for matching keys in disease registries and medical and administrative databases is approximately 15% or less [33-35]. However, the actual prevalence of errors is unknown, as changes in culture and society are expected to affect their occurrence rates. Therefore, to create data that would be more difficult to match, the data were rewritten to increase the number of errors to the extent that a data point would have errors in multiple items. Errors were made more prevalent in the colorectal cancer data set than in the breast cancer data set such that the colorectal cancer data set would be more difficult to match than the breast cancer data set. Subsequently, the remaining pseudodata were added, and finally, a pseudo–data set of 2000 colorectal cancer screenings, 17,866 colorectal cancers, 1048 breast cancer screenings, and 29,949 breast cancers was created. Pseudodata items other than matching keys included serial numbers and pseudoidentification numbers for each database in all data sets. The following pseudodata were randomly added to the colorectal cancer-screening data set: test date, test results, and risk assessment of fecal occult blood test, which is commonly used in Japan. The diagnosis name; International Classification of Diseases, Tenth Revision code; and date of diagnosis were added to the cancer registration data set. Pseudodata items other than these matching keys were only decorative and did not affect the matching experiment. Table 1 lists the errors and orthographic variants added to the data set. The examples of errors specific to Japanese in the data sets used in the experiments in this study are shown in Figure S1 in Multimedia Appendix 2.

Table 1. Errors and orthographic variants included in the data set.

Class, error type, and matching key	Number of data points, n (%)	
	Colorectal cancer (n=60)	Breast cancer (n=62)
Data entry errors		
Typing errors		
Name	3 (5)	1 (2)
Birth date	15 (25)	0 (0)
Address	6 (10)	2 (3)
Sex	5 (8)	0 (0)
Kanji conversion errors		
Name	5 (8)	6 (10)
Address	2 (3)	0 (0)
Misreading		
Name	10 (17)	8 (13)
Missing letters		
Name	2 (3)	1 (2)
Omission		
Address	4 (7)	0 (0)
Name	10 (17)	1 (2)
Orthographic variants		
Variant kanji		
Name	7 (12)	4 (6)
Format		
Address	5 (8)	15 (24)
Data change		
Name change		
Name	2 (3)	1 (2)
Alias		
Name	2 (3)	0 (0)
Moving		
Address	2 (3)	8 (13)
Unmatched on multiple keys	25 (42)	14 (23)
Total	51 (85)	36 (59)

In the experiment, 6 pieces of information—family name (kanji or kana), first name (kanji or kana), date of birth, and gender—were used. In this experiment, matching was performed by combining ≥2 images. In the case of colorectal cancer, 57 combinations were possible: ${}_6C_2 + {}_6C_3 + {}_6C_4 + {}_6C_5 + {}_6C_6$. For breast cancer, outside of a small number of exceptional cases, all screening targets were females, and thus, only 26 combinations were possible: ${}_5C_2 + {}_5C_3 + {}_5C_4 + {}_5C_5$.

In the PDDI protocol, a data array called a Bloom filter is encrypted element by element. More than 90% of the total execution time is spent on this encryption process. The encryption of an element of the data array is independent of that of other elements, and parallelization is easy. The multiprocessing module in Python Standard Library (version 3.9; Python Software Foundation) was used for this parallelization. The PC environment used in the experiment was as follows: central processing unit (CPU), Intel (R) Xeon (R) CPU E5-2690 v4@2.60GHz (28 cores), memory 48 GB. The programs of all the institutions were executed on 1 PC.

Evaluation

Items related to matching accuracy are referred to below with "matching" to distinguish them from the accuracy of cancer screening. To calculate the matching accuracy, the pseudo–cancer screening data were used as a reference point, and when the data matched the specified matching-key conditions in the pseudocancer registration data, the match was considered *positive*. The case in which no matching data were present was defined as *negative*. This matching experiment was conducted between data sets in which the same persons were simulated in both data sets in advance. Therefore, the trueness and falseness of matching were determined as follows: cases in which the matching result correctly matched data belonging to the same person were considered *true* and those in which the matching result did not correctly match data belonging to the same person were considered *false*. In other words, a *false positive* means that data originally registered under separate individuals were erroneously matched, and a *false negative* means that data that should have been matched (because they belong to the same person) were not matched. In an environment in which matching keys that uniquely identify an individual are completely error-free, matching is perfectly accurate. In this experiment, as an evaluation of matching accuracy, the correspondence between positive and negative matches and their trueness or falseness was cross-tabulated to calculate the matching sensitivity and matching specificity. On the basis of this, a combination of matching keys with high matching sensitivity and matching specificity, that is, good matching accuracy, was extracted.

For the estimation of the effect of matching accuracy on the assessment of cancer-screening accuracy, we referred to past studies and assumed 2 scenarios: one in which the true accuracy of cancer screening involved a sensitivity of 90% and a specificity of 90% and the other with a sensitivity of 60% and a specificity of 90% [36-38]. Errors between true and estimated values were calculated to assess screening sensitivity, screening specificity, and screening positive predictive value. For matching accuracy, simulations were carried out in the following manner: values were changed in a stepwise manner in scenarios in which the matching sensitivity was 100%, the matching specificity was 100%, and each parameter was equivalent to the corresponding value observed in the matching experiment. The estimation assumed a group that underwent cancer screening in a certain year. The prevalence of new cancer incidence was set at 775.7 of 100,000 person-years based on the average prevalence in Japan. The data size did not affect the estimation, but at the time of calculation, it was set to 1000 people according to the parameters of this experiment.

In the performance evaluation experiment, we attempted to simulate a scenario in which the system is used by the institutions that are geographically distant from one another. Therefore, we used 6 computers installed at Osaka University and Yamaguchi University (4 of which simulated data-holding institutions). In the experiment, we measured CPU use, memory use, and network traffic for 3 data sizes: 2^{10}, 2^{12}, and 2^{14}. We also implemented multiprocess parallelization and measured its speedup ratio.

Ethics Approval

This study was approved by the institutional review board of the Kanagawa Cancer Center (2021 epidemiology-135).

Results

Data Protection

In our experiments, 2 distributed institutes independently held cancer screening and cancer registration data, in which each data set included the terms of birth date, first name, family name, and sex. These terms were used for matching keys. In our system, in addition to the use of probabilistic encryption, all matching keys and information through a network outside the institute are encrypted, and no server deals with raw data were stored in different distributed institutes. In other words, no institute has a decryption key and reveals all information. This implies that our system does not move any privacy information from any institute and thus avoids privacy risk.

Matching Accuracy

The results of matching using PDDI are shown in subsequent sections. From the preliminary experiments, when only 1 matching key is used, the number of false positives for matching increases and the specificity decreases significantly (Table S2 in Multimedia Appendix 3). Figure 2 shows the results of false positives and false negatives in which pseudodata of colorectal cancer and breast cancer were matched using various combinations of information. In the case of colorectal cancer data, the minimum number of false negatives for matching was 27 and the minimum number of false positives for matching was 0. It is desirable that the common data for all 60 items be output. However, up to 33 (60 – 27) cases are output correctly. For breast cancer data, the minimum number of false negatives for matching was 7, and the minimum number of false positives for matching was 0. Similarly, it is desirable that 62 common data items are output but a maximum of 55 (62 – 7) cases were output correctly.

Figure 2. Number of false positives and false negatives. The points are placed according to the number of false positives and false negatives by the setting of each experiment conducted. Part A shows the result of data simulating colorectal cancer and Part B shows the result of data simulating breast cancer.

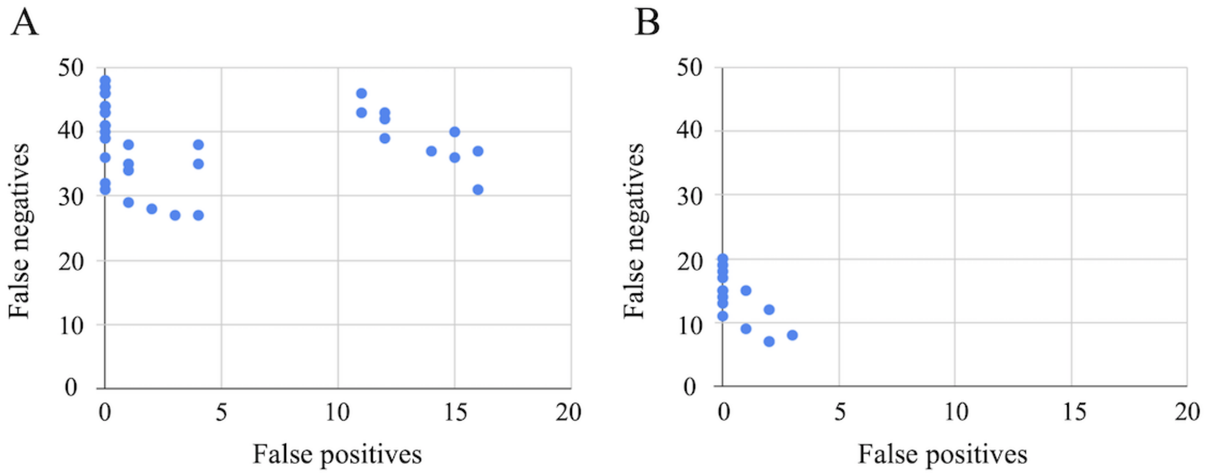


Table 2 presents an excerpt of the matching results. Only combinations with a specificity of ≥99% are shown. In this pseudo–data set, it can be inferred that the combination of matching keys, including the date of birth, is particularly effective. In the colorectal cancer pseudodata, the combination with a specificity of ≥99%, the highest matching sensitivity was the one that used the date of birth and first name (kana) as keys; the matching sensitivity was 55.00%, and the matching specificity was 99.85%. For breast cancer pseudodata, the highest matching sensitivity was obtained when the date of birth and family name (kana or kanji) were used as keys: the matching sensitivity was 88.71%, and the matching specificity was 99.80%. In combination with 100% matching specificity, the matching sensitivity was 48.33% for the data simulating colorectal cancer and 82.26% for the data simulating breast cancer.

Table 2. Matching result between cancer-screening and cancer-registration data (excerpt).

Class[a] and matching key	False positive, n	False negative, n	Sensitivity (%)	Specificity (%)
Colorectal cancer				
Birth date, first name (kana)	3	27	55.00	99.85
Birth date, first name (kana), family name (kana)	0	31	48.33	100
Birth date, sex, first name (kana)	2	28	53.33	99.90
Birth date, sex, family name (kana)	1	29	51.67	99.95
Breast cancer				
Birth date, family name (kana)	2	7	88.71	99.80
Birth date, family name (kanji)	2	7	88.71	99.80
Birth date, first name (kanji)	1	9	85.48	99.90
Birth date, first name (kana), family name (kanji)	0	11	82.26	100

[a] Results of the matching experiment between cancer-screening and cancer registration data for each matching key used. Cases in which all key data shown in the matching-key column successfully corresponded were considered positive matches.

Table 3 shows the effect of matching accuracy on the estimation of sensitivity and specificity of cancer screening based on the model used in this experiment, an assessment of the accuracy of cancer screening. The matching sensitivities were approximately 85%, 50%, and 90%, and the matching specificities were 99.9%, 99.8%, and 99.99%. Assuming that the original values of both screening sensitivity and specificity are both 90% if the matching specificity is set to 100% and the matching sensitivity values are reduced to 90%, 85%, and 50%, the apparent screening specificity values become 89.94% (−0.06%), 89.91% (−0.10%), and 89.69% (−0.34%), respectively. Thus, as the matching sensitivity decreases, the screening specificity is underestimated. If the matching specificity decreases, the screening sensitivity is underestimated. On the basis of the experimental results of the data set simulating breast cancer, when calculated with a matching sensitivity of 88.71% and matching specificity of 99.80%, the apparent value of the screening sensitivity was 72.09% (−19.9%) and that of the screening specificity was 89.93% (−0.08%), and the rate of change in the apparent value of the screening sensitivity was large. However, when using the results of another combination and calculating with a matching sensitivity of 82.26% and matching specificity of 100%, the apparent value of screening sensitivity is 90% (no decrease), and the apparent value of screening specificity is 89.89% (−0.12%). In other words, when the matching specificity is sufficiently large, even if the matching sensitivity is a little low, the change from the original value for both screening sensitivity and screening specificity remains small. As shown in Table 3, this tendency was maintained, even in the estimation assuming the original screening sensitivity of 60%. In addition, regarding the positive predictive value of screening, a decrease in matching sensitivity makes the positive predictive value of screening appear smaller than the original value, and a decrease in matching specificity makes the positive predictive value of screening appear larger than the original value. The effect of matching specificity is also greater for the positive predictive value of screening.

Table 3. Estimation of the impact of matching accuracy on the screening accuracy[a].

Assumption of matching accuracy (%)		Screening sensitivity (%)		Screening specificity (%)		Positive predictive value (%)	
Sensitivity	Specificity	True	Estimate	True	Estimate	True	Estimate
90	100	90	NA[b]	90	89.94	6.6	5.92
85	100	90	NA	90	89.91	6.6	5.59
50	100	90	NA	90	89.69	6.6	3.29
100	99.99	90	88.99	90	NA	6.6	6.58
100	99.90	90	80.93	90	NA	6.6	6.67
100	99.80	90	73.70	90	NA	6.6	6.76
88.71	*99.80*	*90*	*90.00*	*90*	*89.89*	*6.6*	*6.02*
82.26	*100*	*90*	*72.09*	*90*	*89.93*	*6.6*	*5.41*
90	100	60	NA	90	89.96	4.5	4.03
85	100	60	NA	90	89.94	4.5	3.81
50	100	60	NA	90	89.81	4.5	2.24
100	99.99	60	59.37	90	NA	4.5	4.49
100	99.90	60	54.33	90	NA	4.5	4.58
100	99.80	60	49.81	90	NA	4.5	4.67
88.71	*99.80*	*60*	*48.81*	*90*	*89.96*	*4.5*	*4.17*
82.26	*100*	*60*	*60.00*	*90*	*89.68*	*4.5*	*3.18*

[a] The table shows the impact of matching accuracy on cancer-screening accuracy estimates when the true sensitivity of cancer screening is set at 90% and 60%, and the true specificity is set at 90%. The cancer incidence rate is approximately 775.7 person per year, which is the national average in Japan.

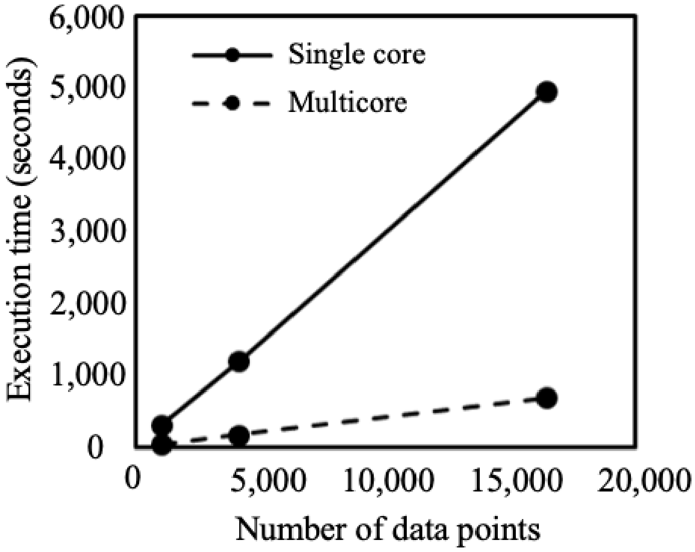
[b] NA: not affected. “NA” represents that no change occurred between the true and estimated values. The italicized values show the estimates obtained using the experimental data.

In principle, when the matching sensitivity is 100%, even if the matching specificity is reduced, both true-negative and false-positive cancer screenings are misidentified as having cancer at the same rate. Therefore, the specificity of cancer screening does not change. Similarly, when the matching specificity is 100%, even if the matching sensitivity decreases, both true-positive and false-negative cancer screening will be misidentified as “no cancer” at the same rate. Therefore, the sensitivity of cancer screening does not change. Therefore, these values are not shown and are depicted as not affected, except when the matching sensitivity and matching specificity obtained from the matching experiment are used.

Performance

The results of the performance evaluation experiment are in subsequent sections. The specifications of the computer used in the experiment are listed in Table S1 in Multimedia Appendix 1. Figure 3 shows the relationship between the amount of data and execution time.

Figure 3. Execution time. The graph shows the relationship between the amount of data and the execution time. The solid line shows the execution time without parallelization, and the dashed line shows the execution time with parallelization.



As shown in Figure 3, the amount of data and the execution time are almost proportional. Furthermore, with 2^{14} (16,384) data points, the nonparallelized execution time was 82 minutes and 26 seconds, whereas with parallelization, the execution time was 11 minutes 38 seconds; hence, a 7.1-fold speedup was observed with parallelization. Figure 4 shows the changes in CPU use of the PDDI server and data-holding institutions when the process is executed on 2^{14} data points without parallelization. As can be observed in this graph, 80.67% of the execution time is processed by the PDDI server, and the calculation time of the data-holding institutions is only 19.33%.

Figure 4. Changes in central processing unit (CPU) usage. The graphs show the changes in CPU usage of the privacy-preserving distributed data integration (PDDI) server and the data-holding institutions when the process is executed on 214 datapoints without parallelization. Part A represents the results of the PDDI server, and part B represents the results of the data-holding institution.

