JMIR Medical Informatics

Impact Factor (2023): 3.1
Volume 9 (2021), Issue 5   ISSN 2291-9694   Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Viewpoint

A Roadmap for Automating Lineage Tracing to Aid Automatically Explaining Machine Learning Predictions for Clinical Decision Support (e27778)
Gang Luo. . . . 4

Original Papers

Anomaly Detection Algorithm for Real-World Data and Evidence in Clinical Research: Implementation, Evaluation, and Validation Study (e27172)
Vendula Churová, Roman Vyškovský, Kate ina Maršálová, David Kudlá ek, Daniel Schwarz. . . . 21

An Algorithm (LaD) for Monitoring Childbirth in Settings Where Tracking All Parameters in the World Health Organization Partograph Is Not Feasible: Design and Expert Validation (e17056)
Michael Balikuddembe, Peter Wakholi, Nazarius Tumwesigye, Thorkild Tylleskar. . . . 38

Automating Stroke Data Extraction From Free-Text Radiology Reports Using Natural Language Processing: Instrument Validation Study (e24381)
Amy Yu, Zhongyu Liu, Chloe Pou-Prom, Kaitlyn Lopes, Moira Kapral, Richard Aviv, Muhammad Mamdani. . . . 49

Extracting Drug Names and Associated Attributes From Discharge Summaries: Text Mining Study (e24678)
Ghada Alfattni, Maksim Belousov, Niels Peek, Goran Nenadic. . . . 57

An Attention Model With Transfer Embeddings to Classify Pneumonia-Related Bilingual Imaging Reports: Algorithm Development and Validation (e24803)
Hyung Park, Min Song, Eun Lee, Bo Seo, Chang Choi. . . . 74

Combining External Medical Knowledge for Improving Obstetric Intelligent Diagnosis: Model Development and Validation (e25304)
Kunli Zhang, Linkun Cai, Yu Song, Tao Liu, Yueshu Zhao. . . . 84

A Multimodal Imaging–Based Deep Learning Model for Detecting Treatment-Requiring Retinal Vascular Diseases: Model Development and Validation Study (e28868)
Eugene Kang, Ling Yeung, Yi-Lun Lee, Cheng-Hsiu Wu, Shu-Yen Peng, Yueh-Peng Chen, Quan-Ze Gao, Chihung Lin, Chang-Fu Kuo, Chi-Chun Lai. . . . 96

Pathway-Driven Coordinated Telehealth System for Management of Patients With Single or Multiple Chronic Diseases in China: System Development and Retrospective Study (e27228)

Zheyu Wang, Jiye An, Hui Lin, Jiaqiang Zhou, Fang Liu, Juan Chen, Huilong Duan, Ning Deng. 109

Telemedicine for Follow-up Management of Patients After Liver Transplantation: Cohort Study (e27175)

Min Tian, Bo Wang, Zhao Xue, Dinghui Dong, Xuemin Liu, Rongqian Wu, Liang Yu, Junxi Xiang, Xiaogang Zhang, Xufeng Zhang, Yi Lv. . . .
1 3 5

Neural Network–Based Retinal Nerve Fiber Layer Profile Compensation for Glaucoma Diagnosis in Myopia: Model Development and Validation (e22664)

Lei Li, Haogang Zhu, Zhenyu Zhang, Liang Zhao, Liang Xu, Rahul Jonas, David Garway-Heath, Jost Jonas, Ya Wang. 147

Transforming a Patient Registry Into a Customized Data Set for the Advanced Statistical Analysis of Health Risk Factors and for Medication-Related Hospitalization Research: Retrospective Hospital Patient Registry Study (e24205)

Zhivko Taushanov, Henk Verloo, Boris Wernli, Saviana Di Giovanni, Armin von Gunten, Filipa Pereira. 160

Predicting Intensive Care Unit Length of Stay and Mortality Using Patient Vital Signs: Machine Learning Model Development and Validation (e21347)

Khalid Alghatani, Nariman Ammar, Abdelmounaam Rezgui, Arash Shaban-Nejad. 185

Automated Generation of Personalized Shock Wave Lithotripsy Protocols: Treatment Planning Using Deep Learning (e24721)

Zhipeng Chen, Daniel Zeng, Ryan Seltzer, Blake Hamilton. 208

Deciphering the Efficacy and Mechanisms of Chinese Herbal Medicine for Diabetic Kidney Disease by Integrating Web-Based Biochemical Databases and Real-World Clinical Data: Retrospective Cohort Study (e27614)

Chien-Wei Wu, Hsing-Yu Chen, Ching-Wei Yang, Yu-Chun Chen. 219

Federated Learning for Thyroid Ultrasound Image Analysis to Protect Personal Information: Validation Study in a Real Health Care Environment (e25869)

Haeyun Lee, Young Chai, Hyunjin Joo, Kyungsu Lee, Jae Hwang, Seok-Mo Kim, Kwangsoon Kim, Inn-Chul Nam, June Choi, Hyeong Yu, Myung-Chul Lee, Hiroo Masuoka, Akira Miyauchi, Kyu Lee, Sungwan Kim, Hyoun-Joong Kong. 239

Predicting Prolonged Length of Hospital Stay for Peritoneal Dialysis–Treated Patients Using Stacked Generalization: Model Development and Validation Study (e17886)

Guilan Kong, Jingyi Wu, Hong Chu, Chao Yang, Yu Lin, Ke Lin, Ying Shi, Haibo Wang, Luxia Zhang. 248

Automatically Diagnosing Disk Bulge and Disk Herniation With Lumbar Magnetic Resonance Images by Using Deep Convolutional Neural Networks: Method Development Study (e14755)

Qiong Pan, Kai Zhang, Lin He, Zhou Dong, Lei Zhang, Xiaohang Wu, Yi Wu, Yanjun Gao. 264

Improving Current Glycated Hemoglobin Prediction in Adults: Use of Machine Learning Algorithms With Electronic Health Records (e25237)

Zakhriya Alhassan, Matthew Watson, David Budgen, Riyad Alshammari, Ali Alessa, Noura Al Moubayed. 274

Effective Training Data Extraction Method to Improve Influenza Outbreak Prediction from Online News Articles: Deep Learning Model Study (e23305)

Beakcheol Jang, Inhwan Kim, Jong Kim. 291

Leveraging Genetic Reports and Electronic Health Records for the Prediction of Primary Cancers: Algorithm Development and Validation Study (e23586)

Nansu Zong, Victoria Ngo, Daniel Stone, Andrew Wen, Yiqing Zhao, Yue Yu, Sijia Liu, Ming Huang, Chen Wang, Guoqian Jiang. 305

Predicting Semantic Similarity Between Clinical Sentence Pairs Using Transformer Models: Evaluation and Representational Analysis (e23099)
Mark Ormerod, Jesús Martínez del Rincón, Barry Devereux. 326

Application of Intelligent Computer-Assisted Taylor 3D External Fixation in the Treatment of Tibiofibular Fracture: Retrospective Case Study (e21455)
Hongfeng Sheng, Weixing Xu, Bin Xu, Hongpu Song, Di Lu, Weiguo Ding, Henry Mildredl. 343

Evaluation of the Privacy Risks of Personal Health Identifiers and Quasi-Identifiers in a Distributed Research Network: Development and Validation Study (e24940)
SeHee Oh, MinDong Sung, Yumie Rhee, Namki Hong, Yu Park. 353

Use of Machine Learning Algorithms to Predict the Understandability of Health Education Materials: Development and Evaluation Study (e28413)
Meng Ji, Yanmeng Liu, Mengdan Zhao, Ziqing Lyu, Boren Zhang, Xin Luo, Yanlin Li, Yin Zhong. 370

Corrigenda and Addendas

Correction: Extracting Family History Information From Electronic Health Records: Natural Language Processing Analysis (e30153)
Maciej Rybinski, Xiang Dai, Sonit Singh, Sarvnaz Karimi, Anthony Nguyen. 323

Correction: A Patient Journey Map to Improve the Home Isolation Experience of Persons With Mild COVID-19: Design Research for Service Touchpoints of Artificial Intelligence in eHealth (e29794)
Qian He, Fei Du, Lianne Simonse. 324

Viewpoint

A Roadmap for Automating Lineage Tracing to Aid Automatically Explaining Machine Learning Predictions for Clinical Decision Support

Gang Luo[1], DPhil

[1] Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, United States

Corresponding Author:
Gang Luo, DPhil
Department of Biomedical Informatics and Medical Education
University of Washington
UW Medicine South Lake Union
850 Republican Street, Building C, Box 358047
Seattle, WA, 98195
United States
Phone: 1 206 221 4596
Fax: 1 206 221 2671
Email: gangluo@cs.wisc.edu

Abstract

Using machine learning predictive models for clinical decision support has great potential in improving patient outcomes and reducing health care costs. However, most machine learning models are black boxes that do not explain their predictions, thereby forming a barrier to clinical adoption. To overcome this barrier, an automated method was recently developed to provide rule-style explanations of any machine learning model's predictions on tabular data and to suggest customized interventions. Each explanation delineates the association between a feature value pattern and an outcome value. Although the association and intervention information is useful, the user of the automated explaining function often requires more detailed information to better understand the patient's situation and to aid in decision making. More specifically, consider a feature value in the explanation that is computed by an aggregation function on the raw data, such as the number of emergency department visits related to asthma that the patient had in the prior 12 months. The user often wants to rapidly drill through to see certain parts of the related raw data that produce the feature value. This task is frequently difficult and time-consuming because the few pieces of related raw data are submerged by many pieces of raw data of the patient that are unrelated to the feature value. To address this issue, this paper outlines an automated lineage tracing approach, which adds automated drill-through capability to the automated explaining function, and provides a roadmap for future research.

(JMIR Med Inform 2021;9(5):e27778) doi:10.2196/27778

KEYWORDS

clinical decision support; database management systems; forecasting; machine learning; electronic medical records

Introduction

Machine learning has won almost all data science competitions [1] and is a hot topic these days. It is about computer algorithms that automatically learn from data, such as extreme gradient boosting, support vector machine, and random forest [2]. Using machine learning predictive models for clinical decision support has great potential in improving patient outcomes and reducing health care costs [3-10]. However, most machine learning models are black boxes that do not explain their predictions. This creates a barrier to clinical adoption. To overcome this barrier, we recently developed an automated method to offer rule-style explanations of any machine learning model's predictions on tabular data and to suggest customized interventions without reducing the model's performance measures [11-14]. Each rule-style explanation delineates the association between a feature value pattern and an outcome value. A feature is also called an independent variable. For the prediction of future emergency department (ED) visits or inpatient stays for asthma for a patient with asthma, one example of the explanation is as follows:

- The patient had 2 ED visits related to asthma in the prior 12 months

> AND the patient's average respiratory rate recorded in the prior 12 months is >25 and ≤28 breaths per minute
> →the patient will likely have at least 1 ED visit or inpatient stay for asthma in the next 12 months [13,14].

An ED visit is related to asthma if the ED visit has an asthma diagnosis code. For the item in the explanation showing that the patient had 2 ED visits related to asthma in the prior 12 months, 1 intervention suggested by the automatic explanation method [12-14] is to apply control procedures that decrease the likelihood that the patient will need emergency care.

The association and intervention information provided by the automatic explanation method for machine learning predictions is useful. However, the user of the automated explaining function often requires more detailed information to better understand the patient's situation and to aid in decision making. More specifically, consider a feature value on the left-hand side of a rule-style explanation that is computed by an aggregation function on the raw data. The user often wants to rapidly drill through to see certain parts of the related raw data producing the feature value. In the context of a relational database, these parts refer to the most relevant attributes of the most essential source tuples producing the feature value. Which attributes are most relevant and which source tuples are most essential depend on both the concrete feature type and the clinical decision support application's need and are illustrated by several examples throughout this paper. The patterns embedded in these parts could provide additional information on the patient that was lost during the aggregation process to compute the feature value. This drill-through task is frequently difficult and time-consuming because the few pieces of related raw data are submerged by many pieces of raw data of the patient that are unrelated to the feature value. For example, as Table 1 shows, the list of encounters of a patient with asthma displayed on the standard interface of an electronic medical record system includes much information that is irrelevant to the feature value "2 of the number of ED visits related to asthma that the patient had in the prior 12 months."

Table 1. An example list of encounters of a patient with asthma displayed on the standard interface of an electronic medical record system.[a]

Visit date	Primary diagnosis[b]	Visit type	Department	Provider	Facility
Dec 20, 2020	Cough (R05)	Outpatient	HMC[c] family medicine clinic	John Smith	HMC
Dec 18, 2020	Dysphagia, unspecified (R13.10)	Outpatient	HMC family medicine clinic	David Wong	HMC
…	…	…	…	…	…
Oct 15, 2020	Cystitis, unspecified without hematuria (N30.90)	Inpatient	UWMC[d] 8SE	Leslie Hurdle	UWMC
Oct 12, 2020 [e]	*Viral infection, unspecified (B34.9)*	*Emergency*	*HMC HEDUCC* [f]	*Patricia Sward*	*HMC*
Oct 09, 2020	Dizziness and giddiness (R42)	Outpatient	HMC family medicine clinic	Eve Johnson	HMC
…	…	…	…	…	…
Feb 11, 2020	Posttraumatic stress disorder, unspecified (F43.10)	Outpatient	HMC psychotherapy clinic	Amy Jiang	HMC
Feb 08, 2020	*Syncope and collapse (R55)*	*Emergency*	*HMC HEDUCC*	*Peter Shavlik*	*HMC*
Feb 03, 2020	Headache, unspecified (R51.9)	Outpatient	HMC family medicine clinic	Jude Lake	HMC
…	…	…	…	…	…

[a] This example list is made based on a similar list seen in real electronic medical record data at the University of Washington Medicine.

[b] This column does not show up on the standard interface. This column is included because it will be discussed in this paper.

[c] HMC: Harborview Medical Center.

[d] UWMC: University of Washington Medical Center.

[e] For the feature value "2 of the number of emergency department visits related to asthma that the patient had in the prior 12 months," the related rows in the list producing the feature value are marked in italics.

[f] HEDUCC: Harborview Emergency Department Urgent Care Center.

For instance, in the rule-style explanation shown above, the first item on the left-hand side is the feature value "2 of the number of ED visits related to asthma that the patient had in the prior 12 months." Asthma may or may not be the primary diagnosis of either of these 2 visits. For this feature value, the user of the automated explaining function wants to see the relevant parts of these 2 visits (visit date, primary diagnosis, department handling the visit, admitting provider, facility where the visit occurred) in the reverse chronological order (see Table 2), like the way encounters are displayed on the standard interface of an electronic medical record system. The patterns embedded in these parts give additional information on the patient not shown by the feature value, such as the time between these 2 visits, how long ago these 2 visits occurred, the primary diagnoses in these 2 visits, and whether these 2 visits occurred at the same facility. However, finding these parts is nontrivial. As seen in real electronic medical record data at the University of Washington Medicine, Intermountain Healthcare, and Kaiser Permanente Southern California, the patient could have had over 100 encounters in the prior 12 months. Only a few of these encounters are ED visits, and even fewer of them are ED visits related to asthma. To find the ED visits of the patient in the prior 12 months, the user would need some manual effort even if aided by the search function for the electronic medical record system. To figure out which of these visits are related to asthma, a task with which the search function often cannot provide much help, the user would need much more manual effort.

Table 2. An example of the parts of the related raw data that should be displayed for a feature value.[a]

Visit date	Primary diagnosis	Department	Provider	Facility
Oct 12, 2020	Viral infection, unspecified (B34.9)	HMC[b] HEDUCC[c]	Patricia Sward	HMC
Feb 08, 2020	Syncope and collapse (R55)	HMC HEDUCC	Peter Shavlik	HMC

[a] For the example list shown in Table 1 and the feature value “2 of the number of emergency department visits related to asthma that the patient had in the prior 12 months,” the parts that the user of the automated explaining function wants to see are in the related raw data producing the feature value.

[b] HMC: Harborview Medical Center.

[c] HEDUCC: Harborview Emergency Department Urgent Care Center.

In practice, numerous possible features computed by various aggregation functions on all kinds of longitudinal attributes in the electronic medical records could be used for predictive modeling and automatic explanation. Examples of such features include whether the most recent asthma diagnosis of the patient is a primary diagnosis, the patient’s average respiratory rate recorded in the prior 12 months, the total number of distinct asthma medications ordered for the patient in the prior 12 months, the total number of units of asthma relievers that were ordered for the patient in the prior 12 months and were neither systemic corticosteroids nor short-acting beta-2 agonists, the number of distinct asthma medication prescribers of the patient in the prior 12 months, and the number of no-shows by the patient in the prior 12 months [13,14]. Most of the possible features are unanticipated by the developers of the search function for the electronic medical record system beforehand. The search function supports only a few fixed types of search. For only a small portion of possible features, the search function can aid drilling through the raw data that produce a given feature value.

This creates a problem for the widespread adoption of the automatic explanation method for machine learning predictions. Frequently, this method gives multiple rule-style explanations for a patient predicted to be at high risk of incurring a poor outcome [11,12]. The user of the automated explaining function is typically a busy clinician having no time to do laborious manual drill-through regularly. However, to better understand the patient’s situation and to make better clinical decisions, the user often wants to drill through multiple feature values of the patient appearing in the explanations. If done manually, this is a challenging task. A patient often has extensive records with numerous variables and hundreds of pages of content accumulated over a long period of time [15]. Further, the relevant raw data producing the feature values are frequently scattered in several places in the electronic medical record system.

This study makes 2 contributions toward solving this problem:

1. We articulate this problem for the first time in the literature. This is done in the “Introduction” section.
2. To address this problem, an automated lineage tracing approach is outlined to add automated drill-through capability to the automated explaining function. This is done in the “Outline of the proposed automated lineage tracing approach” section. Further, a roadmap for future research is provided in the “Directions for future research” section.

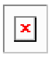
The automated drill-through capability is intended to be offered to help the user of the automated explaining function save time, better understand the patient’s situation, and make better clinical decisions. The discussion in this paper focuses on structured electronic medical record data, a specific method commonly used to build clinical machine learning predictive models, and the automatic explanation method for machine learning predictions [11,12]. Nevertheless, the automated lineage tracing approach is not limited to them. Instead, when automatically explaining machine learning predictions and after appropriate extension, the principle of this approach can be applied to facilitate drilling through any feature value computed by an aggregation function on longitudinal structured data, regardless of whether the data came from electronic medical records, whether the feature is specified by a human expert or semiautomatically extracted from longitudinal data using the method outlined in the prior paper [16], which method is used to build the machine learning predictive model, or which automatic explanation method is used.

Running Example

To illustrate this approach, a running example is used throughout this paper: automatically explaining the predictions of future ED visits or inpatient stays for individual patients with asthma. Our prior papers [12-14,17-19] detail this use case and the features used to make predictions in it.

Base Tables

Below are the schemas of 5 tables in a relational database used in the running example:



The underlined fields mark the key to each table. The *encounter* table includes 1 row per encounter listing its information. The *diagnosis* table includes 1 row per diagnosis code of an encounter. Primary diagnoses are signified by *dx_sequence_number*=1. The *diagnosis_code_master* table includes 1 row per unique diagnosis code giving its description. The *ordered_medication* table includes 1 row per medication appearing in a medication order. The *medication_master* table includes 1 row per unique medication listing its information.

Intermediate Result Tables

Besides the above 5 base tables, 4 intermediate result tables computed on the new data are also used in the running example: *enc_features_1*, *enc_features_2*, *enc_features_3*, and *med_features_1*. The trained machine learning predictive model is applied to the new data to make predictions on individual patients.

The intermediate result table *enc_features_1* contains 3 temporal features on encounters: the number of ED visits, the number of inpatient stays, and the number of outpatient visits that the patient had in the prior 12 months. Let *today_date* denote today's date. *enc_features_1* is computed from the *encounter* base table using the following structured query language (SQL) query.

The intermediate result table *enc_features_2* contains 1 temporal feature on encounters: the number of outpatient visits with a primary diagnosis of asthma that the patient had in the prior 12 months. Recall that the International Classification of Diseases, Tenth Revision diagnosis codes of asthma are J45.x. *enc_features_2* is computed by joining the *encounter* and *diagnosis* base tables using the following SQL query.

The intermediate result table *enc_features_3* contains 2 temporal features on encounters: the number of ED visits related to asthma and the number of inpatient stays related to asthma that the patient had in the prior 12 months. *enc_features_3* is computed by joining the *encounter* and *diagnosis* base tables using the following SQL query.

The intermediate result table *med_features_1* contains 2 temporal features on medications: the total number of medications and the total number of distinct medications ordered for the patient in the prior 12 months. *med_features_1* is computed from the *ordered_medication* base table using the following SQL query.

Relational Algebra Operators

This paper uses the following relational algebra operators with the bag semantics unless otherwise specified: join , left semijoin , selection σ, projection π, duplicate elimination δ, and grouping γ [20]. Commercial database management systems implement relations using the bag semantics.

Review of a Typical Method to Build a Clinical Machine Learning Predictive Model and Our Automated Method to Explain the Model's Predictions

In this section, a typical method to build a machine learning predictive model on structured electronic medical record data as well as the automated method to explain the model's predictions [11-14] are reviewed. In the next section, the automated lineage tracing approach based on these 2 methods is outlined.

A health care system usually has an enterprise data warehouse. It stores in a relational database a copy of the structured electronic medical record data of the health care system, often after some transformations such as pivoting [21,22] and denormalization to facilitate data analysis. For predictive modeling with automated explanation, the overall workflow is to execute database SQL queries to extract features from the electronic medical record data, to build a machine learning predictive model on the training data, to apply the model on new data to make predictions on individual patients, and then to use the automated method to explain the predictions. In the following sections, each of these steps is described sequentially.

Extracting Features From the Electronic Medical Record Data and Building the Clinical Machine Learning Predictive Model

The structured electronic medical record data contain both static attributes (eg, gender) and longitudinal attributes (eg, encounters, diagnoses). Most attributes are longitudinal. As Figure 1 shows, the following operations are performed on the training data:

1. The static features are computed from the static attribute values. The results are stored in 1 or more intermediate result tables. Typically, each of these intermediate result tables is computed by running a select-project-join SQL query on 1 or more base tables.
2. By aggregating longitudinal attribute values and sometimes also using some static attribute values, the patient cohort of interest in the training data is computed. The result is stored in 1 intermediate result table. This is typically done by running a complex SQL query on several base tables. An example patient cohort is the set of all patients with asthma who visited any of the facilities of the health care system during a specific time period.
3. By aggregating longitudinal attribute values, temporal features and the outcome variable are computed and stored in 1 or more intermediate result tables. Typically, each of these intermediate result tables is computed by running a select-project-join-aggregate SQL query on 1 or more base tables. For example, 1 intermediate result table is similar to *enc_features_1* and contains multiple temporal features on encounters computed from the *encounter* base table. A second intermediate result table is similar to *enc_features_2* and contains multiple temporal features on encounters computed by joining the *encounter* and *diagnosis* base tables. A third intermediate result table contains multiple temporal features on medications computed by joining the *ordered_medication* and *medication_master* base tables,

such as the total number of distinct asthma medications and the total number of units of asthma medications ordered for the patient in the prior 12 months. The logical query plan for a select-project-join-aggregate query includes 1 or more select-project-join-aggregate segments [23]. Each segment has a grouping or duplicate elimination operator at its end following a bunch of join, selection, and projection operators.