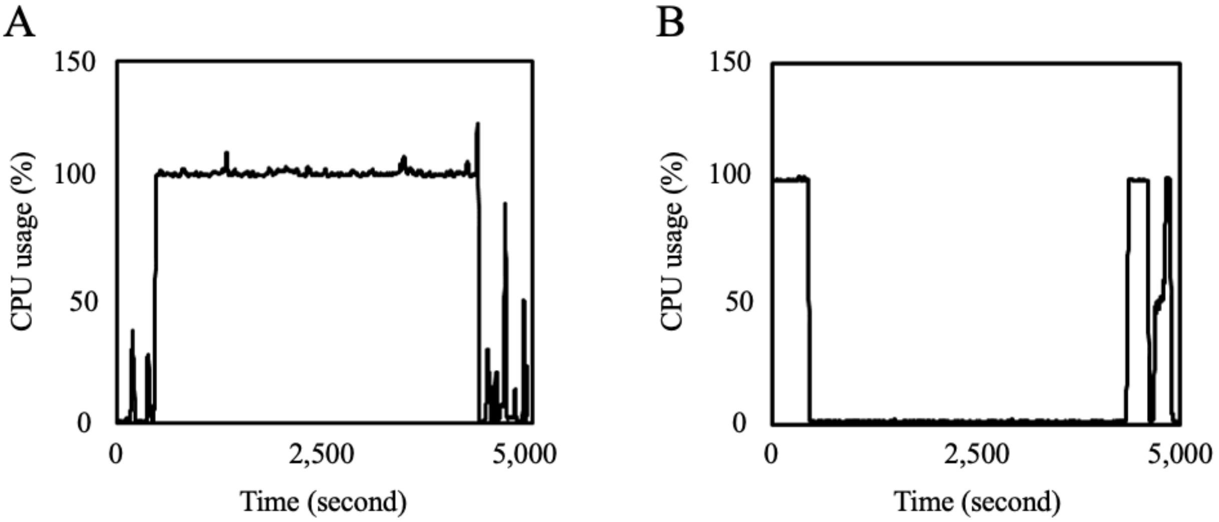
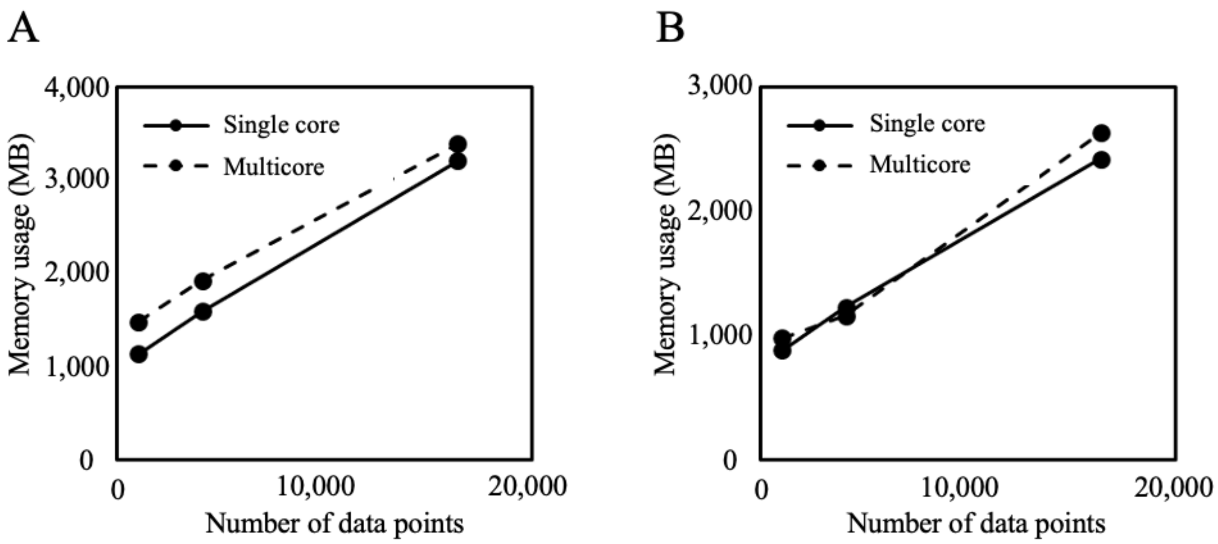


Figure 5 shows the relationship between the amount of data and memory use of the PDDI server and data-holding institutions. Memory use increases linearly with the amount of data. However, even during parallelization for 2^{14} data, which uses a large amount of memory, the PDDI server required no more than 3.4 GB of memory, and the data-holding institutions required no more than 2.7 GB of memory.

Figure 5. Memory usage. The graphs show the relationship between the amount of data and the memory usage of the privacy-preserving distributed data integration (PDDI) server and the data-holding institutions. Part A represents the results of the PDDI server, and part B represents the results of the data-holding institution.



Discussion

Evaluation of Matching Experiment

In this study, we conducted a matching experiment using the accuracy assessment of cancer screening as a model by matching the cancer-screening and cancer registration data.

In the experiment, any matching information is transformed into Bloom filters, encrypted within each institution, and then sent to the PDDI server. Probabilistic encryption was used in this study. This implies that the same matching key is compressed and randomly encrypted to different ciphertext, for example, each birth date of patients A and B in cancer registration data set is 19970911, but that compressed and randomly encrypted are not equal to each other. Unlike simple matching using a hash value [13], our scheme is secure against dictionary attacks because the same value is encrypted into different values owing to the probabilistic encryption.

The matching keys used for multiple combinations, which were particularly excellent with few false positives and false negatives, were all registered in most databases in Japan. It is highly likely that these keys can be applied to existing databases. The matching sensitivity remained in the 50% range for simulated colorectal cancer data containing 85% matching-key errors, but in the case of simulated breast cancer data, which contained 59% matching-key errors, the matching sensitivity value was approximately 85%. This experiment was conducted in a manner that intentionally created a data set that was difficult to match owing to a high prevalence of errors and a large amount of data containing errors in multiple matching keys. The errors contained in the 2 data sets differ as shown in Table 1, and these results cannot be simply compared, but, in general, the fewer the number of errors in the matching keys, the better the matching accuracy. Although cultural backgrounds and times vary, previous studies have shown that the number of errors and omissions in disease registries, medical, and government databases is <15% for matching-key data such as name, zip code, and date of birth [33-35]. On the basis of the opinions of staff with abundant experience in registry management, we predicted that up to approximately 10% of the actual data used for cancer-screening accuracy assessment in Japan includes an error in the matching key. In principle, the false-negative rate cannot be greater than the percentage of data with errors contained in the data set; therefore, it is estimated that a matching sensitivity of ≥90% can be obtained in verification experiments using actual data. The error distributions of the 2 data sets in this experiment were the same, and the prevalence was set at 10%. In the colorectal cancer data, the matching sensitivity was 94.70% when the date of birth and first name (kana) were used as the matching key. In breast cancer data, the matching sensitivity was 98.09% when the date of birth and family name (kana or kanji) were used as the matching key. Regarding the specificity of matching, the combination of keys shown in Table 2 maintained a high specificity of ≥99% in this estimation.

In practical use, the influence on the outcome and evaluation index to be obtained by performing matching is more important than the numerical value of the matching accuracy. As shown in Table 3, when assessing test accuracy for infrequent events, such as cancer, changes in matching specificity values have a significant effect on the apparent value of test accuracy. In our model, a slight decrease in matching sensitivity had a relatively small effect on screening sensitivity and screening specificity. In other words, it is highly important to keep the matching specificity as high as possible to prevent underestimation of the screening sensitivity and screening specificity. The estimation shows that a combination of matching keys with 100% matching specificity has a small effect on the sensitivity and specificity of cancer screening, even if the matching sensitivity is low. Assuming that the original screening sensitivity and screening specificity are 90%, even when the matching specificity is not 100% if the matching specificity is ≥99.97%, the screening sensitivity maintains within 5% even if the matching sensitivity is 85%. Therefore, when considering the accurate calculation of sensitivity estimates for cancer screening, it is desirable to select a matching-key or matching algorithm that can improve matching sensitivity as much as possible without reducing matching specificity. Matching specificity has a greater effect than matching sensitivity on the positive predictive value of screening. However, it is more susceptible to matching sensitivity than screening sensitivity or screening specificity. Therefore, when focusing on the positive predictive value of screening as the index, it is necessary to select the matching key in consideration to not only the matching specificity but also the decrease in matching sensitivity.

Matching specificity in this experiment is defined as the value obtained by dividing the number of people who are determined not to have cancer as a result of matching by the number of people who do not have cancer among the data included in the cancer-screening data set. Therefore, the specificity of the match is affected by the ratio of the data size of the cancer registration data set to the cancer-screening data set and the percentage of true patients with cancer included in the cancer-screening data set. The cancer-screening and cancer registration data sets used in this experiment were approximately 1000 to 2000 and approximately 17,000 to 30,000, respectively. In Japan, where the cancer-screening rate is low, this is roughly equivalent to the number of cancer screenings in small municipalities and the number of cancers in large prefectures; cancer-screening data are managed for each municipality that is the implementing body, and cancer registration data are managed by each prefecture. Epidemiological studies may have to deal with even larger cancer-screening data. In this case, the difference in data size from the cancer registration data set is smaller than that in this experiment. Therefore, matching specificity is expected to be higher. As the errors of the data set in this experiment do not necessarily reflect the actual prevalence, the sensitivity and specificity in this experiment are just reference values. Even so, it is expected that the PDDI system can be used for the assessment of cancer-screening accuracy using matching with cancer registration data by appropriately adjusting the matching conditions.

Performance evaluation experiments verified that the execution time of the PDDI system was almost proportional to the amount of data, and the execution time in parallel execution was 43 seconds per 1000 data samples. With the pseudodatabase used, the execution was completed in approximately 21 minutes, which is sufficient performance for epidemiological studies. The effect of the performance of the computer installed in the data-holding organization on the execution time is relatively small, approximately 20% of the total, and the memory use is <1 GB. Therefore, it was proven that the processing speed was acceptable even with the performance of a normal laptop PC. The maximum network traffic of the PDDI system in this experiment was 858 Mbps. Even so, the execution time consumed by communication is small, and if the communication speed of the data-holding organization is ≥10 Mbps, we do not believe that there will be any problems using this system.

Challenges for Next Experiments Using Practical Data

On the basis of this study, we plan to conduct a verification experiment using actual cancer-screening and cancer registration data. In this experiment, the number of errors in the actual data were unknown. Therefore, the experiment was conducted using a data set with a large number of errors. In the next matching experiment using actual data, we plan to determine the degree of matching accuracy that can be obtained in comparison to a method that partly uses matching based on human judgment. On the basis of this, it is possible to realistically estimate the extent to which matching can cause errors in examination accuracy. Therefore, it is possible to perform higher quality evaluations for practical use. Regarding performance evaluation, as shown in the results of this experiment, the calculation time and memory consumption of the terminal depend on the amount of data. The main purpose of this experiment was to evaluate the feasibility, and the data set used was with a smaller number of items than those contained in the actual data. Therefore, in the next stage, we will confirm the performance using data on the scale of municipalities and prefectures that may actually be used. On the basis of these results, it is necessary to perform a trial calculation to determine the size of the data set that can be matched.

Implementation for Practical Epidemiological Studies

Through this experiment and estimation, we demonstrated that the use of matching using the PDDI system for cancer-screening accuracy assessment deserves consideration. This system is expected to be applied to other types of epidemiological research because it assists in data matching between databases managed by different institutions. We considered the applicability based on matching sensitivity and specificity using cohort studies and case-control studies, which are typical epidemiological studies, as examples.

Assuming that a cohort study examining the association between a factor and cancer incidence will determine the risk ratio of cancer incidence with people who have the factor compared with those who do not have, each person’s data in the cohort are matched with cancer registration data to record cancer incidence. The estimation of this setting is presented in Table S3 in Multimedia Appendix 4. The risk ratio does not change from the true value only by the decrease in matching sensitivity. If the matching specificity is reduced, the risk ratio is underestimated. However, it can be seen from the estimation that the decrease in the risk ratio is approximately 10% in the matching sensitivity and matching specificity equivalent to this matching experiment, even when the prevalence of the factor is 75%. Next, let us assume a case-control study using a data set that links the factors to be examined with data on the presence or absence of a disease by matching. Table S4 in Multimedia Appendix 4 shows a common disease with a high prevalence, here a trial calculation for diabetes, and Table S5 in Multimedia Appendix 4 shows a trial calculation for ulcerative colitis as an example of a disease with a low prevalence. Poor matching accuracy causes systematic errors in factor exposure in populations and control populations, which tends to underestimate odds ratio estimates. Occasionally, this has a greater effect on the odds ratios in diseases with low prevalence. Therefore, when assuming the use of the PDDI system in cohort and case-control studies, care must be taken in selecting the target disease and underestimating the odds ratio. However, if appropriate calculations are made, it appears that a large variety of applications can be fully examined.

The advantage of the PDDI system is that it can provide data to users in an already-matched state, even among ≥3 databases. Currently, in research that integrates data managed by different institutions without a unique identification key, a step-by-step process is necessary, such as collecting data from all target institutions and then performing a match or narrowing down the target audience and repeating the match. However, in the PDDI system, although the data are distributed and stored in different institutions, it is possible to retrieve matched data that meet these conditions. As in other methods [39], it does not assume prior linkage. Therefore, the PDDI system is particularly useful when data obtained from the databases of ≥3 institutions are combined and analyzed. Owing to this characteristic, this system enables the safe and efficient integration of data even in an environment such as Japan, that is, an environment where cancer-screening data are distributed and stored in many municipalities and, therefore, requires multiple movements of private information.

Limitations

This study has several limitations. This study was conducted as a preliminary step in the experiments using real-life data. The data set used in this experiment is a pseudo–data set created using software that is open to the public and does not reflect the amount or ratio of errors mixed in the actual data, nor does it cover all types of errors contained in real-world data. As the types and number of errors contained in actual data depend on the input style of each database and the ability of the input person, subsequent verification experiments using actual data are required. In this study, we dealt only with matching under the condition that all the selected matching keys matched and did not use complicated algorithms for partial matches. We did not examine the extent to which the matching sensitivity and matching specificity shown in this study can be improved by further improvements in matching methods. The experiment used a local database in Japan as the environment, and we noted that the error format is also influenced by language, culture, and institution. Therefore, it is unlikely that this result can be applied directly to other countries and regions.

Conclusions

As a first step toward implementing PDDI in epidemiological studies, we evaluated its feasibility in a model of cancer-screening accuracy assessment in terms of safety, matching accuracy, and performance through a matching experiment using dummy data. This system makes it possible to collate only the information related to the shared data without disclosing the data distributed and managed by multiple institutions and without using a third party. In the matching experiment and the estimation of the effect on the cancer-screening accuracy index using the matching sensitivity and matching specificity obtained by the experiment, it was shown that screening sensitivity and screening specificity can be assessed with minimal errors by keeping the matching specificity high. Because of its characteristics, this system reduces the labor and costs required for personal information management and collation work for both researchers and data providers in many epidemiological studies and is expected to further improve the efficiency and speed of research activities. In future, we will carry out further verification for practical use by using existing data and comparing it with existing methods.

Acknowledgments

This research was supported in part by the Ministry of Education, Culture, Sports, Science and Technology’s 2018 “Society 5.0 Realization Research Center Support Project” and the Japan Society for the Promotion of Science’s Grant-in-Aid for Scientific Research (JP21H034438) and supported by Editage for English language editing and translation. AM, YT, and KN are the developers of the privacy-preserving distributed data integration system discussed in this study. Osaka University has patent rights related to the technology.

Authors' Contributions

AM, YT, and KN were responsible for the development of the privacy-preserving distributed data integration (PDDI) system and environment. AM, YT, KN, and HN designed the study. KW and HN provided the simulated data used in the experiments, and YT and KN conducted buttress experiments using these data. The results were analyzed and interpreted by all authors. In writing the manuscript, YT was responsible for the PDDI system and matching experiments; KN for performance evaluation; AM for the PDDI system and engineering considerations; and KW for the epidemiological background, simulations, and epidemiological considerations. SN and YW provided a critical review and advice on the manuscript from epidemiological and engineering perspectives, respectively. AM was responsible for the overall supervision and oversight of the study in the engineering field, and HN, in the epidemiological field. AM and KW contributed equally to the preparation of this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Privacy-preserving distributed data integration (PDDI) implementation environment, environment construction and usability.

[DOCX File , 23 KB - medinform_v10i12e38922_app1.docx]

Multimedia Appendix 2
Cultural background of practical data-matching failures and examples of the errors specific to Japanese in the dataset of the experiment.
[DOCX File , 185 KB - medinform_v10i12e38922_app2.docx]

Multimedia Appendix 3
Matching-key combinations and the matching results that were not described in the text.
[DOCX File , 21 KB - medinform_v10i12e38922_app3.docx]

Multimedia Appendix 4
Estimating the impact of matching accuracy on outcome evaluation in epidemiological studies.
[DOCX File , 33 KB - medinform_v10i12e38922_app4.docx]

References

1. Matsuda T, Sobue T. Recent trends in population-based cancer registries in Japan: the Act on Promotion of Cancer Registries and drastic changes in the historical registry. Int J Clin Oncol 2015 Feb;20(1):11-20. [doi: 10.1007/s10147-014-0765-4] [Medline: 25351534]
2. Anazawa T, Miyata H, Gotoh M. Cancer registries in Japan: national clinical database and site-specific cancer registries. Int J Clin Oncol 2015 Feb;20(1):5-10. [doi: 10.1007/s10147-014-0757-4] [Medline: 25376769]
3. Rare Disease Data Registry of Japan (in Japanese). Japan Agency for Medical Research and Development. URL: https://www.raddarj.org [accessed 2022-03-03]
4. Tsugane S, Sawada N. The JPHC study: design and some findings on the typical Japanese diet. Jpn J Clin Oncol 2014 Sep 07;44(9):777-782. [doi: 10.1093/jjco/hyu096] [Medline: 25104790]
5. Takeuchi K, Naito M, Kawai S, Tsukamoto M, Kadomatsu Y, Kubo Y, et al. Study profile of the Japan multi-institutional collaborative cohort (J-MICC) study. J Epidemiol 2021 Dec 05;31(12):660-668 [FREE Full text] [doi: 10.2188/jea.JE20200147] [Medline: 32963210]
6. Emery J, Boyle D. Data linkage. Aust Fam Physician 2017;46(8):615-619. [Medline: 28787562]
7. Pratt NL, Mack CD, Meyer AM, Davis KJ, Hammill BG, Hampp C, et al. Data linkage in pharmacoepidemiology: a call for rigorous evaluation and reporting. Pharmacoepidemiol Drug Saf 2020 Jan;29(1):9-17. [doi: 10.1002/pds.4924] [Medline: 31736248]
8. Hagger-Johnson G. Opportunities for longitudinal data linkage in Scotland. Scott Med J 2016 Aug;61(3):136-145. [doi: 10.1177/0036933015575214] [Medline: 25886907]
9. An overview of record linkage methods. In: Linking Data for Health Services Research: A Framework and Instructional Guide. Rockville, MD: Agency for Healthcare Research and Quality (US); 2014.
10. Ludvigsson JF, Almqvist C, Bonamy AE, Ljung R, Michaëlsson K, Neovius M, et al. Registers of the Swedish total population and their use in medical research. Eur J Epidemiol 2016 Feb;31(2):125-136. [doi: 10.1007/s10654-016-0117-y] [Medline: 26769609]
11. Laugesen K, Ludvigsson JF, Schmidt M, Gissler M, Valdimarsdottir UA, Lunde A, et al. Nordic health registry-based research: a review of health care systems and key registries. Clin Epidemiol 2021;13:533-554 [FREE Full text] [doi: 10.2147/CLEP.S314959] [Medline: 34321928]
12. Christen P. Data Matching Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Berlin, Heidelberg: Springer; 2012.
13. Kho AN, Cashy JP, Jackson KL, Pah AR, Goel S, Boehnke J, et al. Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. J Am Med Inform Assoc 2015 Sep;22(5):1072-1080 [FREE Full text] [doi: 10.1093/jamia/ocv038] [Medline: 26104741]
14. Godlove T, Ball AW. Patient matching within a health information exchange. Perspect Health Inf Manag 2015;12(Spring):1g [FREE Full text] [Medline: 26755901]
15. Kissner L, Song D. Privacy-preserving set operations. In: Proceedings of the 25th annual international conference on Advances in Cryptology. 2005 Presented at: CRYPTO'05: Proceedings of the 25th annual international conference on Advances in Cryptology; Aug 14 - 18, 2005; Santa Barbara California. [doi: 10.21236/ada457144]
16. Many D, Burkhart M, Dimitropoulos X. Fast private set operations with SEPIA. TIK Report. 2012 Mar. URL: https://www.research-collection.ethz.ch/handle/20.500.11850/58312 [accessed 2022-04-04]
17. Ion M, Kreuter B, Nergiz A, Patel S, Raykova M, Saxena S, et al. On deploying secure computing: private intersection-sum-with-cardinality. In: Proceedings of the 2020 IEEE European Symposium on Security and Privacy (EuroS&P). 2020 Presented at: 2020 IEEE European Symposium on Security and Privacy (EuroS&P); Sep 07-11, 2020; Genoa, Italy. [doi: 10.1109/eurosp48549.2020.00031]

18. Miyaji A, Nakasho K, Nishida S. Privacy-preserving integration of medical data : a practical multiparty private set intersection. J Med Syst 2017 Mar 16;41(3):37 [FREE Full text] [doi: 10.1007/s10916-016-0657-4] [Medline: 28093660]
19. Miyaji A, Mimoto T. Security Infrastructure Technology for Integrated Utilization of Big Data Applied to the Living Safety and Medical Fields. Cham: Springer; 2020.
20. Winkler W. Matching and record linkage. WIREs Comp Stat 2014 Jul 02;6(5):313-325. [doi: 10.1002/wics.1317]
21. Sorensen HT, Sabroe S, Olsen J. A framework for evaluation of secondary data sources for epidemiological research. Int J Epidemiol 1996 Apr;25(2):435-442. [doi: 10.1093/ije/25.2.435] [Medline: 9119571]
22. Tromp M, Ravelli AC, Bonsel GJ, Hasman A, Reitsma JB. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. J Clin Epidemiol 2011 May;64(5):565-572. [doi: 10.1016/j.jclinepi.2010.05.008] [Medline: 20952162]
23. Sayers A, Ben-Shlomo Y, Blom AW, Steele F. Probabilistic record linkage. Int J Epidemiol 2016 Jun;45(3):954-964 [FREE Full text] [doi: 10.1093/ije/dyv322] [Medline: 26686842]
24. Jaro MA. Probabilistic linkage of large public health data files. Stat Med 1995;14(5-7):491-498. [doi: 10.1002/sim.4780140510] [Medline: 7792443]
25. Page on promoting cancer screening based on scientific evidence (in Japanese). National Cancer Center Institute for Cancer Control. URL: http://canscreen.ncc.go.jp [accessed 2022-03-03]
26. Screening and earlier diagnosis. NHS England. URL: https://www.england.nhs.uk/cancer/early-diagnosis/screening-and-earlier-diagnosis/ [accessed 2022-03-03]
27. American cancer society guidelines for the early detection of cancer. American Cancer Society. URL: https://www.cancer.org/healthy/find-cancer-early/american-cancer-society-guidelines-for-the-early-detection-of-cancer.html [accessed 2022-03-03]
28. Tanaka R, Matsukata M. Report on the model project for the accurate management of cancer screening by utilizing cancer registry data in FY2017 - Aomori Prefecture Commissioned Project (in Japanese). Aomori prefecture 2018 Mar [FREE Full text]
29. 2017 by utilizing cancer registry data Accuracy control project report for cancer screening. Ministry of Health, Labor and Welfare research group. 2018. URL: https://www.pref.wakayama.lg.jp/prefg/041200/h_sippei/gannet/04/05_d/fil/houkokusyo.pdf [accessed 2022-12-19]
30. Pseudo personal information data generation service. hogehoge.tk. URL: http://hogehoge.tk/personal/ [accessed 2021-05-29]
31. Personal information. Kazina. URL: http://kazina.com/dummy/ [accessed 2021-05-29]
32. Test Data Generator (in Japanese). Yamagata. URL: http://yamagata.int21h.jp/tool/testdata/ [accessed 2021-05-29]
33. Muse AG, Mikl J, Smith PF. Evaluating the quality of anonymous record linkage using deterministic procedures with the New York State AIDS registry and a hospital discharge file. Stat Med 1995;14(5-7):499-509. [doi: 10.1002/sim.4780140511] [Medline: 7792444]
34. Howe GR. Use of computerized record linkage in cohort studies. Epidemiol Rev 1998;20(1):112-121. [doi: 10.1093/oxfordjournals.epirev.a017966] [Medline: 9762514]
35. Setoguchi S, Zhu Y, Jalbert JJ, Williams LA, Chen C. Validity of deterministic record linkage using multiple indirect personal identifiers. Circ Cardiovasc Qual Outcomes 2014 May;7(3):475-480. [doi: 10.1161/circoutcomes.113.000294]
36. Ladabaum U, Dominitz JA, Kahi C, Schoen RE. Strategies for colorectal cancer screening. Gastroenterology 2020 Jan;158(2):418-432. [doi: 10.1053/j.gastro.2019.06.043] [Medline: 31394083]
37. Koliopoulos G, Nyaga VN, Santesso N, Bryant A, Martin-Hirsch PP, Mustafa RA, et al. Cytology versus HPV testing for cervical cancer screening in the general population. Cochrane Database Syst Rev 2017 Aug 10;8(8):CD008587 [FREE Full text] [doi: 10.1002/14651858.CD008587.pub2] [Medline: 28796882]
38. Hamashima C, Ohta K, Kasahara Y, Katayama T, Nakayama T, Honjo S, et al. A meta-analysis of mammographic screening with and without clinical breast examination. Cancer Sci 2015 Jul;106(7):812-818 [FREE Full text] [doi: 10.1111/cas.12693] [Medline: 25959787]
39. Kawamoto Y, Shirai T, Kamio K, Tanaka Y, Sakumoto K. Information processing apparatus, information processing method, program, and information processing system. Google Patents. 2014. URL: https://patents.google.com/patent/US20140012862A1/en?oq=US20140012862A1 [accessed 2022-04-06]

Abbreviations

CPU: central processing unit
PDDI: privacy-preserving distributed data integration

Edited by C Lovis; submitted 17.05.22; peer-reviewed by C Sun, SY Shin; comments to author 07.10.22; revised version received 04.11.22; accepted 29.11.22; published 30.12.22.

Please cite as:
Miyaji A, Watanabe K, Takano Y, Nakasho K, Nakamura S, Wang Y, Narimatsu H
A Privacy-Preserving Distributed Medical Data Integration Security System for Accuracy Assessment of Cancer Screening: Development Study of Novel Data Integration System
JMIR Med Inform 2022;10(12):e38922
URL: https://medinform.jmir.org/2022/12/e38922
doi:10.2196/38922
PMID:36583931

Original Paper

Implementation of Machine Learning Pipelines for Clinical Practice: Development and Validation Study

Lara J Kanbar[1], PhD; Benjamin Wissel[2], PhD; Yizhao Ni[2,3], PhD; Nathan Pajor[1,2,3], MD; Tracy Glauser[3,4], MD; John Pestian[2,3], PhD; Judith W Dexheimer[2,3,5], PhD

[1]Division of Pulmonary Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States
[2]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States
[3]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, United States
[4]Division of Neurology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States
[5]Division of Emergency Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States

Corresponding Author:
Judith W Dexheimer, PhD
Division of Biomedical Informatics
Cincinnati Children's Hospital Medical Center
3333 Burnet Avenue
Cincinnati, OH, 45229
United States
Phone: 1 5138032962
Email: judith.dexheimer@cchmc.org

Abstract

Background: Artificial intelligence (AI) technologies, such as machine learning and natural language processing, have the potential to provide new insights into complex health data. Although powerful, these algorithms rarely move from experimental studies to direct clinical care implementation.

Objective: We aimed to describe the key components for successful development and integration of two AI technology–based research pipelines for clinical practice.

Methods: We summarized the approach, results, and key learnings from the implementation of the following two systems implemented at a large, tertiary care children’s hospital: (1) epilepsy surgical candidate identification (or epilepsy ID) in an ambulatory neurology clinic; and (2) an automated clinical trial eligibility screener (ACTES) for the real-time identification of patients for research studies in a pediatric emergency department.

Results: The epilepsy ID system performed as well as board-certified neurologists in identifying surgical candidates (with a sensitivity of 71% and positive predictive value of 77%). The ACTES system decreased coordinator screening time by 12.9%. The success of each project was largely dependent upon the collaboration between machine learning experts, research and operational information technology professionals, longitudinal support from clinical providers, and institutional leadership.

Conclusions: These projects showcase novel interactions between machine learning recommendations and providers during clinical care. Our deployment provides seamless, real-time integration of AI technology to provide decision support and improve patient care.

(JMIR Med Inform 2022;10(12):e37833) doi:10.2196/37833

KEYWORDS

electronic health record; natural language processing; epilepsy; clinical decision support; machine learning; emergency medicine; artificial intelligence

Introduction

With the rampant growth in health data, artificial intelligence (AI) technologies, such as machine learning and natural language processing (NLP), provide a powerful means to extract meaningful associations from big data sets [1]. Applications of machine learning are far-reaching and have included patient identification, computer vision, speech recognition, web searches, and phenotype discovery [2-9].

The electronic health record (EHR) captures data relating to clinical encounters, but as much as 30%-50% of these data are available only in free text [10]. As such, one particularly valuable means to understand health care data involves NLP. NLP is a technique of incorporating free-text analysis and statistical methods into computerized algorithms to derive linguistic features (eg, physicians' diagnosis) from human language input [11]. Clinical care and research can benefit from using this unstructured text information [12,13]. NLP has been used for surveillance, adverse event detection [14-18], medication identification [19], and extraction of data from radiology reports [20-22]. NLP has also successfully been applied to evaluate clinical notes and provide recommendations as part of clinical decision support (CDS) tools [23].

These CDS tools can change user behavior; however, to ensure successful implementation, user involvement in CDS design is critical [24-30]. CDS tools using AI and NLP technologies remain less implementable directly into real-time clinical care with long-term success [31-34]. The reason integration of these AI pipelines within a clinical health system is challenging is that it requires coordination with the following: (1) key stakeholders and expected end users of the CDS tools; (2) biomedical informatics professionals who design the AI; (3) research information technology (IT) professionals who design the CDS tools with stakeholders in mind; and (4) operational IT professionals who are responsible for maintenance, uptime, and EHR integration [35].

In this work, we report the main modifications implemented to improve the development and real-time integration of two AI technology–based pipelines using NLP in a tertiary pediatric health care institution. These modifications contributed to the successful deployment and ongoing utilization of these pipelines.

Methods

Objective

The objective of our case studies was to create functional AI technology–based CDS tools effective in research settings and integrate them into clinical workflow without sacrificing care quality, speed of clinical care delivery, and labor requirement.

Setting and Participants

Cincinnati Children's Hospital Medical Center is a large tertiary care center with more than 1.2 million patient encounters annually. It has a large epilepsy clinic (over 6,400 patients and 12,000 epilepsy visits per year) and a high volume of epilepsy surgery cases (50 per year). The division of pediatric emergency medicine oversees 5 urgent cares and 2 emergency departments (EDs) with an annual census of 170,000 visits. The ED employs 8 full-time clinical research coordinators (CRCs) who enroll patients in research studies during clinical visits.

Case 1: Automated Epilepsy Interventions

Background

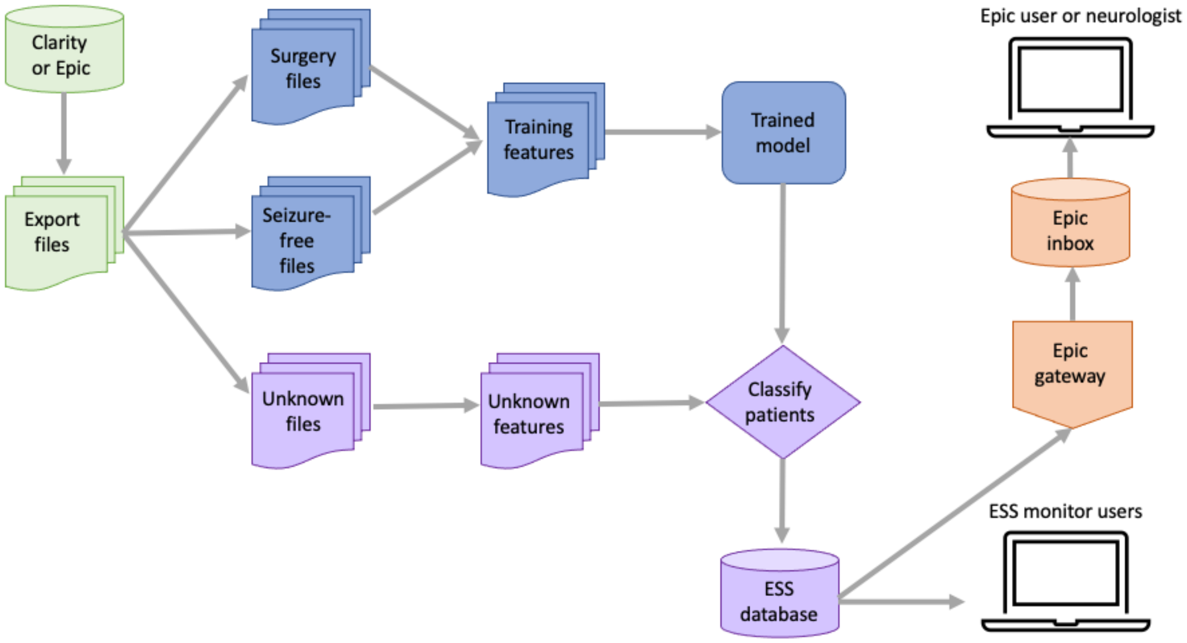
The first case study aimed to facilitate early surgical intervention in patients with intractable epilepsy, as it has been shown to improve cognitive outcomes, mental health, and quality of life [36], as well as increase quality-adjusted life years [37] in a relatively safe procedure for the patient [38]. National guidelines state that patients who continue to have debilitating seizures after 2 or more adequate trials of antiepileptic medications should be considered for a presurgical evaluation referral [39]. From the time of first seizure, on average, patients receive surgery after having epilepsy for 7 years in pediatrics and 20 years in adults [40,41]. Only 0.5%-1.5% of patients received surgery within 2 years of fulfilling clinical criteria for surgical candidacy [42]. Indeed, improving the use of surgery has proven to be difficult [42] because this highly specialized but critical clinical knowledge is not ubiquitously available in clinical care.

Approach

A corpus of notes from patients with a diagnosis of epilepsy who were seizure free or had a history of resective epilepsy surgery was used to devise NLP features. The NLP generated surgical candidacy scores for each patient, with higher scores indicating a higher likelihood of surgical candidacy and lower scores indicating a higher likelihood of seizure freedom. Next, naïve Bayes, support vector machine, and random forest models were developed using retrospective data as described in previous work [43]. Figure 1 describes the system pipeline from input data to the output recommendation.

To ensure the recommendations from the NLP system would be accepted into practice, we validated the algorithm's classifications by comparing them head-to-head against manual labels from epileptologists [2]. Prior to implementation into clinical care, we prospectively evaluated the system for 1 year to test the accuracy in a clinical setting [44].

Figure 1. Epilepsy surgical pipeline architecture. From left to right: a series of Oracle PL/SQL queries extract epilepsy patient data and export them in CSV format to bare meta installation servers. The data are divided into the following 3 groups: patients with surgery, seizure-free patients, and patients with unknown outcome. The feature extraction module (ie, 'training features') analyzes the free-text notes and exports machine-readable feature vectors in SVM light format. Surgery and seizure-free patient features are sent to the classifier training module to train the support vector machine model. Unknown patient features are fed into the final trained classifier, which outputs a surgery candidacy score for each patient. All patients with unknown outcome and their scores are then loaded into the Epilepsy Surgery Software (ESS) database. The highest scoring patients are sent to an Epic web service that generates the in-basket message alerts. All patients and their notes can be viewed and searched in the ESS web application. This entire process is run on a weekly basis, to continually incorporate new electronic health record data into the algorithm training.