Figure 1. The flow chart for building a clinical machine learning predictive model on the training data, making predictions on the new data, and using our automated method to explain the model's predictions.

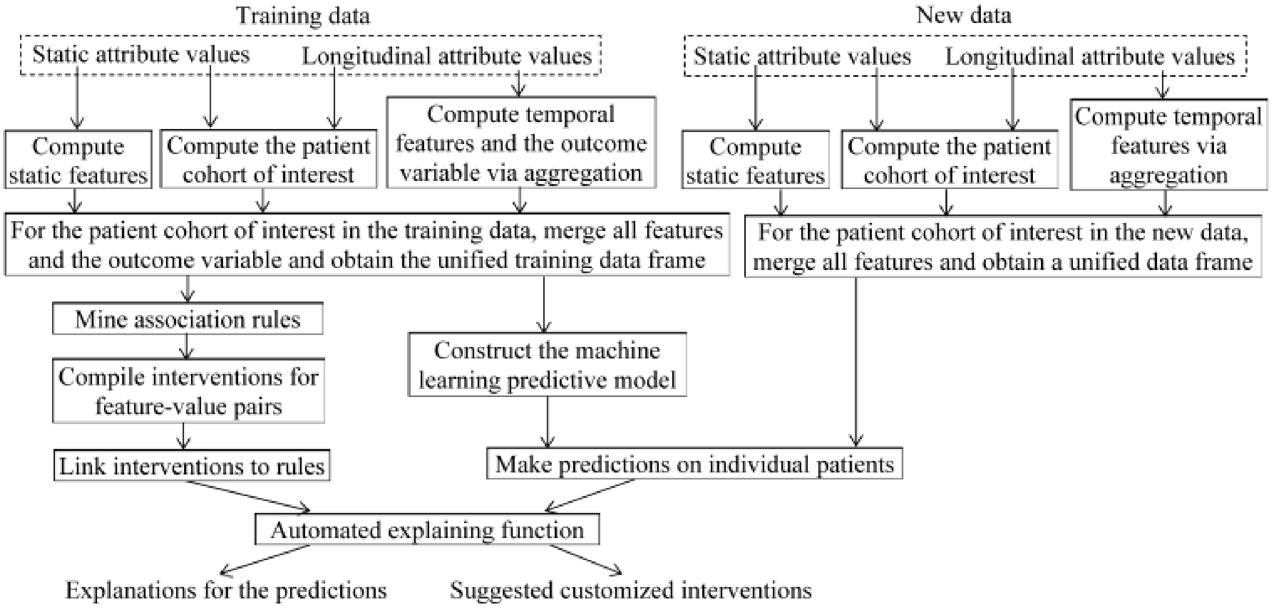
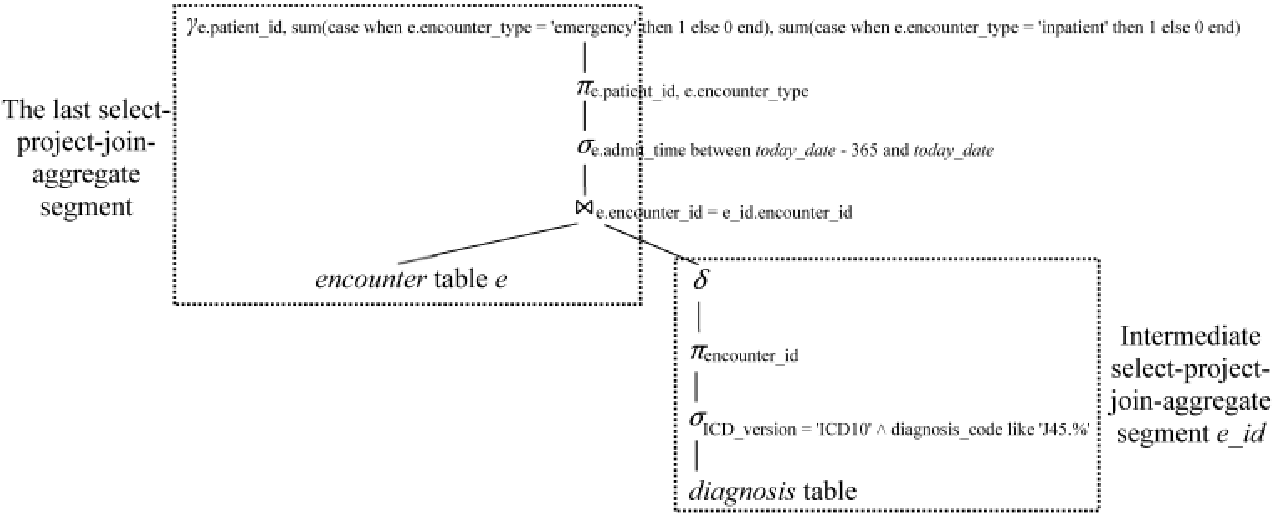

Training data
Static attribute values
Longitudinal attribute values
Compute static features
Compute the patient cohort of interest
Compute temporal features and the outcome variable via aggregation
For the patient cohort of interest in the training data, merge all features and the outcome variable and obtain the unified training data frame
Mine association rules
Compile interventions for feature-value pairs
Link interventions to rules
Construct the machine learning predictive model
New data
Static attribute values
Longitudinal attribute values
Compute static features
Compute the patient cohort of interest
Compute temporal features via aggregation
For the patient cohort of interest in the new data, merge all features and obtain a unified data frame
Make predictions on individual patients
Automated explaining function
Explanations for the predictions
Suggested customized interventions

Figure 2 shows the logical query plan for a select-project-join-aggregate query. By joining the intermediate result tables containing the patient cohort of interest, the static and temporal features, and the outcome variable in the training data, a table containing the unified training data frame is obtained. For the patient cohort of interest, this table includes 1 column for the outcome variable and a separate column for each feature. Then a machine learning predictive model is trained on this table.

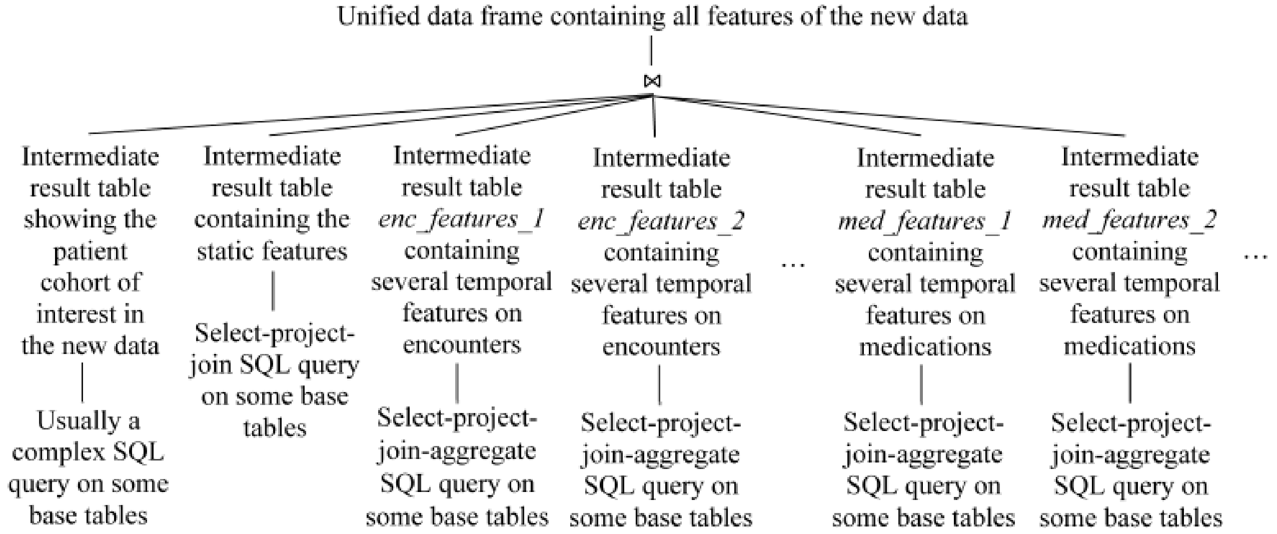
Figure 2. A logical query plan for the select-project-join-aggregate query Q_3 given in the "Intermediate result tables" section.


γe.patient_id, sum(case when e.encounter_type = 'emergency' then 1 else 0 end), sum(case when e.encounter_type = 'inpatient' then 1 else 0 end)
πe.patient_id, e.encounter_type
σe.admit_time between today_date - 365 and today_date
⋈e.encounter_id = e_id.encounter_id
encounter table e
The last select-project-join-aggregate segment
δ
πencounter_id
σICD_version = 'ICD10' ∧ diagnosis_code like 'J45.%'
diagnosis table
Intermediate select-project-join-aggregate segment e_id

Applying the Machine Learning Predictive Model to New Data to Make Predictions on Individual Patients

As Figure 3 shows, similar to the procedure mentioned above, the patient cohort of interest and the static and temporal features in the new data are computed. The results are stored in several intermediate result tables. By joining these tables, a table containing the unified data frame for the new data is obtained. For the patient cohort of interest, this table includes a separate column for each feature. We then apply the machine learning predictive model to this table to make predictions on individual patients.

Figure 3. The high-level logical query plan for computing the unified data frame that contains all the features of the new data. SQL: structured query language.


Unified data frame containing all features of the new data
⋈
Intermediate result table showing the patient cohort of interest in the new data
Usually a complex SQL query on some base tables
Intermediate result table containing the static features
Select-project-join SQL query on some base tables
Intermediate result table enc_features_1 containing several temporal features on encounters
Select-project-join-aggregate SQL query on some base tables
Intermediate result table enc_features_2 containing several temporal features on encounters
Select-project-join-aggregate SQL query on some base tables
…
Intermediate result table med_features_1 containing several temporal features on medications
Select-project-join-aggregate SQL query on some base tables
Intermediate result table med_features_2 containing several temporal features on medications
Select-project-join-aggregate SQL query on some base tables
…

Automatically Explaining the Machine Learning Model's Predictions

At the same time of building the clinical machine learning predictive model, the training data are used to create the knowledge base of the automated explaining function. We do automated discretization [24,25] to convert continuous features to categorical features. Then class-based association rules [24,26] are mined from the unified training data frame. Each rule delineates the association between a feature value pattern and a poor outcome value *c* and is of the form

$$i_1 \text{ AND } i_2 \text{ AND } \ldots \text{ AND } i_t \rightarrow c.$$

This rule shows that a patient satisfying i_1, i_2, …, and i_t tends to have an outcome value *c*. The values of *t* and *c* can change across rules. Each item i_k ($1 \le k \le t$) is a (feature, value) pair showing that a feature has a specific value or a value within a specific range. One example item of the former is that the patient had 2 ED visits related to asthma in the prior 12 months. One example item of the latter is that the patient's average respiratory rate recorded in the prior 12 months is >25 and ≤28 breaths per minute. An example rule containing both items is given in the Introduction.

For each (feature, value) pair item used to create association rules, 0 or more interventions are precompiled. The interventions precompiled for any item on a rule's left-hand side are automatically linked to the rule.

At prediction time, to avoid reducing the machine learning predictive model's performance measures, the model's predictions are used with no change. The mined association rules are used to explain these predictions rather than to make predictions. More specifically, for each patient whom the model predicts to have a poor outcome value, we find and display the rules that have this value on their right-hand sides and whose left-hand sides are fulfilled by the patient. Each rule offers 1 explanation for the prediction. The interventions linked to the rule are displayed next to it as the suggested candidate interventions.