Case 2: Automated Clinical Trial Eligibility Screener (ACTES)

Background

The second case study aimed to identify participants who may meet eligibility criteria for clinical trial recruitment in the ED. In current practice, CRCs and physicians at the site of the hospitals do trial eligibility screening manually [45]. For patients presenting during clinical visits, screening would ideally take place early enough in the visit such that eligible candidates could be approached for enrollment without prolonging their length of stay. However, given the large volume of data documented in EHRs, it is labor-intensive for the staff to screen relevant information, particularly within the time frame of a single visit. As such, automatically screening and identifying eligible patients for a trial based on EHR information promises great benefits for clinical research.

Approach

To facilitate participant identification, we developed a machine learning NLP-based system—ACTES [23,46]—which analyzed structured data and unstructured narrative notes automatically to determine patients' suitability for clinical trial enrollment. For development, we evaluated historical trial-patient enrollment decisions in a pediatric ED and extracted EHR data including clinical notes that were commonly reviewed by CRCs. We then customized the machine learning and NLP algorithms based on the trial-patient data. The ACTES was integrated into the institutional workflow to support real-time patient screening in our recent work [44]; details of system development have been previously reported [46].

Implementation Strategy

We hypothesized that successful implementation of the AI solutions relied on 5 key steps, as follows:

1. Integration of industry standard software pertinent to the implementation site. Specifically, the systems needed to be adapted to use industry standard software libraries.
2. Automation of the process to access the EHR data. The systems need to be linked to the EHR to extract the input data without manual intervention.
3. Encouragement of user feedback to inform the final design of the AI solution.
4. Integration of the AI solutions into typical clinical workflow.
5. Performance evaluations and regular maintenance to continue to evaluate the utility of the AI solution.

After building the AI technology, we implemented the AI solutions using these 5 strategies to facilitate successful deployment of the tools.

Results

After creation and validation of the algorithms in a research setting, we implemented these 2 AI solutions as NLP pipelines. Both pipelines follow a step-by-step process that extracts data from the EHR, processes it, and provides a recommendation in the form of automated alerts that could be sent from the research systems to the EHR (Epic Systems) in real time. To do this, the research systems had to be modified to integrate into clinical workflow, as described in this section.

Industry Standard Software

After reviewing over 20 different libraries for managing NLP pipelines, it was decided that the Java NLP library LingPipe [47] would be used for feature extraction and preprocessing, and the LIBSVM Python implementation from scikit-learn [48] would be used for the classifier [49]. The NLP component in ACTES was built upon the clinical Text Analysis and Knowledge Extraction System [50], and the machine learning component was coded in Java (Oracle Corporation).

Automation of EHR Data Access

For the epilepsy intervention AI, Oracle PL/SQL queries from the EHR relational database were used to extract patient data. For ACTES, RESTful and SOAP web services were developed to extract EHR data, such as demographics, medication orders, and clinical notes in real time, which were stored in an Oracle SQL database. An interactive web-based dashboard was developed to visualize the recommendations and receive feedback from CRCs.

User Feedback Informed the Final Design

AI solutions were designed and integrated with feedback from end users. The epilepsy and ACTES corpora were created by manual annotation of patient notes by providers. Throughout the algorithm design and implementation process, providers were included in the build and ultimate integration. First, the biomedical informatics team shadowed providers for workflow observation. Second, the biomedical informatics team attended clinical meetings that included faculty, staff, and clinical research coordinators for a minimum of 10 hours to get feedback and ensure the design was appropriate. Third, mock-up designs were shared at a minimum of 3 meetings to discuss the process of using and interacting with the AI solution in the form of a CDS tool. In cases where the CDS tool could provide an alert, the providers were consulted on their preferred alert method (eg, email or text message alerts). In both AI technologies, the providers were able to directly interact with the machine learning recommendations as follows:

- For epilepsy surgical intervention, these results are displayed in clinical care to suggest surgical consults, and the subsequent actions resulting from the recommendations are fed back into the application to improve performance.
- For ACTES, the clinical research coordinators' entry of eligibility is used to help train and improve the classifier. Additionally, ACTES was assessed and improved for usability and satisfaction by providers and was found to be easy to use and learn.

Integration Into Clinical Workflow

Both AI technologies were integrated into clinical workflow to support clinical practice. For patients with intractable epilepsy and an upcoming visit, surgical eligibility is evaluated in advance. For patients who are classified as potential surgical candidates, EHR in-basket messages are sent to the provider they are scheduled to see via web services.

We integrated the ACTES into the CRCs' workflow to support real-time patient screening [51]. The system ran continuously on a secured, The Health Insurance Portability and Accountability Act (HIPAA)–compliant server to extract structured and unstructured EHR data for current ED patients. For each clinical trial, the ranked list of patients recommended by the system, along with their demographics and clinical information, were displayed on the dashboard available to the CRCs. The information was refreshed at 10-minute increments to accommodate real-time updates. Given the recommended patients as potential participants for a clinical trial, the CRCs performed additional EHR screening to confirm the candidates' eligibility. When an eligible candidate was identified, the CRCs approached him or her for enrollment as per standard clinical workflow.

Performance Evaluation

The epilepsy AI technology went live on April 12, 2016, as part of the EHR release cycle and runs weekly. On Sundays, the system trains on notes from patients who have been seizure free for 1 year or previously underwent resective epilepsy surgery. This trained classifier evaluates all other 'unknown' patients with epilepsy who have had at least one seizure within the last year but have not had a presurgical evaluation. Thus, the tables of training and test patients are updated weekly. The system performs as well as board-certified neurologists in identifying surgical candidates (with a sensitivity of 71% and positive predictive value of 77%) and improves with additional training, identifying surgical candidates faster than neurologists [2]. As part of the ongoing algorithmic development, the number of patients with a history of surgery included in the training set increased from 102 patients on April 10, 2016, to 195 patients on October 6, 2019.

The ACTES patient identification system went live on October 1, 2017. ACTES was prospectively evaluated using a time-and-motion study, quantitative assessments of enrollment, and postevaluation usability surveys collected from the CRCs [52]. During the time-and-motion study, an observer monitored the activities a CRC was engaged in at 30-second increments for 2 hours. The time spent per activity was compared to that prior to the use of ACTES. This study was repeated monthly for 4 months, and it was distributed among CRCs and shifts. After the implementation of ACTES, the CRCs spent 12.9% (P<.001) less time on electronic screening [52]. The quantitative assessments of enrollment evaluated the number of patients screened, the number of patients approached, and the number of patients enrolled. The use of ACTES significantly improved the number of screened patients for the majority of trials and improved the number of approached patients and enrolled patients, with statistical significance in 2 of 7 trials [52]. Finally, results from the System Usability Survey and additional open-ended questions were analyzed on a monthly basis to improve ACTES [52].

Maintenance

The epilepsy system was operational more than 90% of the time through the first 150 weeks. Throughout this time, issues were addressed by the biomedical informatics research and production IT staff. There were 10 changes made to the NLP system and

6 errors executing the pipeline of scripts. Issues extracting patient notes from the EHR were the largest reason for delays in running the NLP system, which occurred 12 out of 150 (8%) weeks.

Miscellaneous adjustments were made to the ACTES tool during the pilot phase (2017-2018) to accommodate CRC needs. ACTES was also updated 3 times because of significant updates on the institutional EHR system and its web services for real-time data extraction. Updates on the institutional EHR system and the research IT environment caused multiple system breakdowns during the evaluation period that interrupted less than 2 out of 52 (4%) weeks of operation.

Discussion

Principal Results

This work highlighted the major modifications for the integration and deployment of CDS tools from the research setting to clinical practice. We successfully added AI-based technology to the following 2 distinct clinical workflows at our institution: an automated epilepsy surgical intervention tool and an automated clinical trial eligibility screener (ACTES) system. Throughout the process, we determined that successful integration of these tools into clinical care requires adaptation to industry standards, automation of data access, logical integration into clinical workflow, and continual user feedback.

This work has several important strengths. We implemented novel, automated machine learning tools to provide decision support in a tangible fashion at our institution. These tools were well received and streamlined clinical care in the identification of qualified patients for surgery or clinical trials. Our experience with the deployment of these tools agreed with the suggestions made by Kawamoto et al [53] for successful implementation of CDS tools. Our CDS tools were implemented in real time to provide support at a natural point in the clinical workflow, so as not to disrupt or extend the timeline of care. As with their findings, our CDS tools use automatically available EHR data, where possible, to ensure clinical scalability and effective usability. In our case, we added an extra layer of testing whereby we implemented our CDS tools in a localized clinical setting in parallel to clinical care to test accuracy prior to full deployment, which allowed for continued fine-tuning of the CDS tool before it became part of clinical workflow.

Evaluation of Bias

We evaluated both tools for potential bias to ensure that the CDS recommendations were not influenced by racial disparities. The AI technology behind epilepsy surgical candidacy recommendation was evaluated for bias in terms of patient demographics, socioeconomic characteristics, and language [54]. Patient race, gender, and primary language did not bias the AI's surgical candidacy scores (*P*>.35 for all).

Considerations and Limitations

Several concerns should be considered in the implementation of a research tool into real-time clinical settings. As with most record keeping systems, the EHR systems require regular upgrades and bug fixes. This necessitates ongoing IT support to keep the pipeline operational. EHR algorithm extractions and pipeline characteristics should be placed into the EHR upgrade queue to ensure their evaluation with each upgrade cycle. To account for this, resources for both operational and research IT should be set aside to ensure a consistent system when integrated with clinical practice.

The successful deployment and continued use of these systems also required close collaboration with the stakeholders embedded in the respective clinical system. This collaboration was crucial in allowing seamless integration of the research output into daily clinical practice. Without input from the effective end users, it would be difficult to fully understand the current process, needs, as well as limitations related to workflow and data to allow for optimization of the prediction.

Conclusions

The formulation, development, and real-time implementation of two AI solutions in a clinical setting required the development of a CDS tool and pipeline using public, industry-standard programs and existing EHR web interfaces prior to integration. In our work, we found that a CDS tool's success was largely dependent upon the collaboration between machine learning experts, research collaborators, and operational IT professionals. Furthermore, longitudinal support from clinical providers and institutional leadership is necessary for continued maintenance of the deployed CDS tool with careful consideration for its long-term use.

Acknowledgments

The authors would like to thank the neurology providers, emergency medicine providers, and biomedical informatics staff at Cincinnati Children's Hospital Medical Center for providing countless feedback and support.

This work was supported by internal funds from Cincinnati Children's Hospital Medical Center, the National Library of Medicine (5R01LM011124-03), and the Agency for Healthcare Research and Quality (1R21HS024977-01). BW's work was supported by the Ruth L. Kirschstein Predoctoral Individual National Research Service Award (5F31NS115447-02).

Conflicts of Interest

JP and TG report a patent pending on the identification of surgical candidates using natural language processing, licensed to Cincinnati Children's Hospital Medical Center and a patent pending on processing clinical text with domain-specific spreading activation methods, licensed to Cincinnati Children's Hospital Medical Center.

References

1. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019 Apr 04;380(14):1347-1358. [doi: 10.1056/nejmra1814259]
2. Cohen KB, Glass B, Greiner HM, Holland-Bouley K, Standridge S, Arya R, et al. Methodological issues in predicting pediatric epilepsy surgery candidates through natural language processing and machine learning. Biomed Inform Insights 2016 May 22;8:BII.S38308. [doi: 10.4137/bii.s38308]
3. Matykiewicz P, Cohen K, Holland KD, Glauser TA, Standridge SM, Verspoor KM, et al. Earlier identification of epilepsy surgery candidates using natural language processing. 2013 Presented at: Proceedings of the 2013 Workshop on Biomedical Natural Language Processing; August 8; Sofia, Bulgaria p. 1-9.
4. Zhang X, Kim J, Patzer RE, Pitts SR, Patzer A, Schrager JD. Prediction of emergency department hospital admission based on natural language processing and neural networks. Methods Inf Med 2017 Oct 26;56(5):377-389. [doi: 10.3414/ME17-01-0024] [Medline: 28816338]
5. Zeng Z, Shi H, Wu Y, Hong Z. Survey of natural language processing techniques in bioinformatics. Comput Math Methods Med 2015;2015:674296-674210 [FREE Full text] [doi: 10.1155/2015/674296] [Medline: 26525745]
6. Milea D, Najjar RP, Jiang Z, Ting D, Vasseneix C, Xu X, et al. Artificial intelligence to detect papilledema from ocular fundus photographs. N Engl J Med 2020 Apr 30;382(18):1687-1695. [doi: 10.1056/nejmoa1917130]
7. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016 Dec 13;316(22):2402-2410. [doi: 10.1001/jama.2016.17216] [Medline: 27898976]
8. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature 2019 Aug 31;572(7767):116-119 [FREE Full text] [doi: 10.1038/s41586-019-1390-1] [Medline: 31367026]
9. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Jan 25;542(7639):115-118. [doi: 10.1038/nature21056]
10. Hicks J. The potential of claims data to support the measurement of health care quality. RAND Corporation. 2003. URL: https://www.rand.org/pubs/rgs_dissertations/RGSD171.html [accessed 2022-11-17]
11. Hirschberg J, Manning CD. Advances in natural language processing. Science 2015 Jul 17;349(6245):261-266. [doi: 10.1126/science.aaa8685] [Medline: 26185244]
12. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc 2005 Jul 01;12(4):448-457. [doi: 10.1197/jamia.m1794]
13. 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]
14. Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G. Detecting adverse events using information technology. J Am Med Inform Assoc 2003 Mar 01;10(2):115-128 [FREE Full text] [doi: 10.1197/jamia.m1074] [Medline: 12595401]
15. Petratos GN, Kim Y, Evans RS, Williams SD, Gardner RM. Comparing the effectiveness of computerized adverse drug event monitoring systems to enhance clinical decision support for hospitalized patients. Appl Clin Inform 2017 Dec 16;01(03):293-303. [doi: 10.4338/aci-2009-11-ra-0009]
16. Tinoco A, Evans RS, Staes CJ, Lloyd JF, Rothschild JM, Haug PJ. Comparison of computerized surveillance and manual chart review for adverse events. J Am Med Inform Assoc 2011 Jul 01;18(4):491-497 [FREE Full text] [doi: 10.1136/amiajnl-2011-000187] [Medline: 21672911]
17. Conway M, Dowling JN, Chapman WW. Using chief complaints for syndromic surveillance: a review of chief complaint based classifiers in North America. J Biomed Inform 2013 Aug;46(4):734-743 [FREE Full text] [doi: 10.1016/j.jbi.2013.04.003] [Medline: 23602781]
18. Ye Y, Tsui F, Wagner M, Espino JU, Li Q. Influenza detection from emergency department reports using natural language processing and Bayesian network classifiers. J Am Med Inform Assoc 2014 Sep 01;21(5):815-823 [FREE Full text] [doi: 10.1136/amiajnl-2013-001934] [Medline: 24406261]
19. Savova GK, Olson JE, Murphy SP, Cafourek VL, Couch FJ, Goetz MP, et al. Automated discovery of drug treatment patterns for endocrine therapy of breast cancer within an electronic medical record. J Am Med Inform Assoc 2012 Jun 01;19(e1):e83-e89 [FREE Full text] [doi: 10.1136/amiajnl-2011-000295] [Medline: 22140207]
20. Dublin S, Baldwin E, Walker RL, Christensen LM, Haug PJ, Jackson ML, et al. Natural Language Processing to identify pneumonia from radiology reports. Pharmacoepidemiol Drug Saf 2013 Aug 01;22(8):834-841 [FREE Full text] [doi: 10.1002/pds.3418] [Medline: 23554109]
21. Elkin PL, Froehling D, Wahner-Roedler D, Trusko B, Welsh G, Ma H, et al. NLP-based identification of pneumonia cases from free-text radiological reports. AMIA Annu Symp Proc 2008 Nov 06:172-176 [FREE Full text] [Medline: 18998791]
22. Friedman C, Alderson PO, Austin JHM, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994 Mar 01;1(2):161-174 [FREE Full text] [doi: 10.1136/jamia.1994.95236146] [Medline: 7719797]

23. Deleger L, Brodzinski H, Zhai H, Li Q, Lingren T, Kirkendall ES, et al. Developing and evaluating an automated appendicitis risk stratification algorithm for pediatric patients in the emergency department. J Am Med Inform Assoc 2013 Dec 01;20(e2):e212-e220 [FREE Full text] [doi: 10.1136/amiajnl-2013-001962] [Medline: 24130231]
24. Branch-Elliman W, Strymish J, Kudesia V, Rosen AK, Gupta K. Natural language processing for real-time catheter-associated urinary tract infection surveillance: results of a pilot implementation trial. Infect Control Hosp Epidemiol 2015 Sep 26;36(9):1004-1010. [doi: 10.1017/ice.2015.122] [Medline: 26022228]
25. Tso GJ, Tu SW, Oshiro C, Martins S, Ashcraft M, Yuen KW, et al. Automating guidelines for clinical decision support: knowledge engineering and implementation. AMIA Annu Symp Proc 2016;2016:1189-1198 [FREE Full text] [Medline: 28269916]
26. Klein ME, Parvez MM, Shin J. Clinical implementation of pharmacogenomics for personalized precision medicine: barriers and solutions. J Pharm Sci 2017 Sep;106(9):2368-2379. [doi: 10.1016/j.xphs.2017.04.051] [Medline: 28619604]
27. Kilsdonk E, Peute LW, Jaspers MWM. Factors influencing implementation success of guideline-based clinical decision support systems: a systematic review and gaps analysis. Int J Med Inform 2017 Dec;98:56-64. [doi: 10.1016/j.ijmedinf.2016.12.001] [Medline: 28034413]
28. Castillo RS, Kelemen A. Considerations for a successful clinical decision support system. Comput Inform Nurs 2013 Jul;31(7):319-26; quiz 327-8. [doi: 10.1097/NXN.0b013e3182997a9c] [Medline: 23774450]
29. Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005 Mar 9;293(10):1223-1238. [doi: 10.1001/jama.293.10.1223] [Medline: 15755945]
30. Wright A, Ash JS, Aaron S, Ai A, Hickman TT, Wiesen JF, et al. Best practices for preventing malfunctions in rule-based clinical decision support alerts and reminders: results of a Delphi study. Int J Med Inform 2018 Oct;118:78-85 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.08.001] [Medline: 30153926]
31. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med 2018 Aug 28;1(1):39 [FREE Full text] [doi: 10.1038/s41746-018-0040-6] [Medline: 31304320]
32. Hollon TC, Pandian B, Adapa AR, Urias E, Save AV, Khalsa SSS, et al. Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. Nat Med 2020 Jan;26(1):52-58 [FREE Full text] [doi: 10.1038/s41591-019-0715-9] [Medline: 31907460]
33. Titano JJ, Badgeley M, Schefflein J, Pain M, Su A, Cai M, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. Nat Med 2018 Sep 13;24(9):1337-1341. [doi: 10.1038/s41591-018-0147-y] [Medline: 30104767]
34. Wang P, Liu X, Berzin T, Glissen Brown J, Liu P, Zhou C, et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. Lancet Gastroenterol Hepatol 2020 Apr;5(4):343-351 [FREE Full text] [doi: 10.1016/s2468-1253(19)30411-x]
35. Trinkley KE, Kahn MG, Bennett TD, Glasgow RE, Haugen H, Kao DP, et al. Integrating the practical robust implementation and sustainability model with best practices in clinical decision support design: implementation science approach. J Med Internet Res 2020 Oct 29;22(10):e19676 [FREE Full text] [doi: 10.2196/19676] [Medline: 33118943]
36. Engel J, McDermott MP, Wiebe S, Langfitt JT, Stern JM, Dewar S, Early Randomized Surgical Epilepsy Trial (ERSET) Study Group. Early surgical therapy for drug-resistant temporal lobe epilepsy: a randomized trial. JAMA 2012 Mar 07;307(9):922-930 [FREE Full text] [doi: 10.1001/jama.2012.220] [Medline: 22396514]
37. Choi H, Sell RL, Lenert L, Muennig P, Goodman RR, Gilliam FG, et al. Epilepsy surgery for pharmacoresistant temporal lobe epilepsy: a decision analysis. JAMA 2008 Dec 03;300(21):2497-2505. [doi: 10.1001/jama.2008.771] [Medline: 19050193]
38. Engel J, Wiebe S, French J, Sperling M, Williamson P, Spencer D, et al. Practice parameter: temporal lobe and localized neocortical resections for epilepsy. Epilepsia 2003 Jun;44(6):741-751 [FREE Full text] [doi: 10.1046/j.1528-1157.2003.48202.x] [Medline: 12790886]
39. Cross JH, Jayakar P, Nordli D, Delalande O, Duchowny M, Wieser HG, International League against Epilepsy, Subcommission for Paediatric Epilepsy Surgery, Commissions of NeurosurgeryPaediatrics. Proposed criteria for referral and evaluation of children for epilepsy surgery: recommendations of the Subcommission for Pediatric Epilepsy Surgery. Epilepsia 2006 Jun;47(6):952-959 [FREE Full text] [doi: 10.1111/j.1528-1167.2006.00569.x] [Medline: 16822241]
40. Choi H, Carlino R, Heiman G, Hauser WA, Gilliam FG. Evaluation of duration of epilepsy prior to temporal lobe epilepsy surgery during the past two decades. Epilepsy Res 2009 Oct;86(2-3):224-227 [FREE Full text] [doi: 10.1016/j.eplepsyres.2009.05.014] [Medline: 19581072]
41. Kwan P, Schachter SC, Brodie MJ. Drug-resistant epilepsy. N Engl J Med 2011 Sep 08;365(10):919-926. [doi: 10.1056/nejmra1004418]
42. Englot DJ, Ouyang D, Garcia PA, Barbaro NM, Chang EF. Epilepsy surgery trends in the United States, 1990-2008. Neurology 2012 Mar 21;78(16):1200-1206. [doi: 10.1212/wnl.0b013e318250d7ea]

43. Tsochantaridis I, Hofmann T, Thorsten J, Altun Y. Support vector machine learning for interdependent and structured output spaces. In: Proceedings of the twenty-first international conference on Machine learning. 2004 Presented at: ICML '04; July 4-8; Banff, Alberta, Canada. [doi: 10.1145/1015330.1015341]
44. Wissel BD, Greiner HM, Glauser TA, Holland-Bouley KD, Mangano FT, Santel D, et al. Prospective validation of a machine learning model that uses provider notes to identify candidates for resective epilepsy surgery. Epilepsia 2020 Jan 29;61(1):39-48 [FREE Full text] [doi: 10.1111/epi.16398] [Medline: 31784992]
45. Embi PJ, Payne PRO. Clinical research informatics: challenges, opportunities and definition for an emerging domain. JAMIA 2009 May 01;16(3):316-327. [doi: 10.1197/jamia.m3005]
46. Ni Y, Kennebeck S, Dexheimer JW, McAneney CM, Tang H, Lingren T, et al. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. J Am Med Inform Assoc 2015 Jan;22(1):166-178 [FREE Full text] [doi: 10.1136/amiajnl-2014-002887] [Medline: 25030032]
47. Baldwin B, Dayanidhi K. Natural language processing with Java and LingPipe Cookbook. Birmingham, UK: Packt Publishing Ltd; 2014.
48. 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. [doi: 10.5555/1953048.2078195]
49. Joachims T. Text categorization with support vector machines: learning with many relevant features. : Springer Berlin Heidelberg; 1998 Presented at: Machine Learning: ECML-98; 1998; Berlin, Heidelberg p. 137-142. [doi: 10.1007/BFb0026683]
50. 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]
51. Dexheimer JW, Tang H, Kachelmeyer A, Hounchell M, Kennebeck S, Solti I, et al. A time-and-motion study of clinical trial eligibility screening in a pediatric emergency department. Pediatr Emerg Care 2019 Dec;35(12):868-873 [FREE Full text] [doi: 10.1097/PEC.0000000000001592] [Medline: 30281551]
52. Ni Y, Bermudez M, Kennebeck S, Liddy-Hicks S, Dexheimer J. A real-time automated patient screening system for clinical trials eligibility in an emergency department: design and evaluation. JMIR Med Inform 2019 Jul 24;7(3):e14185 [FREE Full text] [doi: 10.2196/14185] [Medline: 31342909]
53. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 2005 Apr 2;330(7494):765 [FREE Full text] [doi: 10.1136/bmj.38398.500764.8F] [Medline: 15767266]
54. Wissel BD, Greiner HM, Glauser TA, Mangano FT, Santel D, Pestian JP, et al. Investigation of bias in an epilepsy machine learning algorithm trained on physician notes. Epilepsia 2019 Sep 23;60(9):e93-e98 [FREE Full text] [doi: 10.1111/epi.16320] [Medline: 31441044]

Abbreviations

ACTES: automated clinical trial eligibility screener
AI: artificial intelligence
CDS: clinical decision support
CRC: clinical research coordinator
ED: emergency department
EHR: electronic health record
IT: information technology
NLP: natural language processing

Edited by T Hao; submitted 08.03.22; peer-reviewed by A Keniston, H Monday, G Nneji; comments to author 11.07.22; revised version received 01.09.22; accepted 19.09.22; published 16.12.22.

Please cite as:
Kanbar LJ, Wissel B, Ni Y, Pajor N, Glauser T, Pestian J, Dexheimer JW
Implementation of Machine Learning Pipelines for Clinical Practice: Development and Validation Study
JMIR Med Inform 2022;10(12):e37833
URL: https://medinform.jmir.org/2022/12/e37833
doi:10.2196/37833
PMID:36525289

Original Paper

Boosting Delirium Identification Accuracy With Sentiment-Based Natural Language Processing: Mixed Methods Study

Lu Wang[1,2], PhD; Yilun Zhang[1], MSc; Mark Chignell[1], PhD; Baizun Shan[1], MSc; Kathleen A Sheehan[3,4], MSc, MD, PhD; Fahad Razak[3,5], MPhil, MD; Amol Verma[3,5], MPhil, MD

[1]Department of Mechanical & Industrial Engineering, University of Toronto, Toronto, ON, Canada

[2]Department of Computer Science, Texas State University, San Marcos, TX, United States

[3]GEMINI - The General Medicine Inpatient Initiative, Unity Health Toronto, Toronto, ON, Canada

[4]Department of Psychiatry, University of Toronto, Toronto, ON, Canada

[5]Faculty of Medicine & Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Mark Chignell, PhD
Department of Mechanical & Industrial Engineering
University of Toronto
RM 8171A, Bahen Building
40 St George Rd
Toronto, ON, M5S 2E4
Canada
Phone: 1 6473898951
Email: chignel@mie.utoronto.ca

Abstract

Background: Delirium is an acute neurocognitive disorder that affects up to half of older hospitalized medical patients and can lead to dementia, longer hospital stays, increased health costs, and death. Although delirium can be prevented and treated, it is difficult to identify and predict.

Objective: This study aimed to improve machine learning models that retrospectively identify the presence of delirium during hospital stays (eg, to measure the effectiveness of delirium prevention interventions) by using the natural language processing (NLP) technique of sentiment analysis (in this case a feature that identifies sentiment toward, or away from, a delirium diagnosis).

Methods: Using data from the General Medicine Inpatient Initiative, a Canadian hospital data and analytics network, a detailed manual review of medical records was conducted from nearly 4000 admissions at 6 Toronto area hospitals. Furthermore, 25.74% (994/3862) of the eligible hospital admissions were labeled as having delirium. Using the data set collected from this study, we developed machine learning models with, and without, the benefit of NLP methods applied to diagnostic imaging reports, and we asked the question “can NLP improve machine learning identification of delirium?”

Results: Among the eligible 3862 hospital admissions, 994 (25.74%) admissions were labeled as having delirium. Identification and calibration of the models were satisfactory. The accuracy and area under the receiver operating characteristic curve of the main model with NLP in the independent testing data set were 0.807 and 0.930, respectively. The accuracy and area under the receiver operating characteristic curve of the main model without NLP in the independent testing data set were 0.811 and 0.869, respectively. Model performance was also found to be stable over the 5-year period used in the experiment, with identification for a likely future holdout test set being no worse than identification for retrospective holdout test sets.

Conclusions: Our machine learning model that included NLP (ie, sentiment analysis in medical image description text mining) produced valid identification of delirium with the sentiment analysis, providing significant additional benefit over the model without NLP.

(JMIR Med Inform 2022;10(12):e38161) doi:10.2196/38161

KEYWORDS

delirium diagnosis; data mining; medical image description; text mining and analysis; sentiment analysis

Introduction

Background

Delirium is described as “acute brain failure” and is considered both a “medical emergency” and “quiet epidemic” [1,2]. It is the most common neuropsychiatric condition among medically ill and hospitalized patients [3]. It is also recognized as a quality of care indicator in Canada, the United States, the United Kingdom, and Australia [4-8]. Symptoms of delirium can be severe and distressing for both patients and caregivers [9,10] and result from a complex interaction between predisposing and precipitating factors [9]. Affecting up to 50% of older hospital patients, those with delirium are more than twice as likely to die in the hospital or require nursing home placement [11-14]. The long-term effects of delirium are serious, as it is associated with worsening cognitive impairment and incident dementia [14-17]. Patients with delirium have longer hospitalizations, increased readmission rates, and more than double the health care costs. The study by Leslie et al [18] indicated that 1-year health costs associated with delirium ranged from US $16,303 to US $64,421 per patient. More recent estimates suggest that it accounts for US $183 billion dollars of annual health care expenditures in the United States [18,19]. Up to 40% of cases are preventable and many of the remaining cases of delirium could be better managed with implementation of standardized multicomponent programs [19,20]. These programs result in up to US $3800 in savings per patient in hospital costs and >US $16,000 in savings per person-year in the year following an episode of delirium [19,20]. However, in routine clinical care, there is a significant practice gap, and most hospitals have not consistently implemented best practices [19-21].

A key barrier in using delirium as a quality indicator is the lack of a reliable and scalable method for early identification of delirium cases. Clinicians are not good at recognizing delirium using clinical gestalt, with corresponding recognition rates ranging between 16% and 35% [22]. The Confusion Assessment Method (CAM) [23] is one of the number of screening tools for delirium, but it takes time and training to use; as a result, tools such as CAM are used relatively infrequently. For instance, Hogan et al [23] found that only 28% of emergency departments with a geriatric focus used delirium screening tools.

As delirium is difficult to recognize in situ, there has been interest in recognizing delirium after it has occurred, either through administrative chart review (ie, looking for evidentiary factors such as the use of antipsychotic drugs) or through retrospective identification. Ideally, identification of delirium would be prospective, proving a method to identify those at the highest risk of developing delirium to target delirium identification interventions for these individuals. However, retrospective identification of delirium can also be useful in determining delirium rates, which can serve as quality indicators and measures of effectiveness for interventions aimed at quality improvement.

Numerous models for predicting delirium have been developed based on known predisposing and precipitating risk factors [18]. However, current models have limitations [24]. First, they rely on variables not routinely collected as part of clinical care such as preexisting cognitive impairment and functional status, making them difficult to scale [25]. For example, the United Kingdom’s National Institute for Clinical Excellence delirium risk identification model requires information on cognitive impairment and sensory impairment to be available in the electronic record [26-28]. Second, a systematic review of delirium identification models highlighted their inadequate identification and numerous methodological concerns regarding how the models were validated such as their accuracy and inadequate predictive ability. The review concluded that model performance was likely exaggerated [26]. Third, prior risk identification models for delirium have tended to use a limited set of machine learning methods [7,29-33] and have tended to neglect text data [34].

With the growing availability of electronic clinical data repositories such as the one used in this study, methods such as data mining and machine learning can supplement or replace conventional statistical models [27,32,34-38]. Natural language processing (NLP) methods for medical text mining are required to extract valuable medical information and derive calculable variables for identification models [39]. NLP has proven to be highly effective in extracting the information from medical text into a computationally useful form that can support clinical decision-making [40-47].

Sentiment analysis analyzes the text for the sentiment of the writer (eg, positive vs negative, or in our case delirium vs non–delirium-related text) using machine learning and NLP [46-48]. We adapted sentiment analysis to predict sentiment concerning delirium status. Thus, positive (with delirium) and negative (without delirium) status was a new (binary) sentiment feature in the subsequent analysis. Using this delirium-based text sentiment analysis, we created a text-derived feature that estimated the delirium status for each admission.

Objective

The overall research goal of our project was to retrospectively identify delirium cases during hospitalization using all data available from admission to discharge to estimate delirium rates and thereby quantify the effect of quality improvement interventions related to delirium. In this study, we focus on the methodological goal of demonstrating the value of incorporating NLP methods in the retrospective identification of delirium.

Methods

Data Source

Overview

The General Medicine Inpatient Initiative (GEMINI) is a multi-institutional research collaboration in Ontario, Canada. GEMINI has developed infrastructure and methods to collect and standardize electronic clinical data from hospitals. The data for this study were obtained from 6 hospitals (St Michael’s Hospital, Toronto General Hospital, Toronto Western Hospital, Trillium Credit Valley Hospital, Trillium Mississauga Hospital, and Sunnybrook Hospital). GEMINI is emerging as a rich resource for clinical research and quality measurement [4,49-52].

A rigorous internal validation process demonstrated 98% to 100% accuracy across key data types [50].

In GEMINI, administrative health data are linked to clinical data extracted from hospital information systems at the individual patient level (Table 1).

Table 1. Data contained in the General Medicine Inpatient Initiative project.

Data type	Patient details	Physician and room	Laboratory	Imaging	Pharmacy	Clinical documentation	Microbiology
Selected variables	• Demographics • Comorbidities • Diagnoses • Procedures • Costs	• Physician details • Transfer details	• Biochemistry • Hematology • Transfusion	• Radiologist reports of diagnostic and interventional imaging	• Medication • Dose • Route	• Physician orders • Vital signs	• Organism • Antimicrobial susceptibility • Collection details

Administrative Data

Patient-level characteristics were collected from hospitals as reported to the Canadian Institute for Health Information Discharge Abstract Database and the National Ambulatory Care Reporting System. Diagnostic data and interventions were coded using the enhanced Canadian International Statistical Classification of Diseases and Related Health Problems and the Canadian Classification of Health Interventions.

Clinical Data

Data from the electronic information systems in GEMINI include laboratory test results (biochemistry, hematology, and microbiology), blood transfusions, in-hospital medications, vital signs, imaging reports, and room transfers. The quality of the key elements of these data was ensured through statistical quality control processes and direct data validation [53]. GEMINI data extraction methods allow access to a wealth of data ideal for text processing methods, including radiologist reports of diagnostic imaging.

The delirium cases in the research reported here were identified through manual medical record review by trained medical professionals using a validated method [54]. This method relies primarily on the identification of delirium or its numerous synonyms (eg, confusion) through a detailed review of physicians, nurses, and interprofessional documentation. The method has good sensitivity (74%) and specificity (83%) compared with clinical assessment and is considered a suitable gold standard for the identification of delirium for research and quality improvement [54].

We used 11 data files from a GEMINI data set that contained 3862 hospital admissions manually labeled according to delirium status. The data files include clinical and administrative data, as described in Table 1. However, labeling delirium is highly labor intensive, with trained reviewers answering the following question as part of the process: “Is there any evidence from the chart of acute confusional state (e.g., delirium, mental status change, inattention, disorientation, hallucinations, agitation, inappropriate behavior, etc.)? Review the entire medical record, including progress notes, nursing notes, consult notes, etc.” Thus, although chart review labels can be used to train more efficient machine learning methods, they are too expensive to use in label all older patients in terms of whether they experienced delirium during their hospital stay.

In our study, we used the chart review method [51] to label a subset of cases in our data set with respect to delirium. Interrater reliability was assessed by having 5% of the charts blindly reviewed by a second abstractor, achieving 90% interrater reliability. This resulted in the 3862 hospital admissions used in the analyses reported in this paper. The data files include clinical and administrative data, as described in Table 1.