Our automatic explanation method for machine learning predictions has been successfully applied to multiple clinical predictive modeling problems [11,12,27,28]. It has several advantages. Among all the automatic explanation methods for machine learning predictions in the literature [29,30], our method is the only one that can automatically suggest customized interventions. The rule-style explanations given by our method are easier to comprehend than the non–rule-style explanations given by many other methods. Unlike many other automatic explanation methods that either lower the machine learning predictive model's performance measures or work for only a specific machine learning algorithm, our automatic explanation method works for any machine learning algorithm on tabular data without lowering the model's performance measures. Unlike several other methods that use rules computed at prediction time to offer explanations [31,32], our method uses rules mined before prediction time to offer explanations. This is essential for our method to automatically suggest customized interventions at prediction time.

Review of the Existing Automated Lineage Tracing Techniques

In this section, the existing automated lineage tracing techniques are reviewed. An overview of such techniques developed in various fields is provided. Then, a specific set of automated lineage tracing techniques most closely related to this work is reviewed.

Overview of the Existing Automated Lineage Tracing Techniques

The lineage or provenance of a given data item *i* refers to the source data items producing *i* and how *i* was derived [33]. The former is called where-lineage. The latter is called how-lineage. Each type of lineage can be at either the schema level or the instance level. An example of where-lineage at the schema level is the set of base tables producing a specific materialized view. An example of where-lineage at the instance level is the set of tuples in the base tables producing a given temporal feature value in a materialized view. Lineage information can be computed in either an eager way or a lazy way. In the former case, lineage information is computed and stored at the same time of producing the output data. In the latter case, lineage information is computed when needed. This paper focuses on where-lineage that is at the instance level and computed in a lazy way.

Ikeda et al surveyed existing lineage tracing techniques in databases [33,34], e-science [35], and scientific data processing [36]. Among all of the lineage tracing techniques in the literature, the techniques Cui et al [23,37] developed are the most closely related to this work. These techniques are used to trace the lineage of a tuple in a materialized view [38] defined by a select-project-join-aggregate query in a relational database. Cui et al [39,40] described lineage tracing techniques for warehouse data computed via a directed acyclic graph of transformations, some of which could involve complex procedural code. Zhang et al [41] described lineage tracing techniques for data computed by arbitrary functions. In general, the more flexibility is allowed on the transformations or functions, the less efficiently lineage can be traced [39].

In big data systems, Ikeda et al [42,43] described lineage tracing techniques for data computed via a directed acyclic graph of map and reduce functions [44]. Amsterdamer et al [45] described lineage tracing techniques for data computed by using Pig Latin [46].

In scientific data processing, lineage tracing is often done on curated databases, which contain scientific data copied from other databases [36,47].

Schelter et al [48] described a method to trace the schema-level lineage of the data sets, features, models, and predictions produced in machine learning experiments.

Review of Cui et al's Automated Lineage Tracing Techniques for Relational Databases

To automatically trace the lineage of a tuple *t* in a materialized view [38] defined by a select-project-join-aggregate query, Cui et al [23,37] proceeded as follows. First, the materialized view's definition query is transformed into a canonical form of the logical query plan. As Figure 2 shows, the canonical form includes 1 or more select-project-join-aggregate segments. Each segment has 0 or 1 join operator, 0 or 1 selection operator, 0 or 1 projection operator, and a grouping or duplicate elimination operator in this particular order. Second, a separate intermediate materialized view is created for each intermediate select-project-join-aggregate segment of the canonical form. The root node of such a segment is not the root node of the canonical form. Third, we recursively trace through the hierarchy of intermediate materialized views in a top-down way. At each level of the hierarchy, the lineage tracing query for a 1-level select-project-join-aggregate materialized view is used to compute the current traced tuples' lineage with respect to each base table and each materialized view at the next lower level. For a 1-level select-project-join-aggregate materialized view $MV = \gamma(\pi_A(\sigma_C(R_1 \boxtimes R_2 \boxtimes \ldots \boxtimes R_n)))$, the lineage of a tuple set $T \subseteq MV$ with respect to the base table or the materialized view R_i ($1 \leq i \leq n$) is $\pi_{Ri}(\sigma_C(R_1 \boxtimes R_2 \boxtimes \ldots \boxtimes R_n) \boxtimes T)$. Here, the projection operator π on R_i has the set semantics, making each selected tuple in R_i appear only once. Further, all attributes of R_i appear in the projection operator and subsequently in the lineage traced on R_i. The final traced lineage of tuple *t* includes the lineage traced on every base table appearing in the canonical form.

We use an example to illustrate Cui et al's [23,37] automated lineage tracing techniques. If "create table enc_features_3" is replaced by "create materialized view enc_features_3_view" in query Q_3 given in the "Intermediate result tables" section, a query Q_{3_v} defining a materialized view *enc_features_3_view* is obtained. To trace the lineage of a tuple *t* in *enc_features_3_view* whose *patient_id* is *asthma_patient_id*, one proceeds as follows.

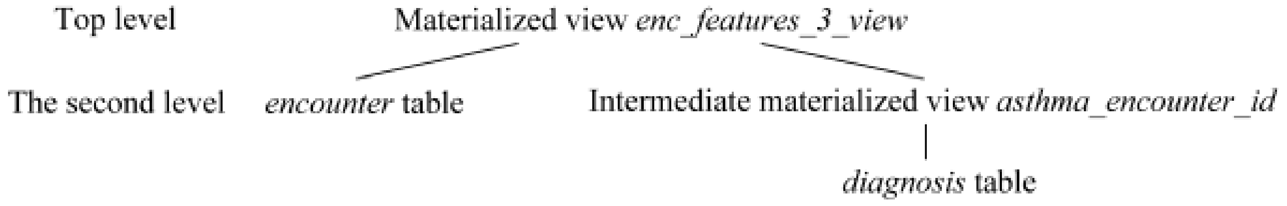
First, the canonical form of the logical query plan for query Q_{3_v} is obtained. The canonical form is the same as the logical query plan for query Q_3 shown in Figure 2.

Second, an intermediate materialized view *asthma_encounter_id* is created for the intermediate select-project-join-aggregate segment *e_id* shown in Figure 2. This is done using the following SQL query.

☒

Figure 4 shows the resulting hierarchy of intermediate materialized views, with the materialized view *enc_features_3_view* at the top and the *encounter* and *diagnosis* base tables at the bottom.

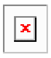
Figure 4. The hierarchy of intermediate materialized views matching the canonical form of the logical query plan for the definition query of the materialized view *enc_features_3_view*.


Top level
Materialized view enc_features_3_view
The second level
encounter table
Intermediate materialized view asthma_encounter_id
diagnosis table

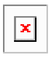
Third, at the top level of the hierarchy of intermediate materialized views, the lineage of tuple *t* with respect to the *encounter* base table is computed using the following SQL query.

☒

The following SQL query is used to compute the lineage of tuple *t* with respect to the intermediate materialized view *asthma_encounter_id* and to store the results in a temporary table *temp*.



Fourth, at the second level of the hierarchy of intermediate materialized views, the lineage of the tuples in the temporary table *temp* with respect to the *diagnosis* base table is computed using the following SQL query.



The final traced lineage of tuple *t* includes both the results of query Q_6 and the results of query Q_8.

Outline of the Proposed Automated Lineage Tracing Approach

In this section, an automated lineage tracing approach is outlined to add automated drill-through capability to the automated explaining function. Our presentation includes 4 subsections. In the first subsection, an overview of the lineage tracing component of the automated explaining function is provided. In the second subsection, the unique requirements on automated lineage tracing are shown for automatically explaining machine learning predictions for clinical decision support. In the third subsection, the proposed automated lineage tracing techniques fulfilling these requirements is outlined. In the fourth subsection, some considerations are presented for future computer coding implementation of the proposed lineage tracing approach.

Overview of the Lineage Tracing Component

At association rule mining time, all (feature, value) pair items used to create association rules are known. Which items involve temporal features computed by aggregation functions on the raw data is also known. For each item that is related to a temporal feature of a patient and on the left-hand side of a rule, a hyperlink is added to the item in the rule. In addition, a parameterized stored procedure is written for the item in the database to retrieve lineage information. The stored procedure typically has 2 parameters: the *patient_id* of the patient being examined and the endpoint of the temporal aggregation period, such as today. When the stored procedure is run for the first time, an execution plan is generated. All subsequent runs will use the same execution plan to avoid runtime query optimization overhead.

At automatic explanation time, the user of the automated explaining function is allowed to do lineage tracing for any item that is on the left-hand side of a rule-style explanation and related to a temporal feature value. When the user clicks the item's hyperlink, the stored procedure prewritten for the item is invoked to retrieve some prespecified parts of the related raw data producing the feature value. Except for the cases with 2 specific aggregation functions described later in the paper, the retrieved data instances are always displayed on a page in the reverse chronological order like that in the electronic medical records.

Unique Requirements for Automated Lineage Tracing

Typically, the user of the automated explaining function is a clinician. To fit the user's busy schedule and to aid timely decision making, the user wants the lineage tracing process for a temporal feature value to be finished quickly, preferably within 1 second. This goal is partially fulfilled by the existing lineage tracing techniques [23,37], whereas the realized lineage tracing speed can be further improved. In addition, the retrieved lineage information should be easy to scan and include the most essential content needed to facilitate decision making. This enables the user to quickly gain useful insights from the information, ideally within 1 or a few seconds. As summarized in Table 3, that goal translates to 5 unique requirements on automated lineage tracing that are unmet by the existing lineage tracing techniques.

Table 3. The 5 unique requirements of automated lineage tracing for automatically explaining machine learning predictions for clinical decision support.

Requirement	Reason for posing the requirement
Retrieving only a small set of attributes	To prevent the user from being overwhelmed by many nonessential or irrelevant attributes
Adding some essential attributes that do not directly produce the feature value	To make the retrieved lineage information include the most essential content
Sorting the retrieved lineage information in an appropriate order	To make the retrieved lineage information easy to scan
Computing the lineage information based on the semantic meaning of the feature	To avoid including irrelevant or nonessential source tuples in the retrieved lineage information
Performing no lineage tracing for any health care system feature value computed by an aggregation function	To avoid including irrelevant data in the retrieved lineage information

Requirement 1: Retrieving Only a Small Set of Attributes

When tracing the lineage of a temporal feature value, one should retrieve from the base tables only a small set of attributes specific to the temporal feature rather than the many attributes involved in deriving all of the features used for automated explanation. This requirement is posed to prevent the user of the automated explaining function from being overwhelmed by many nonessential or irrelevant attributes.

To aid automatic explanation, we want to allow tracing the lineage of a temporal feature value in the form of a small set of attributes specific to the temporal feature (see Table 2 for an example). This cannot be well done using Cui et al's lineage tracing techniques [23,37]. These techniques were developed to trace the lineage of a tuple including all of its attribute values in a select-project-join-aggregate materialized view in a relational database. If the retrieved lineage information ever touches a tuple in a base table, all attribute values of the tuple are included in this information. For automatic explanation, both factors would cause the retrieved lineage information to have an excessive volume, overwhelming the user of the automated explaining function.

To see this, the process of making predictions with automatic explanations is reviewed. Usually, many features are used to make predictions and to automatically explain them. All of the items on the left-hand side of a rule-style explanation come from the same tuple in the unified data frame, which contains all features of the new data. As Figure 3 shows, this unified data frame is obtained by joining many intermediate result tables. Each of them falls into 1 of the 3 categories: (1) a table containing the patient cohort of interest in the new data, (2) a table containing 1 or more static features, and (3) a table containing 1 or more temporal features. Each hyperlinked item on the left-hand side of a rule-style explanation comes from exactly 1 intermediate result table in the third category.

When the user of the automated explaining function clicks the hyperlink for an item on the left-hand side of a rule-style explanation, one could use Cui et al's techniques [23,37] to trace the lineage of the tuple in the unified data frame, from which the item comes. For each intermediate result table mentioned above and each base table used to create it, the retrieved lineage information contains some tuples from the base table including all of their attribute values. Most of the retrieved lineage information is unnecessary for automatic explanation for 3 reasons.

Reason 1

The retrieved lineage information often includes thousands of tuples from several dozen base tables. Most of these base tables are used to compute the other feature values in the tuple in the unified data frame that are unrelated to the item, and include no information that can help the user of the automated explaining function gain useful insights related to the item. In fact, to obtain the lineage information of the item essential for automatic explanation, we need to only trace through the intermediate result table related to the item solely for the item and to examine the base tables used to create this table. The features in this table that are unrelated to the item can be ignored. There is also no need to trace through the intermediate result tables containing the features unrelated to the item. Moreover, at automatic explanation time, we know the *patient_id* of the patient linked to the item. The user usually does not need to know why this patient is in the patient cohort of interest in the new data. Thus, there is no need to trace through the intermediate result table showing the patient cohort.

Reason 2

A base table often has many attributes, only a few of which are essential for the user of the automated explaining function to gain useful insights related to the item. For instance, the *encounter* table often has >100 attributes. The lineage information shown in Table 2 covers only 4 of them: *admit_time* transformed to the date format, *department*, *admitting_provider*, and *facility*.

Reason 3

Certain items are each computed using several base tables and intermediate query results. For the user of the automated explaining function to gain useful insights related to the item, only the attributes and tuples of some of these base tables are essential. Alternatively, none or only some of these intermediate query results need to be traced through.

For example, in query Q_2 given in the “Intermediate result tables” section, both the *encounter* and *diagnosis* base tables are used to compute the feature “the number of outpatient visits with a primary diagnosis of asthma that the patient had in the prior 12 months.” For a value of this feature, we need to use the information in the *diagnosis* table to find the related tuples in the *encounter* table. Nevertheless, the user would expect each encounter shown in the retrieved lineage information to be an outpatient visit with a primary diagnosis of asthma. Thus, there is no need to include any attribute or tuple from the *diagnosis* table in the retrieved lineage information, for example, to give the primary diagnosis of each encounter included in that information.

As a second example, in query Q_3 given in the “Intermediate result tables” section, both the *encounter* base table and the intermediate query result *e_id* are used to compute the feature “the number of ED visits related to asthma that the patient had in the prior 12 months.” For a value of this feature, the user of the automated explaining function would expect each encounter shown in the retrieved lineage information to be an ED visit related to asthma, like that shown in Table 2. Thus, there is no need to trace through *e_id* and to obtain the corresponding tuples in the *diagnosis* table showing that each encounter included in the retrieved lineage information has an asthma diagnosis code.

Requirement 2: Adding Some Essential Attributes That Do Not Directly Produce the Feature Value

For certain temporal features, when acquiring the lineage of a feature value, one should not use only the related raw data that directly produce the feature value. Instead, one needs to add to them some related attributes in the base tables, which are specific to the temporal feature and do not directly produce the feature value. We pose this requirement to make the retrieved lineage information include the most essential content needed to facilitate decision making. For example, as query Q_1 given in the “Intermediate result tables” section shows, the feature “the number of ED visits that the patient had in the prior 12 months” is computed solely from the *encounter* base table. For a value of this feature, we want the retrieved lineage information to be similar to that shown in Table 2 and include a primary diagnosis column. This column is computed using the *diagnosis* and *diagnosis_code_master* base tables unused in Q_1 and is formed by concatenating the *diagnosis_code* and *dx_code_description* columns of the *diagnosis_code_master* base table. The cases for many other temporal features on encounters are similar.

Requirement 3: Sorting the Retrieved Lineage Information in an Appropriate Order

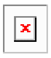
When presenting the lineage information, the related raw data retrieved for a temporal feature value should be sorted in an order specific to the temporal feature. This requirement is posed to make the retrieved lineage information easy to scan. Usually, we want the data instances in the retrieved lineage information to be displayed in the reverse chronological order like that in the electronic medical records. However, there are 2 exceptions. First, when the temporal feature is the maximum value of an attribute of a given patient, we want the related raw data retrieved for a feature value to be displayed in the descending order of the attribute value. For example, for the feature "the highest systolic blood pressure of the patient in the prior 12 months," we want the lineage information retrieved for a feature value to contain the systolic blood pressure of the patient in the prior 12 months sorted in the descending order. Second, when the temporal feature is the minimum value of an attribute of a given patient, we want the related raw data retrieved for a feature value to be displayed in the ascending order of the attribute value. In either of the 2 cases, a resort button could be added to the retrieved lineage information on display. If the user of the automated explaining function clicks this button, the data instances in the retrieved lineage information are rearranged in the reverse chronological order for display.

Requirement 4: Computing the Lineage Information Based on the Semantic Meaning of the Feature

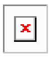
The lineage information of a temporal feature value should be computed based on the semantic meaning of the feature rather than solely on the literal writing of the SQL query used to compute the feature. We pose this requirement to avoid including irrelevant or nonessential source tuples in the retrieved lineage information. For a select-project-join-aggregate materialized view containing 1 or more temporal features, Cui et al [23,37] compute the lineage of a tuple in it based solely on the literal SQL query used to define it. In certain cases, this literal approach is suboptimal for automatic explanation. Instead, we should consider the semantic meanings of the temporal features during lineage tracing. In the following, 2 such cases are described. Each case is presented as a subrequirement.

Subrequirement 4.1

When the temporal feature is the sum of a variable computed by a case statement in SQL including multiple conditions and some of them return 0, only the lineage information related to the other conditions should be retrieved. In SQL, such a temporal feature is written in the form of



As an example of this subrequirement, for the feature "the number of ED visits that the patient had in the prior 12 months," the lineage information retrieved for a value of the feature should be the ED visits that the patient had in the prior 12 months, regardless of whether the feature is computed using SQL query Q_9 or Q_{10} below.



The differences between Q_9 and Q_{10} are highlighted in italics in Q_{10}. If the feature is computed using Q_9, Cui et al's techniques [23,37] would retrieve all the encounters of the patient in the prior 12 months as the lineage information. This could easily overwhelm the user of the automated explaining function, as usually most of these encounters are not ED visits.

Subrequirement 4.2

When the temporal feature is the total number of distinct items, the retrieved lineage information should include only 1 representative data instance for each distinct item. For example, query Q_4 given in the "Intermediate result tables" section computes the feature "the total number of distinct medications ordered for the patient in the prior 12 months." For a value of this feature, Cui et al's techniques [23,37] would retrieve all medications ordered for the patient in the prior 12 months as the lineage information. This information is often overwhelming and not succinct enough for the user of the automated explaining function to quickly find the distinct medications ordered for the patient in the prior 12 months, as the same medication could be ordered for the patient multiple times in a year. To avoid this problem, one could retrieve only the most recent order of each distinct medication ordered for the patient in the prior 12 months as the lineage information. For the user, these distinct medications typically provide enough insight into the patient's status related to the feature value.

Requirement 5: Performing No Lineage Tracing for Any Health Care System Feature Value Computed by an Aggregation Function

We do not trace the lineage of any health care system feature value computed by an aggregation function. We pose this requirement to avoid including irrelevant data in the retrieved lineage information. Like temporal features of a patient, certain health care system features [17-19] such as the number of patients with asthma of the primary care provider of a patient are computed by aggregation functions. These health care system features are each computed using multiple patients' information rather than solely the information of the patient being examined. Since other patients' detailed information does not help the user of the automated explaining function understand this patient's situation, we do not trace the lineage of any value of this feature, even if it appears on the left-hand side of a rule-style explanation.

Outline of the Proposed Techniques to Form the Lineage Tracing Query That Computes the Lineage Information

To perform automated lineage tracing for explaining machine learning predictions for clinical decision support, Cui et al's lineage tracing techniques [23,37] are modified to fulfill the requirements mentioned above. Even without giving any detail on the computer coding implementation and the performance evaluation results, Cui et al [37] already used 49 pages to describe the details of their automated lineage tracing algorithm. The case described in this paper is more complex than Cui et al's case [37]. In the case described in this paper, which attributes are most relevant and which source tuples are most essential for inclusion in the retrieved lineage information depend on both the concrete feature type and the clinical decision support application's need. In comparison, no such dependency exists in Cui et al's case [37]. Thus, it is expected that, once fully worked out, the proposed automated lineage tracing algorithm would be more sophisticated than Cui et al's algorithm [37]. In this viewpoint paper, the goal is not to enumerate all possible feature types and to provide a detailed design or any computer coding implementation of the proposed automated lineage tracing approach. Rather, the goal is to describe the design approach for the proposed automated lineage tracing module and to provide a roadmap for future research. We achieve this goal by outlining the main steps of forming the lineage tracing query, giving 4 example temporal features, and illustrating at a high level how to form the lineage tracing query for each of these 4 features.

Overview of the Lineage Tracing Query Formation Process

Usually, each intermediate result table shown in Figure 3 has a *patient_id* column. It is used as the join column in the join operation to produce the unified data frame containing all features of the new data. As explained in "Reason 1" of the "Requirement 1" section, to obtain the lineage information of a temporal feature value, we need to only trace through the intermediate result table containing this value solely for this value. This intermediate result table is usually computed from some base tables by using a select-project-join-aggregate SQL query S_0. To form the lineage tracing query for a temporal feature value of a patient in the intermediate result table, one proceeds in 4 steps. First, the other temporal features, if any, are removed from S_0 to obtain a simplified query S_1. Second, if applicable, S_1 is transformed to query S_2 to fulfill subrequirement 4.1. Third, Cui et al's techniques [23,37] are modified to address Reasons 2 and 3 given in the "Requirement 1" section. The modified techniques are used to form a preliminary lineage tracing query S_3 based on S_2 and the patient's *patient_id*. Fourth, to obtain the final lineage tracing query, S_3 is transformed to fulfill Requirements 2 and 3 and subrequirement 4.2.

In the following, 4 examples are used to illustrate at a high level how to form the lineage tracing query. In each example, the user of the automated explaining function is examining a patient with asthma whose identifier is *asthma_patient_id* and wants to drill through a temporal feature value of this patient. We outline the main steps of forming the lineage tracing query for the feature value without giving the detailed algorithm.

Example 1: The Number of ED Visits That the Patient Had in the Prior 12 Months

As defined by query Q_1 in the "Intermediate result tables" section, the intermediate result table *enc_features_1* contains 3 temporal features. One of them is the number of ED visits that the patient had in the prior 12 months. To form the lineage tracing query for a value of this feature, one proceeds as follows.

First, the other 2 features are removed from query Q_1 to obtain query Q_9 given in the "Subrequirement 4.1" section.

Second, to fulfill subrequirement 4.1 on handling the sum of a variable computed by a case statement, query Q_9 is transformed to query Q_{10} given in the "Subrequirement 4.1" section.

Third, Cui et al's lineage tracing techniques [23,37] are used to form a draft lineage tracing query Q_{11} based on Q_{10} and *asthma_patient_id*.

The differences between Q_{10} and Q_{11} are highlighted in italics in Q_{11}. To address Reason 2 given in the "Requirement 1" section and retrieve from the *encounter* table only its attributes essential for automatic explanation, Q_{11} is transformed to the following preliminary lineage tracing query.