Ethics Approval

The research ethics board (REB) at the Toronto Academic Health Science Network approved the GEMINI study (REB reference number 15-087). The extension of the REB approval was issued by the Unity Health Toronto REB (reference number 15-087). A separate REB approval was obtained for Trillium Health Partners.

This paper is also part of the GEMINI substudy, named “Using artificial intelligence to identify and predict delirium among hospitalized medical patients,” which was approved by the University of Toronto REB (approved as reference number 38377).

Data Preprocessing

The data tables contained in GEMINI were merged into a single table worksheet suitable for conducting machine learning. Before that, merger relevant variables were selected from the data tables, as described in the following subsections.

Laboratory Tests

A total of 45 medical tests were included in this data file, for example, blood urea nitrogen, mean cell volume, and high sensitivity troponin. Note that in each admission, not all 45 medical tests were performed, although some tests were performed several times in the same patient. In the original laboratory tests data file, each instance of a medical test corresponded to a separate record. We converted the laboratory tests table to one with a single row per admission, where each column represented a different test. As patients typically received a small subset of the available tests, there were many empty cells (ie, sparsity), and some cells had to represent multiple instances of the same test. To address the problem of sparse variables, we converted them to 1 or 0 flag variables (1 for test performed and 0 for test not performed). For frequently performed tests, we recorded the minimum, maximum, median, and frequency of the test results for each admission. If a test was administered at least five times in >50% of the admissions, we calculated the SD of the test results across each admission as an additional summary measure.

Patient Diagnosis

We first mapped the International Classification of Diseases, Tenth Revision (ICD-10) to the Clinical Classification Software (CCS) discharge diagnosis codes in a process that we previously described [4,49,50,55]. We use all available ICD-10 codes, including those assigned retrospectively, and this should not be considered data leakage but rather leveraging all available data to serve the use. The physician team identified 240 unique CCS codes potentially relevant to delirium. We then created flag variables (Boolean) for these 240 unique CCS codes that indicated whether the admission involved each of the diagnoses. Note that we did not create flag variables for ICD-10 codes because this would have dramatically increased the number of features in the analysis.

Clinical Interventions

This set of features covered a range of clinical interventions including surgical and endoscopic procedures as coded by the Canadian Classification of Health Interventions. Two variables were used to record the number of interventions for each admission. The first derived variable was the number of interventions performed per admission (including repetitions of the same intervention). The second derived variable counted the number of unique interventions per admission. No other information regarding interventions was used in the data file.

Room Transfers

We calculated the number of room transfers for each admission, which was the only variable used from this data table.

Clinical Risk Scores

We used the following clinical scores, which are markers of illness severity and patient risk of adverse outcomes: Charlson Comorbidity Index [56], Laboratory-Based Acute Physiology Score [57], and Kidney Disease: Improving Global Outcomes Acute Kidney Injury stages [58].

Emergency Department Triage Score

We applied one-hot encoding on the feature representing the patient's illness severity at the time of emergency department triage with a 5-point scale, as measured by the Canadian Triage and Acuity Scale [59].

Administrative Admission and Discharge Data

We applied one-hot encoding on the feature representing the type of medical services that the patient was admitted to and discharged from as per the hospital admission, discharge, and transfer system. We also calculated hospital length of stay and derived a feature to indicate where the patient was discharged to.

Medications

This file had 1 row per admission and was used as is.

Special Care Unit

Only 320 admissions had special care unit information, so we created a flag variable with binary coding to indicate whether patients were cared for in a special care unit at any point during the admission.

Blood Transfusion

This medical data table contained only 429 admissions that included blood transfusion information; therefore, we created 1 column with binary coding to represent its presence or absence.

NLP on Radiologist Reports of Diagnostic Imaging

The medical imaging data table contained the text description of magnetic resonance images and computed tomography scans, which were filtered to include only brain or head imaging. Similar to the laboratory tests data file, there was 1 row per imaging test; therefore, there could be multiple rows per admission. If there were multiple tests per admission, we first concatenated the text descriptions across the tests and then used text mining on this file by cleaning, tokenizing, and vectorizing.

The data set used for machine learning represented data integrated from multiple sources, for example, laboratory results, medications, radiologist reports, and administrative data. We adapted sentiment analysis to predict sentiment concerning delirium status. Thus positive (with delirium) and negative (without delirium) status was a binary sentiment that then formed a new feature in the subsequent analysis. Using this delirium-based text sentiment analysis, we created a text-derived feature that estimated the delirium status for each admission.

Preliminary text analysis was carried out before the sentiment analysis. Text cleaning included uppercase transformation, stop words removal, punctuation removal, intraword splitting, tokenization, and lemmatization and was performed using the *nltk* [39] and *sklearn* [60] packages. Next, term frequency–inverse document frequency, word count representation, and *n*-gram methods were applied for text vectorization.

A total of 8 baseline machine learning classification models were then trained to perform sentiment analysis, that is, logistic regression, Naive Bayes, support vector machine (SVM), decision tree, random forest, gradient boosting, *XGboost*, and multilayer perceptron. Hyperparameter tuning was applied using *RandomSearchCV* (ie, a randomized search on hyperparameters optimized by cross-validated search over parameter settings) [60].

Gradient boosting was selected as the final sentiment analysis method because its F_1-score was the highest among the 8 classifiers. The final model was a stochastic gradient boosting (with a 0.8 subsample) that used 200 estimators, with Friedman mean square error as the criterion and a maximum depth of 3. We then created a feature with the predicted binary sentiment from the description of the medical images in the text using the selected gradient boosting model.

We integrated this new feature with 10 laboratory tests and electronic health record data to create a complete data file for training and testing machine learning identification models.

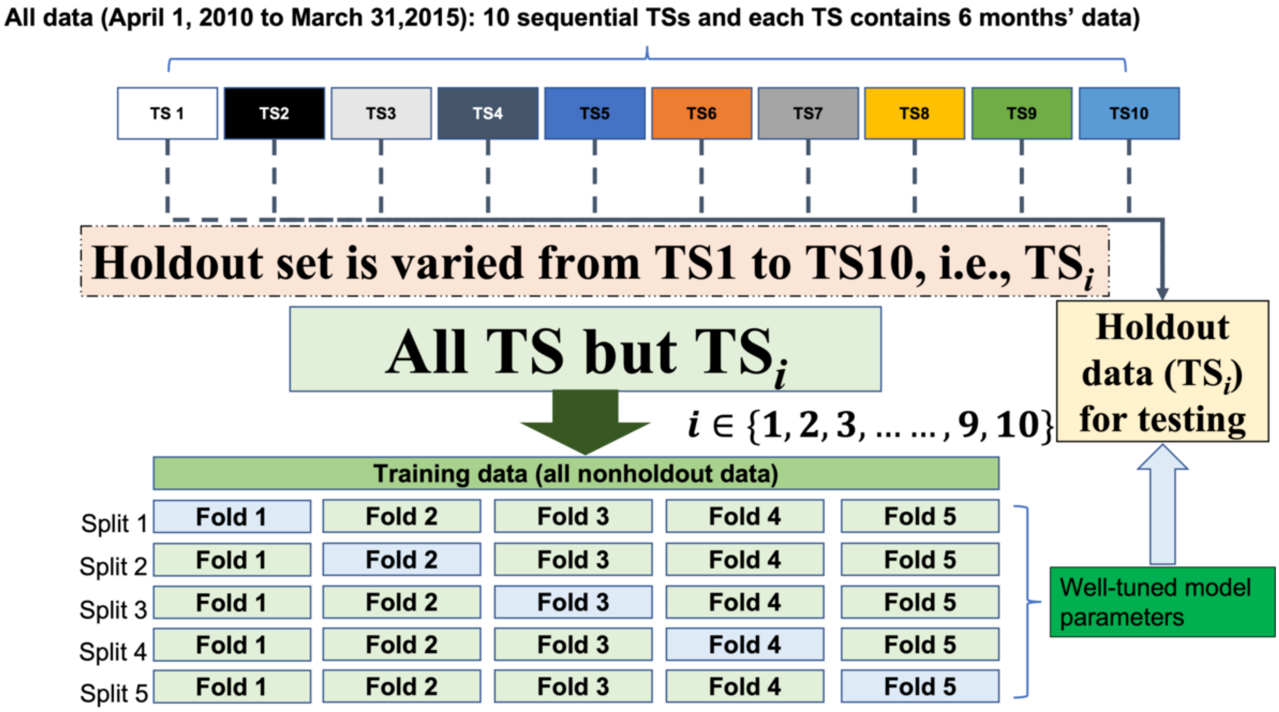
Model Construction and Training

A total of 12 supervised classification algorithms with the task of predicting delirium status were implemented. The 12 machine learning algorithms covering most types of machine learning models were as follows:

- Ensemble machine learning models: gradient boosting classifier, AdaBoost classifier, random forest, and voting classifier soft
- Nonparametric machine learning models: K-nearest neighbor and decision tree
- Linear parametric machine learning models: logistic regression, linear SVM, and linear discriminant analysis
- Nonlinear parametric machine learning models: quadratic discriminant analysis, neural network: multilayer perceptron classifier in deep learning
- Bayesian-based machine learning models: Gaussian Naive Bayes.

For the modeling, we split our integrated complete data into 2 parts, a training set and a testing set. As shown in Figure 1, the data extended over a 5-year period, from April 1, 2010, to March 31, 2015. We divided this period into ten 6-month segments. We treated the first 9 segments, that is, April 1, 2010, to September 30, 2014, as the training set. The last 6-month period, that is, October 1, 2014, to March 1, 2015, was used as holdout data (ie, testing set) to estimate the likely future performance of the model that was forward in time relative to the data used in building the model. This allowed us to assess whether there was any nonstationarity in the data, which would affect our ability to predict delirium in the future based on models developed on currently available data as transferability to future data.

Figure 1. Data splits for models training and testing on a rolling basis. TS: time segment.


All data (April 1, 2010 to March 31,2015): 10 sequential TSs and each TS contains 6 months' data)
TS 1
TS2
TS3
TS4
TS5
TS6
TS7
TS8
TS9
TS10
Holdout set is varied from TS1 to TS10, i.e., TS_i
All TS but TS_i
$i \in \{1, 2, 3, \dots\dots, 9, 10\}$
Holdout data (TS_i) for testing
Training data (all nonholdout data)
Well-tuned model parameters

Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

In the training set, we used 5-fold cross-validation to tune the model parameters for each of the 12 machine learning algorithms. We then used the tuned parameters from the 5-fold cross-validation to identify delirium status of each admission in the testing or holdout set.

Results

Overview

We tested the model performance on the holdout testing set and calculated 6 evaluation metrics to find the best model, that is, accuracy, precision, recall or sensitivity, F_1-score, specificity, and area under the receiver operating characteristic curve (ROC-AUC).

Accuracy answers the question of how many admissions did we correctly label out of all the admissions.

Precision answers the question of how many of those who we predicted as having delirium actually had delirium.

Sensitivity represents the proportion of people with delirium who were correctly labeled as having delirium.

F_1-score is a weighted average of the precision or recall, where the F_1-score reaches its best value at 1 and worst score at 0.

Specificity answers the question of how many negative instances (ie, people with no delirium) were correctly predicted.

The ROC curve was plotted using the true-positive rate against the false-positive rate at various threshold settings. The calculated ROC-AUC indicated the probability that our binary classifier ranked a randomly chosen positive instance higher than a randomly chosen negative one (assuming “positive” ranks higher than “negative”).

The 12 machine learning algorithms, along with hyperparameter tuning and cross-validation, were implemented in the Python package *Scikit-learn* [60]. Hyperparameter tuning was conducted using the *RandomizedSearchCV* and *GridSearchCV* functions. Cross-validation was used via the *cross_val_score*, *cross_validate*, and *cross_val_predict* functions.

The gradient boosting classifier was trained using the *GradientBoostingClassifier* function. The AdaBoost classifier used the *AdaBoostClassifier* function. The neural network classifier was implemented using the *MLPClassifier* function. The decision tree classifier was implemented using the *DecisionTreeClassifier* function. K-nearest neighbor classification was trained using the *KNeighborsClassifier* function. The logistic regression classifier used the *LogisticRegression* function. The random forest classifier was implemented using the *RandomForest* classifier function. The SVM method used the *svm* function. The Gaussian Naive Bayes method implemented the *GaussianNB* function. The linear discriminant analysis classifier was trained using the *LinearDiscriminantAnalysis* function. The quadratic discriminant analysis classifier used the *QuadraticDiscriminantAnalysis* function. The voting classifiers with soft settings were implemented using the *Voting Classifier* function.

Experimental Results

We trained these models using hyperparameter tuning and 5-fold cross-validation on the first 9 time segments. We present the results from the 3 best-performing models in Table 2, and the results from the other 9 models are presented in Multimedia Appendix 1. In both tables, we report the average performance over 5 folds for the data from the first 9 time segments.

We then tested our delirium identification (sentimental or +NLP) model, which incorporated NLP in the training process. We compared the results of the +NLP model with the results obtained for the unsentimental (–NLP) delirium identification model that was trained, without NLP, on the last 6-month data in the GEMINI data set. The performance of the 3 best-performing models in predicting delirium labels in the last 6 months of the data is shown in Table 3. A similar presentation of the results is shown for the other 9 models in Multimedia Appendix 2. It should be noted that we used well-tuned parameters from the best-performing models of the training data on the testing data.

Table 2. Comparison of models in the 3 best-performing algorithms: average training results using 5-fold cross-validation on training set (April 1, 2010, to September 30, 2014).

Models	Gradient boosting classifier	AdaBoost classifier	Random forest
Accuracy			
Delirium (+NLP[a])	*0.868* [b]	0.866	0.826
Delirium (–NLP)	0.797	0.795	0.768
Precision			
Delirium (+NLP)	0.78	0.794	*0.833*
Delirium (–NLP)	0.747	0.75	0.8
Recall			
Delirium (+NLP)	*0.678*	0.649	0.398
Delirium (–NLP)	0.341	0.329	0.141
Specificity			
Delirium (+NLP)	0.935	0.942	0.975
Delirium (–NLP)	0.957	0.958	*0.988*
ROC-AUC[c]			
Delirium (+NLP)	*0.91*	0.895	0.897
Delirium (–NLP)	0.83	0.834	0.83
***F_1*-score**			
Delirium (+NLP)	*0.722*	0.712	0.529
Delirium (–NLP)	0.463	0.452	0.239

[a] NLP: natural language processing.

[b] Highest performance values are italicized.

[c] ROC-AUC: area under the receiver operating characteristic curve.

Table 3. Comparison of 3 types of models in the 3 best-performing algorithms: model performance on holdout set 10 (October 1, 2014, to March 31, 2015).

Models	Gradient boosting classifier	AdaBoost classifier	Random forest
Accuracy			
Delirium (+NLP[a])	*0.853* [b]	0.835	0.835
Delirium (–NLP)	0.807	0.811	0.776
Precision			
Delirium (+NLP)	0.742	0.725	*0.866*
Delirium (–NLP)	0.74	0.747	0.806
Recall			
Delirium (+NLP)	*0.669*	0.594	0.436
Delirium (–NLP)	0.406	0.421	0.188
Specificity			
Delirium (+NLP)	0.918	0.92	0.976
Delirium (–NLP)	0.949	0.949	*0.984*
ROC-AUC[c]			
Delirium (+NLP)	0.922	0.917	*0.93*
Delirium (–NLP)	0.848	0.849	0.869
F_1-score			
Delirium (+NLP)	*0.704*	0.653	0.58
Delirium (–NLP)	0.524	0.538	0.305

[a] NLP: natural language processing.

[b] Highest performance values are italicized.

[c] ROC-AUC: area under the receiver operating characteristic curve.

In the training set, our proposed delirium (+NLP) models performed the best in terms of accuracy, precision, recall or sensitivity, rate, ROC-AUC, and F_1-score, whereas delirium (–NLP) models generated the best specificity. In the testing set, the performances in both delirium (+NLP) and delirium (–NLP) models continued the same trend.

Note that F_1-score is the balance of sensitivity and precision, and ROC-AUC is generated by sensitivity and specificity so that our delirium (+NLP) models performed the best in terms of balancing sensitivity, precision, and specificity. In acute diseases such as delirium, sensitivity is particularly important because the cost of failed identification of a disease (a miss) is higher than the cost of a false alarm. Thus, the present results indicate that the sentimental (vs unsentimental) delirium identification model should be more useful in clinical practice.

We also tested the +NLP and –NLP models across time, moving the holdout set across each of the 9 time segments one at a time, before using the most recent time segment as the holdout set. Thus, each of the time segments was used as the testing set, whereas the other 9 time segments were treated as the training set on a rolling basis, as shown in Figure 1. The corresponding data distribution of training and independent holdout or testing data are presented in Table 4. Tables 5 and 6 present the data distribution of patient characteristics of the cohort across the data splits.

Figure 2 shows the identification results for the best-performing machine learning algorithm, that is, the gradient boosting across the 10 time segments. The 8 panels in the figure represent the 8 evaluation metrics used.

Note that the 2 different lines shown in each of the 8 panels within Figure 2 represent the results on the corresponding evaluation metrics for the 2 different types of models (ie, Delirium [+NLP] and Delirium [–NLP]). The 10 data points in each line show how the performance varied as the timing of the holdout time segment varied. Overall, the identification performance of the sentimental (+NLP) model was better than that of the unsentimental (–NLP) model. In addition, the performance of the sentimental (+NLP) model tended to be more stable across the different time segments than the other schemes. It can also be seen that precision, recall, and F_1-score tended to be less stable over time than the other 3 measures, even though these performance measures remained relatively stable for the delirium (+NLP) model.

Figure 3 presents the calibration of the gradient boosting model that was found to provide the best overall performance.

Table 4. Data distribution of training and holdout sets for each time segment (TS). Note that positive admissions indicate that the patients were diagnosed with delirium upon their admissions, whereas negative admissions were not.