The differences between Q_{11} and Q_{12} are highlighted in italics in Q_{12}.

Fourth, to fulfill Requirement 2, a primary diagnosis column needs to be added to the raw data that are retrieved by query Q_{12} and that directly produce the feature value being examined. To fulfill Requirement 3, the retrieved raw data need to be sorted in the reverse chronological order. To meet both demands, Q_{12} is transformed to the following final lineage tracing query.

The differences between Q_{12} and Q_{13} are highlighted in italics in Q_{13}. || is the string concatenation operator in SQL.

Example 2: The Number of Outpatient Visits With a Primary Diagnosis of Asthma That the Patient Had in the Prior 12 Months

As defined by query Q_2 in the "Intermediate result tables" section, the intermediate result table *enc_features_2* contains the temporal feature "the number of outpatient visits with a primary diagnosis of asthma that the patient had in the prior 12 months." To form the lineage tracing query for a value of this feature, one proceeds as follows.

First, to address Reason 2 given in the "Requirement 1" section, only the attributes essential for automatic explanation should be included from the *encounter* table. To address Reason 3 given in the "Requirement 1" section, no attribute or tuple from the *diagnosis* table should be included in the retrieved lineage information. A preliminary lineage tracing query Q_{14} is formed based on query Q_2 and *asthma_patient_id* by using a modified version of Cui et al's lineage tracing techniques [23,37] that meets both demands.

The differences between Q_2 and Q_{14} are highlighted in italics in Q_{14}.

Second, to fulfill Requirement 3 of sorting the related raw data retrieved for the feature value in the reverse chronological order, query Q_{14} is transformed to the following final lineage tracing query.

The differences between Q_{14} and Q_{15} are highlighted in italics in Q_{15}.

Example 3: The Number of ED Visits Related to Asthma That the Patient Had in the Prior 12 Months

As defined by query Q_3 in the "Intermediate result tables" section, the intermediate result table *enc_features_3* contains 2 temporal features. One of them is the number of ED visits related to asthma that the patient had in the prior 12 months. To form the lineage tracing query for a value of this feature, one proceeds as follows.

First, the other feature is removed from query Q_3 to obtain the following simplified query.

Second, to fulfill subrequirement 4.1 on handling the sum of a variable computed by a case statement, query Q_{16} is transformed to the following query.

The differences between Q_{16} and Q_{17} are highlighted in italics in Q_{17}.

Third, to address Reason 2 given in the "Requirement 1" section, only the attributes essential for automatic explanation should be included from the *encounter* table. To address Reason 3 given in the "Requirement 1" section, the intermediate query result *e_id* should not be traced through to include any corresponding tuple in the *diagnosis* table in the retrieved lineage information. A preliminary lineage tracing query Q_{18} is formed based on query Q_{17} and *asthma_patient_id* by using a modified version of Cui et al's lineage tracing techniques [23,37] that meets both demands.

The differences between Q_{17} and Q_{18} are highlighted in italics in Q_{18}.

Cui et al's lineage tracing techniques [23,37,49] are applied to query Q_3 to create a materialized view *asthma_encounter_id*, which is defined by query Q_5 in the "Review of Cui et al's automated lineage tracing techniques for relational databases" section. The *asthma_encounter_id* is used to rewrite the preliminary lineage tracing query Q_{18} as follows.

The differences between Q_{18} and Q_{19} are highlighted in italics in Q_{19}.

Fourth, to fulfill Requirement 2, a primary diagnosis column needs to be added to the raw data that are retrieved by query Q_{19} and that directly produce the feature value being examined. To fulfill Requirement 3, the retrieved raw data need to be sorted in the reverse chronological order. To meet both demands, Q_{19} is transformed to the following final lineage tracing query.

The differences between Q_{19} and Q_{20} are highlighted in italics in Q_{20}.

Example 4: The Total Number of Distinct Medications Ordered for the Patient in the Prior 12 Months

As defined by query Q_4 in the "Intermediate result tables" section, the intermediate result table *med_features_1* contains 2 temporal features. One of them is the total number of distinct medications ordered for the patient in the prior 12 months. To form the lineage tracing query for a value of this feature, one proceeds as follows.

First, the other feature is removed from query Q_4 to obtain the following simplified query.

Second, to address Reason 2 given in the "Requirement 1" section, only the attributes essential for automatic explanation should be included from the *ordered_medication* table. A preliminary lineage tracing query Q_{22} is formed based on query Q_{21} and *asthma_patient_id* by using a modified version of Cui et al's lineage tracing techniques [23,37] that meets this demand.

The differences between Q_{21} and Q_{22} are highlighted in italics in Q_{22}.

Third, to fulfill subrequirement 4.2, one could retrieve only the most recent order of each distinct medication ordered for the patient in the prior 12 months as the lineage information. This is done by transforming query Q_{22} to the following query.

The differences between Q_{22} and Q_{23} are highlighted in italics in Q_{23}.

Fourth, to fulfill requirement 2, a medication name column is added to the raw data that are retrieved by query Q_{23} and directly produce the feature value being examined. To fulfill Requirement 3, the retrieved raw data are sorted in the reverse chronological order. Q_{23} is transformed to the following final lineage tracing query to meet both demands.

The differences between Q_{23} and Q_{24} are highlighted in italics in Q_{24}.

Considerations for Future Computer Coding Implementation of the Proposed Automated Lineage Tracing Approach

Maximizing the Automation Degree of the Lineage Tracing Query Formation Process

For a select-project-join-aggregate materialized view, Cui et al [23,37] used a fully automated approach to analyze its definition query to derive a lineage tracing query for a tuple in it. In the case of automatically explaining machine learning predictions, all temporal features used for making predictions and automatic explanation are known at machine learning model building time. In general, for each temporal feature, we can form a lineage tracing query either manually or semiautomatically, but often not fully automatically, beforehand. Nevertheless, once the query is formed and put into the knowledge base of the automated explaining function, we can use the query to automatically retrieve the lineage information of a value of the feature at prediction time.

As mentioned before, automatic explanation poses several unique requirements on automated lineage tracing. Two of them make it difficult to fully automate the lineage tracing query formation process. First, Requirement 1 says that the lineage information retrieved for a temporal feature value should include only a small set of relevant attributes specific to the temporal feature. Almost infinite attributes and temporal features could possibly be used for clinical machine learning. Thus, it is infeasible to precompile the set of relevant attributes for every possible temporal feature. Second, Requirement 2 says that when acquiring the lineage of a value for certain temporal features, we need to include some attributes that are specific to the temporal feature and do not directly produce the feature value. For a reason similar to the above, it is infeasible to precompile the set of such attributes for every possible such temporal feature.

Although the lineage tracing query formation process cannot be fully automated in the most general case, 2 methods can still be used to maximize the process' automation degree and to reduce the workload of the developers of the automated explaining function. First, for a temporal feature, an approach similar to that of Cui et al [23,37] can be used to automatically form a draft lineage tracing query. The developers of the automated explaining function revise this query as needed to obtain the final lineage tracing query. Second, the same temporal feature is often used for multiple predictive modeling tasks. One can create a library of lineage tracing queries for temporal features to facilitate query reuse across various predictive modeling tasks. This library is formed for a data set in the Observational Medical Outcomes Partnership common data model format [50] using its linked standardized terminologies [51]. This format standardizes administrative and clinical variables from ≥10 large US health care systems [52,53]. For any data set that is put into this format, we can use this library to obtain lineage tracing queries.

Improving the Lineage Tracing Speed

As mentioned before, the user of the automated explaining function wants the lineage tracing process for a temporal feature value to be finished quickly, preferably within 1 second. To expedite tracing the lineage of a tuple in a materialized view defined by a select-project-join-aggregate query *S*, Cui et al [23,37,49] advocated creating a materialized view for each intermediate select-project-join-aggregate segment of the canonical form of the logical query plan for *S*. While this boosts the lineage tracing speed, the resulting speed is still not fast enough to reach a subsecond response time [23,39]. To further improve the lineage tracing speed, we can build indices [39,42] on the selection and join attributes of both the base tables and the materialized views created for the intermediate select-project-join-aggregate segments. For instance, in Example 3, we can build 1 index on the *encounter_id* column of the materialized view *asthma_encounter_id* and another index on the *patient_id* column of the *encounter* base table. We can create indices either manually or by using an automated index design tool provided by a commercial relational database system [54-56]. Typically, each intermediate result table containing 1 or more temporal features is computed on 1 or a few base tables using no more than a small number of join operations. The lineage tracing query for a temporal feature value falls into a similar case. Thus, with appropriate indices, we would expect the lineage tracing query to finish execution quickly. For base tables of moderate sizes and simple materialized views, Cui and Widom [39] showed that lineage tracing can be done within 1 second when indices exist on the keys of the base tables. For large base tables and temporal features computed through more complex procedures, we would expect that more indices are needed to reach a subsecond response time.

The above discussion focuses on the case that the electronic medical record data are stored in a relational database and features are extracted using SQL queries. When the electronic medical record data are stored in a big data system and features are extracted using map and reduce functions [44] or Pig Latin [46], we can modify the corresponding existing lineage tracing techniques [42,43,45] in a similar way to enable lineage tracing to aid automatically explaining machine learning predictions for clinical decision support.

Discussion

Directions for Future Research

The above discussion describes the high-level design approach for the proposed automated lineage tracing module. To complete the detailed design of the proposed automated lineage tracing approach, implement the module in computer code, and test the module's performance, much research is needed along the following directions:

1. We need to compile a list of attributes and temporal feature types most commonly used in building clinical machine learning predictive models. For these attributes and temporal feature types, we need to complete the detailed design and the computer coding implementation of the proposed automated lineage tracing approach.
2. We need to come up with an automated approach to design indices needed for improving the lineage tracing speed. The database research community has developed several automated index design approaches [54-56]. We can modify

these approaches to fit the database querying workload posed by automated lineage tracing.

3. We plan to assess the execution speed of the proposed automated lineage tracing approach after implementing it in computer code.
4. As shown by prior work on automated lineage tracing shown in the "Overview of the existing automated lineage tracing techniques" section, the database research community takes it for granted that automated lineage tracing could help users better understand the data and save time in doing data analysis. To the best of our knowledge, no formal study to date has been published on measuring the impact of automated lineage tracing on users' data analysis and decision-making process. After implementing the proposed automated lineage tracing module, we plan to choose several clinical predictive modeling tasks and assess for each task, the impact of offering the module on the data analysis and decision-making process of the users of the automated explaining function. In particular, we plan to evaluate whether the addition of the module benefits the user and improves outcomes, for example, by saving the user's time, making it easier for the user to understand the predictions given by the machine learning predictive model and helping the user better understand the patient's situation and make better clinical decisions.

Limitations of the Proposed Approach

The proposed automated lineage tracing approach has several limitations:

1. To build clinical machine learning predictive models, we usually use temporal features that are computed by SQL queries of low or moderate complexities. It is possible that some temporal features used to build certain predictive models are computed by rather complex SQL queries. We may not be able to finish the lineage tracing process for a value of such a temporal feature quickly, regardless of how many indices are built to expedite this process. For example, this could happen if the SQL query uses complex procedural code, which has no property that can be used to simplify the lineage tracing process [39]. Having a long lineage tracing time could make the user of the automated explaining function become impatient. Nevertheless, it is still faster and more convenient to do lineage tracing using the automated approach than to let the user do manual drill-through.
2. The proposed automated lineage tracing approach works for any feature values computed by the standard aggregation functions in SQL on longitudinal structured data. For certain deep learning predictive models built on longitudinal structured data, the previously proposed method [16] could be used to semiautomatically extract comprehensible and predictive temporal features from the models and the longitudinal structured data, and then apply the automated approach to trace the lineage of the values of these features. For any other deep learning predictive model that is built directly on longitudinal structured data and that uses incomprehensible features hidden in the neurons of the deep neural network, the proposed automated approach can no longer be used to trace the lineage of the values of these features.
3. Almost infinite attributes and temporal features could possibly be used for clinical machine learning. Further, some attributes are not covered by the Observational Medical Outcomes Partnership common data model. For the reasons given in the "Maximizing the automation degree of the lineage tracing query formation process" section, we could maximize the automation degree of the lineage tracing query formation process for only certain types of temporal features formed on certain attributes. For any other temporal feature, the developers of the automated explaining function could still need a nontrivial amount of time to create the corresponding lineage tracing query.

Conclusions

Automatically explaining machine learning predictions is critical to overcome the model interpretability barrier to using machine learning predictive models in clinical practice. Our previously developed automatic explanation method for machine learning predictions can be used to address this barrier, but a gap remains to fulfill the need of rapidly drilling through a feature value in an explanation that is computed by an aggregation function on the raw data. This paper articulates this gap, outlines an automated lineage tracing approach to close the gap, and provides a roadmap for future research. The automated drill-through capability is intended to be offered to help the user of the automated explaining function save time, better understand the patient's situation, and make better clinical decisions. It would take several people multiple years to work out the detailed design and the computer coding implementation of the proposed automated lineage tracing approach. We hope this paper will make some researchers become interested in and join the research endeavor on this topic. Only after the detailed design and the computer coding implementation of the proposed automated lineage tracing approach are fully worked out, one could deploy the automated lineage tracing module in clinical practice and measure the module's impact on clinicians' decision-making process. The principle of the automated lineage tracing approach generalizes to nonmedical data and other automated methods to explain machine learning predictions.

Acknowledgments

We thank Xiaoyi Zhang and Brian Kelly for the useful discussions. GL was partially supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under award number R01HL142503. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of Interest

None declared.

References

1. Kaggle. URL: https://www.kaggle.com [accessed 2021-04-30]
2. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, 2nd ed. New York, USA: Springer; 2019.
3. Lee G, Wang S, Dipuro F, Hou J, Grover P, Low LL, et al. Leveraging on predictive analytics to manage clinic no show and improve accessibility of care. 2017 Presented at: Proceedings of 2017 IEEE International Conference on Data Science and Advanced Analytics; October 19-21, 2017; Tokyo, Japan p. 429-438. [doi: 10.1109/dsaa.2017.25]
4. Dean NC, Jones BE, Jones JP, Ferraro JP, Post HB, Aronsky D, et al. Impact of an electronic clinical decision support tool for emergency department patients with pneumonia. Ann Emerg Med 2015;66(5):511-520. [doi: 10.1016/j.annemergmed.2015.02.003] [Medline: 25725592]
5. Hsu JC, Chen YF, Chung WS, Tan TH, Chen T, Chiang JY. Clinical verification of a clinical decision support system for ventilator weaning. Biomed Eng Online 2013;12 Suppl 1:S4 [FREE Full text] [doi: 10.1186/1475-925X-12-S1-S4] [Medline: 24565021]
6. Barbieri C, Molina M, Ponce P, Tothova M, Cattinelli I, Ion Titapiccolo J, et al. An international observational study suggests that artificial intelligence for clinical decision support optimizes anemia management in hemodialysis patients. Kidney Int 2016;90(2):422-429 [FREE Full text] [doi: 10.1016/j.kint.2016.03.036] [Medline: 27262365]
7. Brier ME, Gaweda AE, Dailey A, Aronoff GR, Jacobs AA. Randomized trial of model predictive control for improved anemia management. Clin J Am Soc Nephrol 2010 May;5(5):814-820 [FREE Full text] [doi: 10.2215/CJN.07181009] [Medline: 20185598]
8. Gaweda AE, Aronoff GR, Jacobs AA, Rai SN, Brier ME. Individualized anemia management reduces hemoglobin variability in hemodialysis patients. J Am Soc Nephrol 2014 Jan;25(1):159-166 [FREE Full text] [doi: 10.1681/ASN.2013010089] [Medline: 24029429]
9. Gaweda AE, Jacobs AA, Aronoff GR, Brier ME. Model predictive control of erythropoietin administration in the anemia of ESRD. Am J Kidney Dis 2008 Jan;51(1):71-79. [doi: 10.1053/j.ajkd.2007.10.003] [Medline: 18155535]
10. Hamlet KS, Hobgood A, Hamar GB, Dobbs AC, Rula EY, Pope JE. Impact of predictive model-directed end-of-life counseling for Medicare beneficiaries. Am J Manag Care 2010 May;16(5):379-384 [FREE Full text] [Medline: 20469958]
11. Luo G. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. Health Inf Sci Syst 2016;4:2 [FREE Full text] [doi: 10.1186/s13755-016-0015-4] [Medline: 26958341]
12. Luo G, Johnson MD, Nkoy FL, He S, Stone BL. Automatically explaining machine learning prediction results on asthma hospital visits in asthmatic patients: secondary analysis. JMIR Med Inform 2020 Dec 31;8(12):e21965 [FREE Full text] [doi: 10.2196/21965] [Medline: 33382379]
13. Tong Y, Messinger AI, Luo G. Testing the generalizability of an automated method for explaining machine learning predictions on asthma patients' asthma hospital visits to an academic health care system. IEEE Access 2020;8:195971-195979 [FREE Full text] [doi: 10.1109/access.2020.3032683] [Medline: 33240737]
14. Luo G, Nau CL, Crawford WW, Schatz M, Zeiger RS, Koebnick C. Generalizability of an automatic explanation method for machine learning prediction results on asthma-related hospital visits in patients with asthma: quantitative analysis. J Med Internet Res 2021 Apr 15;23(4):e24153 [FREE Full text] [doi: 10.2196/24153] [Medline: 33856359]
15. Halamka JD. Early experiences with big data at an academic medical center. Health Aff (Millwood) 2014 Jul;33(7):1132-1138. [doi: 10.1377/hlthaff.2014.0031] [Medline: 25006138]
16. Luo G. A roadmap for semi-automatically extracting predictive and clinically meaningful temporal features from medical data for predictive modeling. Glob Transit 2019;1:61-82 [FREE Full text] [doi: 10.1016/j.glt.2018.11.001] [Medline: 31032483]
17. Luo G, Nau CL, Crawford WW, Schatz M, Zeiger RS, Rozema E, et al. Developing a predictive model for asthma-related hospital encounters in patients with asthma in a large, integrated health care system: secondary analysis. JMIR Med Inform 2020 Nov 09;8(11):e22689 [FREE Full text] [doi: 10.2196/22689] [Medline: 33164906]
18. Tong Y, Messinger AI, Wilcox AB, Mooney SD, Davidson GH, Suri P, et al. Forecasting future asthma hospital encounters of patients with asthma in an academic health care system: predictive model development and secondary analysis study. J Med Internet Res 2021 Apr 16;23(4):e22796 [FREE Full text] [doi: 10.2196/22796] [Medline: 33861206]
19. Luo G, He S, Stone BL, Nkoy FL, Johnson MD. Developing a model to predict hospital encounters for asthma in asthmatic patients: secondary analysis. JMIR Med Inform 2020 Jan 21;8(1):e16080 [FREE Full text] [doi: 10.2196/16080] [Medline: 31961332]
20. Garcia-Molina H, Ullman JD, Widom J. Database Systems: the Complete Book, 2nd ed. Upper Saddle River, NJ: Pearson; 2008.

21. Cunningham C, Graefe G, Galindo-Legaria CA. PIVOT and UNPIVOT: optimization and execution strategies in an RDBMS. 2004 Presented at: Proceedings of the 30th International Conference on Very Large Data Bases; August 31-September 3, 2004; Toronto, Canada p. 998-1009. [doi: 10.1016/b978-012088469-8.50087-5]
22. Lyman JA, Scully K, Harrison JHJ. The development of health care data warehouses to support data mining. Clin Lab Med 2008 Mar;28(1):55-71. [doi: 10.1016/j.cll.2007.10.003] [Medline: 18194718]
23. Cui Y, Widom J. Practical lineage tracing in data warehouses. 2000 Presented at: Proceedings of the 16th International Conference on Data Engineering; February 28-March 3, 2000; San Diego, CA p. 367-378. [doi: 10.1109/icde.2000.839437]
24. Liu B, Hsu W, Ma Y. Integrating classification and association rule mining. 1998 Presented at: Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining; August 27-31, 1998; New York City, USA p. 80-86.
25. Fayyad UM, Irani KB. Multi-interval discretization of continuous-valued attributes for classification learning. 1993 Presented at: Proceedings of the 13th International Joint Conference on Artificial Intelligence; August 28-September 3, 1993; Chambéry, France p. 1022-1029.
26. Thabtah FA. A review of associative classification mining. The Knowledge Engineering Review 2007 Mar 01;22(1):37-65. [doi: 10.1017/s0269888907001026]
27. Alaa AM, van der Schaar M. Prognostication and risk factors for cystic fibrosis via automated machine learning. Sci Rep 2018 Jul 26;8(1):11242 [FREE Full text] [doi: 10.1038/s41598-018-29523-2] [Medline: 30050169]
28. Alaa AM, van der Schaar M. AutoPrognosis: automated clinical prognostic modeling via Bayesian optimization with structured kernel learning. 2018 Presented at: Proceedings of 35th International Conference on Machine Learning; July 10-15, 2018; Stockholm, Sweden p. 139-148.
29. Molnar C. Interpretable Machine Learning. Morrisville, NC: lulu.com; 2020.
30. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv 2019 Jan 23;51(5):93. [doi: 10.1145/3236009]
31. Rudin C, Shaposhnik Y. Globally-consistent rule-based summary-explanations for machine learning models: application to credit-risk evaluation. 2019 Presented at: Proceedings of INFORMS 11th Conference on Information Systems and Technology; October 19-20, 2019; Seattle, WA p. 1-19. [doi: 10.2139/ssrn.3395422]
32. Ribeiro MT, Singh S, Guestrin C. Anchors: high-precision model-agnostic explanations. 2018 Presented at: Proceedings of the 32nd AAAI Conference on Artificial Intelligence; February 2-7, 2018; New Orleans, LA p. 1527-1535.
33. Ikeda R, Widom J. Data lineage: a survey. Stanford University Technical Report. URL: http://ilpubs.stanford.edu:8090/918/1/lin_final.pdf [accessed 2021-04-30]
34. Cheney J, Chiticariu L, Tan WC. Provenance in Databases: Why, How, and Where. Found Trends Databases 2009;1(4):379-474. [doi: 10.1561/1900000006]
35. Simmhan Y, Plale B, Gannon D. A survey of data provenance in e-science. SIGMOD Rec 2005 Sep;34(3):31-36. [doi: 10.1145/1084805.1084812]
36. Bose R, Frew J. Lineage retrieval for scientific data processing: a survey. ACM Comput Surv 2005 Mar;37(1):1-28. [doi: 10.1145/1057977.1057978]
37. Cui Y, Widom J, Wiener JL. Tracing the lineage of view data in a warehousing environment. ACM Trans Database Syst 2000 Jun;25(2):179-227. [doi: 10.1145/357775.357777]
38. Gupta A, Mumick IS. Materialized Views: Techniques, Implementations, and Applications. Cambridge, MA: The MIT Press; 1999.
39. Cui Y, Widom J. Lineage tracing for general data warehouse transformations. The VLDB Journal The International Journal on Very Large Data Bases 2003 May 1;12(1):41-58. [doi: 10.1007/s00778-002-0083-8]
40. Ikeda R, Sarma AD, Widom J. Logical provenance in data-oriented workflows. 2013 Presented at: Proceedings of the 29th IEEE International Conference on Data Engineering; April 8-12, 2013; Brisbane, Australia p. 877-888. [doi: 10.1109/icde.2013.6544882]
41. Zhang M, Zhang X, Prabhakar S. Tracing lineage beyond relational operators. 2007 Presented at: Proceedings of the 33rd International Conference on Very Large Data Bases; September 23-27, 2007; Vienna, Austria p. 1116-1127.
42. Ikeda R, Park H, Widom J. Provenance for generalized map and reduce workflows. 2011 Presented at: Proceedings of the 5th Biennial Conference on Innovative Data Systems Research; January 9-12, 2011; Asilomar, CA p. 273-283.
43. Park H, Ikeda R, Widom J. RAMP: a system for capturing and tracing provenance in MapReduce workflows. Proc VLDB Endow 2011 Aug;4(12):1351-1354. [doi: 10.14778/3402755.3402768]
44. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. 2004 Presented at: Proceedings of the 6th Symposium on Operating System Design and Implementation; December 6-8, 2004; San Francisco, CA p. 137-150.
45. Amsterdamer Y, Davidson SB, Deutch D, Milo T, Stoyanovich J, Tannen V. Putting Lipstick on Pig: enabling database-style workflow provenance. Proc VLDB Endow 2011 Dec;5(4):346-357. [doi: 10.14778/2095686.2095693]
46. Olston C, Reed B, Srivastava U, Kumar R, Tomkins A. Pig Latin: a not-so-foreign language for data processing. 2008 Presented at: Proceedings of the ACM SIGMOD International Conference on Management of Data; June 10-12, 2008; Vancouver, BC, Canada p. 1099-1110. [doi: 10.1145/1376616.1376726]