Different TS as holdout set on a rolling basis	Training set			Holdout set		
	Number of admissions	Number of negative admissions	Number of positive admissions	Number of admissions	Number of negative admissions	Number of positive admissions
TS1	3541	2635	906	321	233	88
TS2	3531	2627	904	331	241	90
TS3	3494	2581	913	368	287	81
TS4	3488	2596	892	374	272	102
TS5	3526	2620	906	336	248	88
TS6	3479	2585	894	383	283	100
TS7	3446	2560	886	416	308	108
TS8	3476	2580	896	386	288	98
TS9	3424	2536	888	438	332	106
TS10	3353	2492	861	509	376	133

Table 5. Data information in patient characteristics for age and gender of the cohort across in 10 time segments (TSs) in both training and testing data sets. Three adult age groups are defined: young adults aged 18-44 years, middle-aged adults aged 45-64 years, and older adults aged ≥65 years.

TS	Gender				Age					
	Training		Testing		Training			Testing		
	Male, n (%)	Female, n (%)	Male, n (%)	Female, n (%)	Young adults, n (%)	Middle-aged adults, n (%)	Older adults, n (%)	Young adults, n (%)	Middle-aged adults, n (%)	Older adults, n (%)
TS1 (training: n=3541; testing: n=321)	1753 (49.51)	1788 (50.49)	162 (50.5)	159 (49.5)	430 (12.14)	844 (23.84)	2267 (64.02)	36 (11.2)	81 (25.2)	204 (63.5)
TS2 (training: n=3531; testing: n=331)	1736 (49.16)	1795 (50.84)	179 (54.1)	152 (45.9)	421 (11.92)	845 (23.93)	2265 (64.15)	45 (13.6)	80 (24.2)	206 (62.2)
TS3 (training: n=3494; testing: n=368)	1746 (49.97)	1748 (50.03)	169 (45.9)	199 (54.1)	417 (11.93)	845 (24.18)	2232 (63.88)	49 (13.3)	80 (21.7)	239 (64.9)
TS4 (training: n=3488; testing: n=374)	1737 (49.8)	1751 (50.2)	178 (47.6)	196 (52.4)	415 (11.9)	854 (24.48)	2219 (63.62)	51 (13.6)	71 (18.9)	252 (67.4)
TS5 (training: n=3526; testing: n=336)	1748 (49.57)	1778 (50.43)	167 (49.7)	169 (50.3)	423 (12)	838 (23.77)	2265 (64.24)	43 (12.8)	87 (25.9)	206 (61.3)
TS6 (training: n=3479; testing: n=383)	1728 (49.67)	1751 (50.33)	187 (48.8)	196 (51.2)	417 (11.99)	832 (23.91)	2230 (64.1)	49 (12.8)	93 (24.3)	241 (62.9)
TS7 (training: n=3446; testing: n=416)	1700 (49.33)	1746 (50.67)	215 (51.7)	201 (48.3)	415 (12.04)	833 (24.17)	2198 (63.78)	51 (12.3)	92 (22.1)	273 (65.6)
TS8 (training: n=3476; testing: n=386)	1724 (49.6)	1752 (50.4)	191 (49.5)	195 (50.5)	423 (12.17)	826 (23.76)	2227 (64.07)	43 (11.14)	99 (25.65)	244 (63.21)
TS9 (training: n=3424; testing: n=428)	1702 (49.71)	1722 (50.29)	213 (48.6)	225 (51.34)	409 (11.95)	817 (23.86)	2198 (64.19)	57 (13.01)	108 (24.66)	273 (62.33)
TS10 (training: n=3353; testing: n=509)	1661 (49.54)	1692 (50.46)	254 (49.9)	255 (50.1)	424 (12.65)	791 (23.59)	2138 (63.76)	42 (8.25)	134 (26.33)	333 (65.42)

Table 6. Data information in patient characteristics for special care unit (SCU) of the cohort across the data splits.

TS[a]	Training		Testing	
	In SCU file, n (%)	Not in SCU file, n (%)	In SCU file, n (%)	Not in SCU file, n (%)
TS1 (training: n=3541; testing: n=321)	291 (8.22)	3250 (91.78)	27 (8.4)	294 (91.6)
TS2 (training: n=3531; testing: n=331)	292 (8.27)	3239 (91.73)	26 (7.8)	305 (92.1)
TS3 (training: n=3494; testing: n=368)	289 (8.27)	3205 (91.73)	29 (7.9)	339 (92.1)
TS4 (training: n=3488; testing: n=374)	285 (8.17)	3203 (91.83)	33 (8.8)	341 (91.2)
TS5 (training: n=3526; testing: n=336)	290 (8.22)	3236 (91.78)	28 (8.3)	308 (91.7)
TS6 (training: n=3479; testing: n=383)	282 (8.11)	3197 (91.89)	36 (9.4)	347 (90.6)
TS7 (training: n=3446; testing: n=416)	286 (8.3)	3160 (91.7)	32 (7.7)	384 (92.3)
TS8 (training: n=3476; testing: n=386)	282 (8.11)	3194 (91.89)	36 (9.3)	350 (90.7)
TS9 (training: n=3424; testing: n=428)	282 (8.24)	3142 (91.76)	36 (8.2)	402 (91.8)
TS10 (training: n=3353; testing: n=509)	283 (8.44)	3070 (91.56)	35 (6.9)	474 (93.1)

[a]TS: time segment.

Figure 2. The performances of 2 schemes changing over the 10 time segments (TSs) are shown using the gradient boosting classifier, where TS1 to TS10 are as follows: April 1, 2010, to September 30, 2010; October 1, 2010, to March 31, 2011; April 1, 2011, to September 30, 2011; October 1, 2011, to March 31, 2012; April 1, 2012, to September 30, 2012; October 31, 2012, to March 31, 2013; April 1, 2013, to September 30, 2013; October 1, 2013, to March 31, 2014; April 1, 2014, to September 30, 2014; and October 1, 2014, to March 31, 2015. NLP: natural language processing; ROC-AUC: area under the receiver operating characteristic curve.

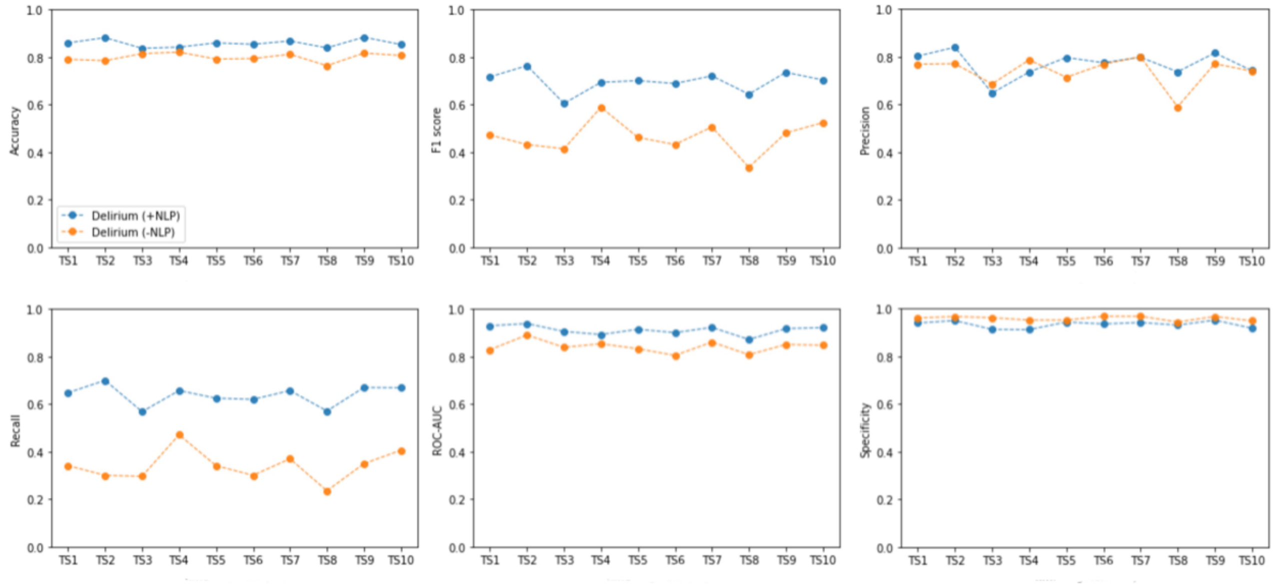
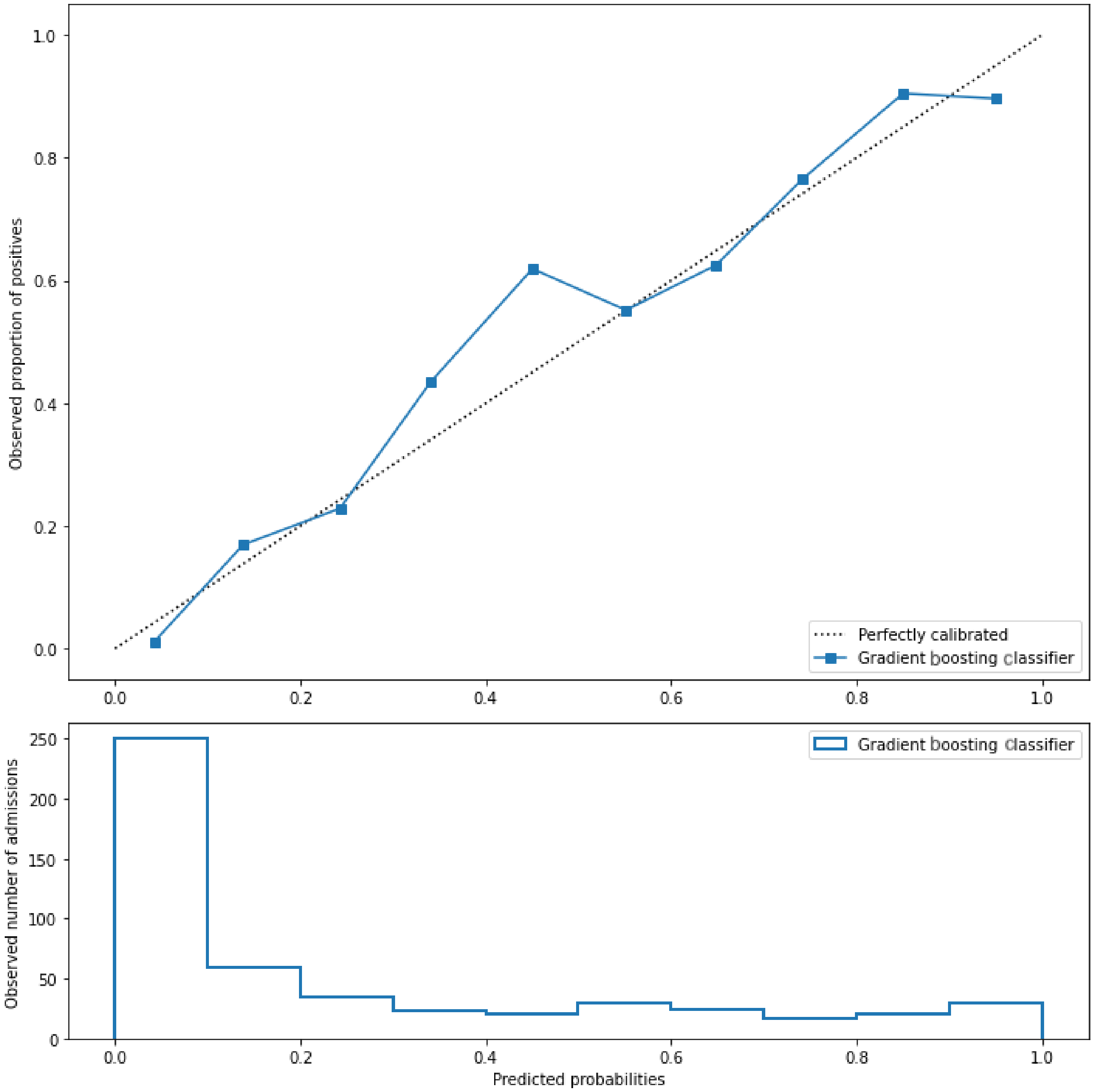


Figure 3. The calibration plot of the gradient boosting classifier.



As with the results for the last 6-month time segment, the delirium (+NLP) model also performed best using data from each of the earlier 9 time segments as the holdout set. The delirium (+NLP) model outperformed the delirium (–NLP) model in terms of accuracy, precision, recall or sensitivity, miss rate, ROC-AUC, and F_1-score.

Discussion

Principal Findings

Overall, machine learning models incorporating NLP either outperformed or were competitive with models that did not incorporate NLP for predicting the presence of delirium. Performance of the delirium (+NLP) model was relatively weaker on the specificity metric, but that metric was highly variable across the different holdout sets suggesting that it is a less reliable measure of performance in this application. As shown in the recall measure, the delirium (+NLP) model was better at detecting true positives, that is, identifying delirium for the admissions or patients who had ground truth delirium labels. The delirium (+NLP) model also performed best out of the 4 schemes in terms of having consistently high performance in terms of sensitivity, F_1-score (balancing sensitivity and precision), and ROC-AUC.

Prior risk identification models for delirium have tended to use a limited set of machine learning methods [7,29-33] and have tended to neglect text data [34]. In addition, most machine learning identification models to identify delirium only evaluate via simple partition of data (randomly partitioned 80%/20% for training and validating the classification model, respectively) or cross-validation [30,32,33]. In contrast, we used independent holdout or testing data (cross-validation in training data and totally separate testing data over time segments on the rolling basis, as shown in Figure 1), providing more rigorous testing of the identification model.

Previous research has found that routine clinical screening, using tools such as CAM, underreports up to 75% of delirium cases compared with clinical assessments for research [61-64]. Although we were not able to directly compare our model's performance with CAM results on the same patients, it is well documented in the literature that routine clinical use of CAM is unreliable for research or quality measurement, reinforcing the need for a model such as the one we developed in this study. Notably, the Montreal Cognitive Assessment is primarily used for the assessment of stable cognitive impairment and not for delirium.

The delirium (+NLP) model provided the best balance between recognizing cases of delirium, where they existed, and not mislabeling nondelirium cases as delirium. The baseline delirium scheme performed better when detecting true negatives. This is likely because our GEMINI data set was unbalanced, with 75% of admissions being nondelirium; thus, a poorly tuned model can achieve better accuracy by being biased toward predicting nondelirium.

One way of dealing with the trade-off between precision and recall is to use the F_1-score, which is the harmonic mean (average) of the precision and sensitivity or recall scores. With this more balanced measure, our proposed delirium (+NLP) model outperformed the one without NLP across all time segments.

Our delirium (+NLP) method integrated an NLP derived feature into multisource medical data to improve the performance and usefulness of models. This approach can also be extended to other medical identification contexts.

This approach has several important applications, including for quality measurement and quality improvement, for statistical risk adjustment in research projects, and for large-scale observational research in retrospective cohorts. There is currently no scalable solution to retrospectively identify the occurrence of delirium in hospital, and CAM is underutilized, perhaps because of the lack of trained clinical resources. We agree that prospective predictions of delirium would be clinically useful, and research on that topic is underway. However, retrospective prediction is also important for quality management purposes and for evaluating the effectiveness of interventions for preventing delirium. Typically, CAM is poorly implemented and used infrequently [23].

One major reason why delirium is underidentified in routine data sources is because it is often inconsistently documented, with the use of various synonyms (eg, confusion and altered level of consciousness). The only validated, high-quality method for retrospectively identifying delirium is the Chart-based Delirium Identification Instrument review method that we used as the gold standard labeling method for training our machine learning models. This method is time intensive and requires up to 1 hour per hospital chart. Thus, it cannot be easily applied to large data sets. Therefore, developing models that can use routinely collected clinical and administrative health care data represents a major contribution to the literature, as they can enable both research and quality applications that rely on retrospective identification of delirium cases.

It would be desirable to build models that could predict delirium risk at the time of hospitalization or in real time during the course of hospital admission. One impediment to developing these models is having sufficiently large data sets on which to train them. Our models, which seek to accurately classify hospitalizations with or without delirium retrospectively could then be used to label (using model predictions) large data sets, which could then be used to generate quality estimates and provide a basis for further model prediction.

Conclusions

Delirium is a highly prevalent, preventable, and treatable neurocognitive disorder, which is associated with very poor outcomes when untreated. It is characterized by an acute onset of fluctuating mental status, psychomotor disturbance, and hallucinations, and it is difficult to spot because the symptoms can often be attributed to other causes. Better delirium prediction will create an opportunity for higher quality care through automated identification of delirium or of delirium risk. In the research reported in this paper, we have shown that incorporation of the NLP approach can significantly improve identification compared with the standard machine learning methods without NLP. We also showed that varying the holdout period over time can estimate the temporal stability of model identification. Another useful feature of this type of stationarity analysis is that it can be used to identify unreliable evaluative criteria that exhibit nonstationarity and to identify models that are nonstationary with respect to their effectiveness over time. In this study, we found that precision was an unreliable criterion, with wide fluctuations over different periods.

The results of this study demonstrate the value of NLP in the identification of an important health care outcome, and we recommend that future research should focus on (1) applying NLP on medical notes to extract more valuable information and (2) augmenting the delirium (+NLP) model by adding explanations so that the resulting models are more consumable and more easily integrated into clinical workflow.

Acknowledgments

The authors would like to thank the Canadian Institutes of Health Research Foundation and the National Science and Engineering Research Council for funding this work through a Collaborative Health Research Projects grant (application number 415033).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Comparison of models with average training results using 5-fold cross-validation on training set (April 1, 2010, to September 30, 2014) in the other 9 algorithms: neural network, decision tree, logistic regression, linear support vector machine, Gaussian Naive Bayes, linear discriminant analysis, quadratic discriminant analysis, and voting classifier.
[DOCX File , 23 KB - medinform_v10i12e38161_app1.docx]

Multimedia Appendix 2
Comparison of 3 types of models in the other 9 algorithms: model performance on holdout set 10 (October 1, 2014, to March 31, 2015).
[DOCX File , 23 KB - medinform_v10i12e38161_app2.docx]

References

1. Maldonado JR. Acute brain failure: pathophysiology, diagnosis, management, and sequelae of delirium. Crit Care Clin 2017 Jul;33(3):461-519. [doi: 10.1016/j.ccc.2017.03.013] [Medline: 28601132]
2. Han JH, Wilson A, Ely EW. Delirium in the older emergency department patient: a quiet epidemic. Emerg Med Clin North Am 2010 Aug;28(3):611-631 [FREE Full text] [doi: 10.1016/j.emc.2010.03.005] [Medline: 20709246]
3. Maldonado JR. Delirium pathophysiology: an updated hypothesis of the etiology of acute brain failure. Int J Geriatr Psychiatry 2018 Nov;33(11):1428-1457. [doi: 10.1002/gps.4823] [Medline: 29278283]
4. Verma AA, Masoom H, Rawal S, Guo Y, Razak F, GEMINI Investigators. Pulmonary embolism and deep venous thrombosis in patients hospitalized with syncope: a multicenter cross-sectional study in Toronto, Ontario, Canada. JAMA Intern Med 2017 Jul 01;177(7):1046-1048 [FREE Full text] [doi: 10.1001/jamainternmed.2017.1246] [Medline: 28492876]
5. Conn DK, Gibson M. Guidelines for the assessment and treatment of mental health issues. In: Conn DK, Herrmann N, Kaye A, Rewilak D, Schogt B, editors. Practical Psychiatry in the Long-Term Care Home: A Handbook for Staff. 3rd revised and expanded edition. Göttingen, Germany: Hogrefe and Huber Publishers; 2007:267-278.
6. Gage L, Hogan DB. 2014 CCSMH Guideline Update: The Assessment and Treatment of Delirium. Canadian Coalition for Seniors' Mental Health. Toronto, Canada: Canadian Coalition for Seniors' Mental Health; 2014. URL: https://ccsmh.ca/wp-content/uploads/2016/03/2014-ccsmh-Guideline-Update-Delirium.pdf [accessed 2022-12-07]
7. Wong K, Tsang A, Liu B, Schwartz R. The Ontario Senior Friendly Hospital Strategy: Delirium and Functional Decline Indicators - A Report of the Senior Friendly Hospital Indicators Working Group. Local Health Integration Networks of Ontario. 2012 Nov. URL: https://www.rgptoronto.ca/wp-content/uploads/2017/12/SFH_Delirium_and_Functional_Decline_Indicators.pdf [accessed 2022-12-07]
8. Australian Commission on Safety and Quality in Health Care. 2012. URL: https://www.safetyandquality.gov.au/ [accessed 2022-12-07]
9. Breitbart W, Gibson C, Tremblay A. The delirium experience: delirium recall and delirium-related distress in hospitalized patients with cancer, their spouses/caregivers, and their nurses. Psychosomatics 2002;43(3):183-194. [doi: 10.1176/appi.psy.43.3.183] [Medline: 12075033]
10. Bruera E, Bush SH, Willey J, Paraskevopoulos T, Li Z, Palmer JL, et al. Impact of delirium and recall on the level of distress in patients with advanced cancer and their family caregivers. Cancer 2009 May 01;115(9):2004-2012 [FREE Full text] [doi: 10.1002/cncr.24215] [Medline: 19241420]
11. Inouye SK. Delirium in older persons. N Engl J Med 2006 Mar 16;354(11):1157-1165 [FREE Full text] [doi: 10.1056/NEJMra052321] [Medline: 16540616]
12. McCusker J, Cole M, Abrahamowicz M, Primeau F, Belzile E. Delirium predicts 12-month mortality. Arch Intern Med 2002 Feb 25;162(4):457-463. [doi: 10.1001/archinte.162.4.457] [Medline: 11863480]
13. Salluh JI, Wang H, Schneider EB, Nagaraja N, Yenokyan G, Damluji A, et al. Outcome of delirium in critically ill patients: systematic review and meta-analysis. BMJ 2015 Jun 03;350:h2538 [FREE Full text] [doi: 10.1136/bmj.h2538] [Medline: 26041151]
14. Yaffe K, Weston A, Graff-Radford NR, Satterfield S, Simonsick EM, Younkin SG, et al. Association of plasma beta-amyloid level and cognitive reserve with subsequent cognitive decline. JAMA 2011 Jan 19;305(3):261-266 [FREE Full text] [doi: 10.1001/jama.2010.1995] [Medline: 21245181]
15. MacLullich AM, Beaglehole A, Hall RJ, Meagher DJ. Delirium and long-term cognitive impairment. Int Rev Psychiatry 2009 Feb;21(1):30-42. [doi: 10.1080/09540260802675031] [Medline: 19219711]
16. Fong TG, Davis D, Growdon ME, Albuquerque A, Inouye SK. The interface between delirium and dementia in elderly adults. Lancet Neurol 2015 Aug;14(8):823-832 [FREE Full text] [doi: 10.1016/S1474-4422(15)00101-5] [Medline: 26139023]
17. Rockwood K, Cosway S, Carver D, Jarrett P, Stadnyk K, Fisk J. The risk of dementia and death after delirium. Age Ageing 1999 Oct;28(6):551-556. [doi: 10.1093/ageing/28.6.551] [Medline: 10604507]

18. Leslie DL, Marcantonio ER, Zhang Y, Leo-Summers L, Inouye SK. One-year health care costs associated with delirium in the elderly population. Arch Intern Med 2008 Jan 14;168(1):27-32 [FREE Full text] [doi: 10.1001/archinternmed.2007.4] [Medline: 18195192]
19. Hshieh TT, Yang T, Gartaganis SL, Yue J, Inouye SK. Hospital elder life program: systematic review and meta-analysis of effectiveness. Am J Geriatr Psychiatry 2018 Oct;26(10):1015-1033 [FREE Full text] [doi: 10.1016/j.jagp.2018.06.007] [Medline: 30076080]
20. Inouye SK, Bogardus Jr ST, Charpentier PA, Leo-Summers L, Acampora D, Holford TR, et al. A multicomponent intervention to prevent delirium in hospitalized older patients. N Engl J Med 1999 Mar 04;340(9):669-676. [doi: 10.1056/NEJM199903043400901] [Medline: 10053175]
21. Teodorczuk A, Reynish E, Milisen K. Improving recognition of delirium in clinical practice: a call for action. BMC Geriatr 2012 Sep 14;12:55 [FREE Full text] [doi: 10.1186/1471-2318-12-55] [Medline: 22974329]
22. Lewis LM, Miller DK, Morley JE, Nork MJ, Lasater LC. Unrecognized delirium in ED geriatric patients. Am J Emerg Med 1995 Mar;13(2):142-145. [doi: 10.1016/0735-6757(95)90080-2] [Medline: 7893295]
23. Hogan TM, Olade TO, Carpenter CR. A profile of acute care in an aging America: snowball sample identification and characterization of United States geriatric emergency departments in 2013. Acad Emerg Med 2014 Mar;21(3):337-346 [FREE Full text] [doi: 10.1111/acem.12332] [Medline: 24628759]
24. McCoy Jr TH, Snapper L, Stern TA, Perlis RH. Underreporting of delirium in statewide claims data: implications for clinical care and predictive modeling. Psychosomatics 2016;57(5):480-488. [doi: 10.1016/j.psym.2016.06.001] [Medline: 27480944]
25. Lindroth H, Bratzke L, Purvis S, Brown R, Coburn M, Mrkobrada M, et al. Systematic review of prediction models for delirium in the older adult inpatient. BMJ Open 2018 Apr 28;8(4):e019223 [FREE Full text] [doi: 10.1136/bmjopen-2017-019223] [Medline: 29705752]
26. Pendlebury ST, Lovett NG, Smith SC, Wharton R, Rothwell PM. Delirium risk stratification in consecutive unselected admissions to acute medicine: validation of a susceptibility score based on factors identified externally in pooled data for use at entry to the acute care pathway. Age Ageing 2017 Mar 01;46(2):226-231 [FREE Full text] [doi: 10.1093/ageing/afw198] [Medline: 27816908]
27. Rudolph JL, Doherty K, Kelly B, Driver JA, Archambault E. Validation of a delirium risk assessment using electronic medical record information. J Am Med Dir Assoc 2016 Mar 01;17(3):244-248. [doi: 10.1016/j.jamda.2015.10.020] [Medline: 26705000]
28. Rudolph JL, Harrington MB, Lucatorto MA, Chester JG, Francis J, Shay KJ, Veterans Affairs and Delirium Working Group. Validation of a medical record-based delirium risk assessment. J Am Geriatr Soc 2011 Nov;59 Suppl 2(Suppl 2):S289-S294 [FREE Full text] [doi: 10.1111/j.1532-5415.2011.03677.x] [Medline: 22091575]
29. Naylor CD. On the prospects for a (deep) learning health care system. JAMA 2018 Sep 18;320(11):1099-1100. [doi: 10.1001/jama.2018.11103] [Medline: 30178068]
30. Jauk S, Kramer D, Großauer B, Rienmüller S, Avian A, Berghold A, et al. Risk prediction of delirium in hospitalized patients using machine learning: an implementation and prospective evaluation study. J Am Med Inform Assoc 2020 Jul 01;27(9):1383-1392 [FREE Full text] [doi: 10.1093/jamia/ocaa113] [Medline: 32968811]
31. Buenviaje B, Bischoff JE, Roncace RA, Willy CJ. Mahalanobis-Taguchi system to identify preindicators of delirium in the ICU. IEEE J Biomed Health Inform 2016 Jul;20(4):1205-1212. [doi: 10.1109/JBHI.2015.2434949] [Medline: 26011872]
32. Corradi JP, Thompson S, Mather JF, Waszynski CM, Dicks RS. Prediction of incident delirium using a random forest classifier. J Med Syst 2018 Nov 14;42(12):261. [doi: 10.1007/s10916-018-1109-0] [Medline: 30430256]
33. Oh J, Cho D, Park J, Na SH, Kim J, Heo J, et al. Prediction and early detection of delirium in the intensive care unit by using heart rate variability and machine learning. Physiol Meas 2018 Mar 27;39(3):035004. [doi: 10.1088/1361-6579/aaab07] [Medline: 29376502]
34. Hercus C, Hudaib AR. Delirium misdiagnosis risk in psychiatry: a machine learning-logistic regression predictive algorithm. BMC Health Serv Res 2020 Feb 27;20(1):151 [FREE Full text] [doi: 10.1186/s12913-020-5005-1] [Medline: 32106845]
35. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2014;2:3 [FREE Full text] [doi: 10.1186/2047-2501-2-3] [Medline: 25825667]
36. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
37. Hinton G. Deep learning-a technology with the potential to transform health care. JAMA 2018 Sep 18;320(11):1101-1102. [doi: 10.1001/jama.2018.11100] [Medline: 30178065]
38. Saeed M, Lieu C, Raber G, Mark RG. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. Comput Cardiol 2002;29:641-644. [Medline: 14686455]
39. Loper E, Bird S. Nltk: the natural language toolkit. arXiv 2002 May 17. [doi: 10.3115/1118108.1118117]
40. Ridgway JP, Uvin A, Schmitt J, Oliwa T, Almirol E, Devlin S, et al. Natural language processing of clinical notes to identify mental illness and substance use among people living with HIV: retrospective cohort study. JMIR Med Inform 2021 Mar 10;9(3):e23456 [FREE Full text] [doi: 10.2196/23456] [Medline: 33688848]

41. Wu H, Hodgson K, Dyson S, Morley KI, Ibrahim ZM, Iqbal E, et al. Efficient reuse of natural language processing models for phenotype-mention identification in free-text electronic medical records: a phenotype embedding approach. JMIR Med Inform 2019 Dec 17;7(4):e14782 [FREE Full text] [doi: 10.2196/14782] [Medline: 31845899]
42. Geng W, Qin X, Yang T, Cong Z, Wang Z, Kong Q, et al. Model-based reasoning of clinical diagnosis in integrative medicine: real-world methodological study of electronic medical records and natural language processing methods. JMIR Med Inform 2020 Dec 21;8(12):e23082 [FREE Full text] [doi: 10.2196/23082] [Medline: 33346740]
43. Nakatani H, Nakao M, Uchiyama H, Toyoshiba H, Ochiai C. Predicting inpatient falls using natural language processing of nursing records obtained from Japanese electronic medical records: case-control study. JMIR Med Inform 2020 Apr 22;8(4):e16970 [FREE Full text] [doi: 10.2196/16970] [Medline: 32319959]
44. 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]
45. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform 2019 Apr 27;7(2):e12239 [FREE Full text] [doi: 10.2196/12239] [Medline: 31066697]
46. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ 2015 Apr 24;350:h1885 [FREE Full text] [doi: 10.1136/bmj.h1885] [Medline: 25911572]
47. 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]
48. Devika MD, Sunitha C, Ganesh A. Sentiment analysis: a comparative study on different approaches. Procedia Comput Sci 2016;87:44-49. [doi: 10.1016/j.procs.2016.05.124]
49. Verma AA, Guo Y, Kwan JL, Lapointe-Shaw L, Rawal S, Tang T, et al. Prevalence and costs of discharge diagnoses in inpatient general internal medicine: a multi-center cross-sectional study. J Gen Intern Med 2018 Nov;33(11):1899-1904 [FREE Full text] [doi: 10.1007/s11606-018-4591-7] [Medline: 30054888]
50. Verma AA, Pasricha SV, Jung HY, Kushnir V, Mak DY, Koppula R, et al. Assessing the quality of clinical and administrative data extracted from hospitals: the General Medicine Inpatient Initiative (GEMINI) experience. J Am Med Inform Assoc 2021 Mar 01;28(3):578-587 [FREE Full text] [doi: 10.1093/jamia/ocaa225] [Medline: 33164061]
51. Wang L, Chignell M, Zhang Y, Pinto A, Razak F, Sheehan K, et al. Physician experience design (PXD): more usable machine learning prediction for clinical decision making. AMIA Annu Symp Proc 2022 May 23;2022:476-485 [FREE Full text] [Medline: 35854747]
52. Verma AA, Guo Y, Kwan JL, Lapointe-Shaw L, Rawal S, Tang T, et al. Patient characteristics, resource use and outcomes associated with general internal medicine hospital care: the General Medicine Inpatient Initiative (GEMINI) retrospective cohort study. CMAJ Open 2017 Dec 11;5(4):E842-E849 [FREE Full text] [doi: 10.9778/cmajo.20170097] [Medline: 29237706]
53. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Use of sentiment analysis for capturing patient experience from free-text comments posted online. J Med Internet Res 2013 Nov 01;15(11):e239 [FREE Full text] [doi: 10.2196/jmir.2721] [Medline: 24184993]
54. Inouye SK, Leo-Summers L, Zhang Y, Bogardus ST, Leslie DL, Agostini JV. A chart-based method for identification of delirium: validation compared with interviewer ratings using the confusion assessment method. J Am Geriatr Soc 2005 Feb;53(2):312-318. [doi: 10.1111/j.1532-5415.2005.53120.x] [Medline: 15673358]
55. Petersen CL, Halter R, Kotz D, Loeb L, Cook S, Pidgeon D, et al. Using natural language processing and sentiment analysis to augment traditional user-centered design: development and usability study. JMIR Mhealth Uhealth 2020 Aug 07;8(8):e16862 [FREE Full text] [doi: 10.2196/16862] [Medline: 32540843]
56. 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]
57. Escobar GJ, Greene JD, Scheirer P, Gardner MN, Draper D, Kipnis P. Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. Med Care 2008 Mar;46(3):232-239. [doi: 10.1097/MLR.0b013e3181589bb6] [Medline: 18388836]
58. Khwaja A. KDIGO clinical practice guidelines for acute kidney injury. Nephron Clin Pract 2012;120(4):c179-c184 [FREE Full text] [doi: 10.1159/000339789] [Medline: 22890468]
59. Bullard MJ, Chan T, Brayman C, Warren D, Musgrave E, Unger B, Members of the CTAS National Working Group. Revisions to the Canadian emergency department triage and acuity scale (CTAS) guidelines. CJEM 2014 Nov;16(6):485-489. [Medline: 25358280]
60. 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.

61. Loftus CA, Wiesenfeld LA. Geriatric delirium care: using chart audits to target improvement strategies. Can Geriatr J 2017 Dec;20(4):246-252 [FREE Full text] [doi: 10.5770/cgj.20.276] [Medline: 29296131]
62. Solberg LM, Plummer CE, May KN, Mion LC. A quality improvement program to increase nurses’ detection of delirium on an acute medical unit. Geriatr Nurs 2013;34(1):75-79 [FREE Full text] [doi: 10.1016/j.gerinurse.2012.12.009] [Medline: 23614146]
63. Rice KL, Bennett M, Gomez M, Theall KP, Knight M, Foreman MD. Nurses' recognition of delirium in the hospitalized older adult. Clin Nurse Spec 2011;25(6):299-311. [doi: 10.1097/NUR.0b013e318234897b] [Medline: 22016018]
64. Lemiengre J, Nelis T, Joosten E, Braes T, Foreman M, Gastmans C, et al. Detection of delirium by bedside nurses using the confusion assessment method. J Am Geriatr Soc 2006 Apr;54(4):685-689. [doi: 10.1111/j.1532-5415.2006.00667.x] [Medline: 16686883]

Abbreviations

CAM: Confusion Assessment Method
CCS: Clinical Classification Software
GEMINI: General Medicine Inpatient Initiative
ICD-10: International Classification of Diseases, Tenth Revision
NLP: natural language processing
REB: research ethics board
ROC-AUC: area under the receiver operating characteristic curve
SVM: support vector machine

Edited by T Hao; submitted 21.03.22; peer-reviewed by M Afshar, F Carini; comments to author 27.06.22; revised version received 22.08.22; accepted 19.09.22; published 20.12.22.

Please cite as:
Wang L, Zhang Y, Chignell M, Shan B, Sheehan KA, Razak F, Verma A
Boosting Delirium Identification Accuracy With Sentiment-Based Natural Language Processing: Mixed Methods Study
JMIR Med Inform 2022;10(12):e38161
URL: https://medinform.jmir.org/2022/12/e38161
doi:10.2196/38161
PMID:36538363

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org