47. Buneman P, Chapman A, Cheney J. Provenance management in curated databases. 2006 Presented at: Proceedings of the ACM SIGMOD International Conference on Management of Data; June 27-29, 2006; Chicago, IL p. 539-550. [doi: 10.1145/1142473.1142534]
48. Schelter S, Böse J, Kirschnick J, Klein T, Seufert S. Automatically tracking metadata and provenance of machine learning experiments. 2017 Presented at: Proceedings of the ML Systems Workshop at NIPS 2017; December 8, 2017; Long Beach, CA p. 1-8.
49. Cui Y, Widom J. Storing auxiliary data for efficient maintenance and lineage tracing of complex views. 2000 Presented at: Proceedings of the Second Intl Workshop on Design and Management of Data Warehouses; June 5-6, 2000; Stockholm, Sweden p. 1-19.
50. Data standardization. Observational Health Data Sciences and Informatics. URL: https://www.ohdsi.org/data-standardization [accessed 2021-04-30]
51. Standardized vocabularies. Observational Health Data Sciences and Informatics. URL: https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:sidebar [accessed 2021-04-30]
52. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
53. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19(1):54-60 [FREE Full text] [doi: 10.1136/amiajnl-2011-000376] [Medline: 22037893]
54. Das S, Grbic M, Ilic I, Jovandic I, Jovanovic A, Narasayya VR, et al. Automatically indexing millions of databases in Microsoft Azure SQL database. 2019 Presented at: Proceedings of the ACM SIGMOD International Conference on Management of Data; June 30-July 5, 2019; Amsterdam, Netherlands p. 666-679. [doi: 10.1145/3299869.3314035]
55. Dageville B, Das D, Dias K, Yagoub K, Zaït M, Ziauddin M. Automatic SQL tuning in Oracle 10g. 2004 Presented at: Proceedings of the 30th International Conference on Very Large Data Bases; August 31-September 3, 2004; Toronto, Canada p. 1098-1109. [doi: 10.1016/b978-012088469-8.50096-6]
56. Zilio DC, Rao J, Lightstone S, Lohman GM, Storm AJ, Garcia-Arellano C, et al. DB2 Design Advisor: integrated automatic physical database design. 2004 Presented at: Proceedings of the 30th International Conference on Very Large Data Bases; August 31-September 3, 2004; Toronto, Canada p. 1087-1097. [doi: 10.1016/b978-012088469-8.50095-4]

Abbreviations

ED: emergency department
SQL: structured query language

Edited by C Lovis; submitted 06.02.21; peer-reviewed by V Rajan; comments to author 21.03.21; revised version received 25.03.21; accepted 14.04.21; published 27.05.21.

Please cite as:
Luo G
A Roadmap for Automating Lineage Tracing to Aid Automatically Explaining Machine Learning Predictions for Clinical Decision Support
JMIR Med Inform 2021;9(5):e27778
URL: https://medinform.jmir.org/2021/5/e27778
doi:10.2196/27778
PMID:34042600

Original Paper

Anomaly Detection Algorithm for Real-World Data and Evidence in Clinical Research: Implementation, Evaluation, and Validation Study

Vendula Churová[1,2], MSc; Roman Vyškovský[1,2], MSc; Kateřina Maršálová[1], MSc; David Kudláček[2], MSc; Daniel Schwarz[1,2], MSc, PhD

[1]Faculty of Medicine, Masaryk University, Brno, Czech Republic

[2]Institute of Biostatistics and Analyses, Ltd, Brno, Czech Republic

Corresponding Author:
Daniel Schwarz, MSc, PhD
Institute of Biostatistics and Analyses, Ltd
Postovska 3
Brno
Czech Republic
Phone: 420 604996753
Email: schwarz@biostatistika.cz

Abstract

Background: Statistical analysis, which has become an integral part of evidence-based medicine, relies heavily on data quality that is of critical importance in modern clinical research. Input data are not only at risk of being falsified or fabricated, but also at risk of being mishandled by investigators.

Objective: The urgent need to assure the highest data quality possible has led to the implementation of various auditing strategies designed to monitor clinical trials and detect errors of different origin that frequently occur in the field. The objective of this study was to describe a machine learning–based algorithm to detect anomalous patterns in data created as a consequence of carelessness, systematic error, or intentionally by entering fabricated values.

Methods: A particular electronic data capture (EDC) system, which is used for data management in clinical registries, is presented including its architecture and data structure. This EDC system features an algorithm based on machine learning designed to detect anomalous patterns in quantitative data. The detection algorithm combines clustering with a series of 7 distance metrics that serve to determine the strength of an anomaly. For the detection process, the thresholds and combinations of the metrics were used and the detection performance was evaluated and validated in the experiments involving simulated anomalous data and real-world data.

Results: Five different clinical registries related to neuroscience were presented—all of them running in the given EDC system. Two of the registries were selected for the evaluation experiments and served also to validate the detection performance on an independent data set. The best performing combination of the distance metrics was that of Canberra, Manhattan, and Mahalanobis, whereas Cosine and Chebyshev metrics had been excluded from further analysis due to the lowest performance when used as single distance metric–based classifiers.

Conclusions: The experimental results demonstrate that the algorithm is universal in nature, and as such may be implemented in other EDC systems, and is capable of anomalous data detection with a sensitivity exceeding 85%.

(JMIR Med Inform 2021;9(5):e27172) doi:10.2196/27172

KEYWORDS

clinical research data; real-world evidence; registry database; data quality; EDC system; anomaly detection

Introduction

Adherence to principles of evidence-based medicine has become the norm in the present-day clinical practice. Such principles include establishing proper guidelines built upon evidence derived from the best available clinical research. Therefore, high quality of input data is of utmost importance, because otherwise biased evidence may be generated, possibly resulting in harmful health decisions.

Clinical registries, defined as a systematic collection of clearly defined set of health and demographic data gathered from patients with specific health characteristics, represent one of many data sources available in health care [1]. The impact of clinical registries on quality of patient care taking account of a clinical research perspective is reviewed in [2], where monitoring health care delivery patterns and compliance with the evidence-based guidelines are also examined. The real-world data (RWD) collected in these registries may, in the context of postmarket research, provide much needed answers to questions unaddressed by existing randomized controlled trials. As patient populations participating in clinical trials are frequently low in numbers and rather homogenous and highly specific, further usage of such obtained data sets for the purpose of predicting medical treatment outcomes or future performance in the real-world, uncontrolled conditions has proved to be difficult [3].

The efficiency of data analysis is heavily dependent on data quality that has the potential to impact clinical research outcomes in both controlled clinical trials and postmarket surveillance practice represented mostly by noninterventional, observational studies and clinical registries. Data quality–related issues, such as high proportion of missing or inaccurate data, bring uncertainty to the final analytics, slow workflows, generate extra work, and thus increase research costs. A review and a generic framework for data quality in medical registries are given in [4], including some types and percentages of various data errors in a case study. In another scoping review [5], which focused on trauma registries, a call for standardization of classification, measurement, and improvement of data quality can be found. In order to mitigate data quality issues, various auditing techniques and monitoring strategies have been put in place (see the review in [6]). Besides extensive monitoring approaches including on-site visits and exhaustive source data verification, other effective risk-based monitoring methods have recently been implemented in the field of data quality assurance. These reduce monitoring costs by utilizing advanced statistical tools capable of identifying medical centers or clinics with atypical data patterns which might signify a quality problem [7]. The statistical concepts underpinning the central statistical monitoring (CSM) designed to detect fraud, that is, fabrication or falsification of data, were proposed 2 decades ago. The incidence of data fraud in clinical research is considered to be relatively low, yet difficult to estimate accurately [8].

Conventional data collection in clinical research involves recording data in paper case report forms (CRFs), followed by a double entry in a relational database. Continuous technological advancements in computer science, life sciences, and health care have given rise to the electronic data capture (EDC) systems, which have proved to be a more efficient [9] and also a cheaper [10] alternative to the paper data capture. EDC systems enable investigators to enter data directly into electronic CRFs (eCRFs) and study coordinators to oversee and control them in real time [11-13] even in multicenter research studies. EDC systems have become predominant because they are not only time- and cost-effective, but also contribute to quality assurance, as they allow data access to be controlled and all changes made to them using audit trail features to be traced. Moreover, they perform automatic edit checks designed to prevent invalid data from being entered [14] into a clinical registry, which is, however, hardly possible to be ruled out completely. When multiple variables need to be constrained by edit checks, the validation procedures, designed by data managers, may become too complicated and prone to error. The alert messages resulting from such complex edit checks may be unintelligible to clinical investigators, who still need to understand their factual content, as the validation procedures form an integral part of the eCRF.

Thus, there is still great potential for further improvements in ensuring high quality of data with the use of EDC systems. Integration with the aforementioned risk-based monitoring tools, such as CSM employing various outlier detection techniques, represents another automated approach to quality control. The review in [15] divides the outlier detection techniques, which have been used for data assurance in health care databases, into several categories: statistical, clustering, classification, nearest neighbor, and mixture models. It reveals that the statistical techniques are used frequently, whereas the other ones associated with data science and data mining are still little used in this context.

The viewpoint presented in [16] questions the benefits of a particular CSM technique which classified clinical sites as outlying based on the data inconsistency score calculated from thousands of statistical tests in a particular multicenter, postmarketing trial, and therefore dismissed the idea that trials could be conducted at lower costs.

This paper describes an algorithm based on machine learning designed to detect anomalous patterns in data created as a consequence of carelessness, systematic error, or intentionally by entering fabricated values. It focuses on the main concepts defining the anomaly detection algorithm and presents a particular EDC system demonstrating its successful implementation. The data sets collected by this system have been used in a number of clinical registries and serve here for pilot testing and calculations of anomaly detection rates. It is important to note that by anomalous data or an anomaly we understand an observation that does not conform to normally gathered data, where an observation refers to a single patient data record entering the detection algorithm.

Methods

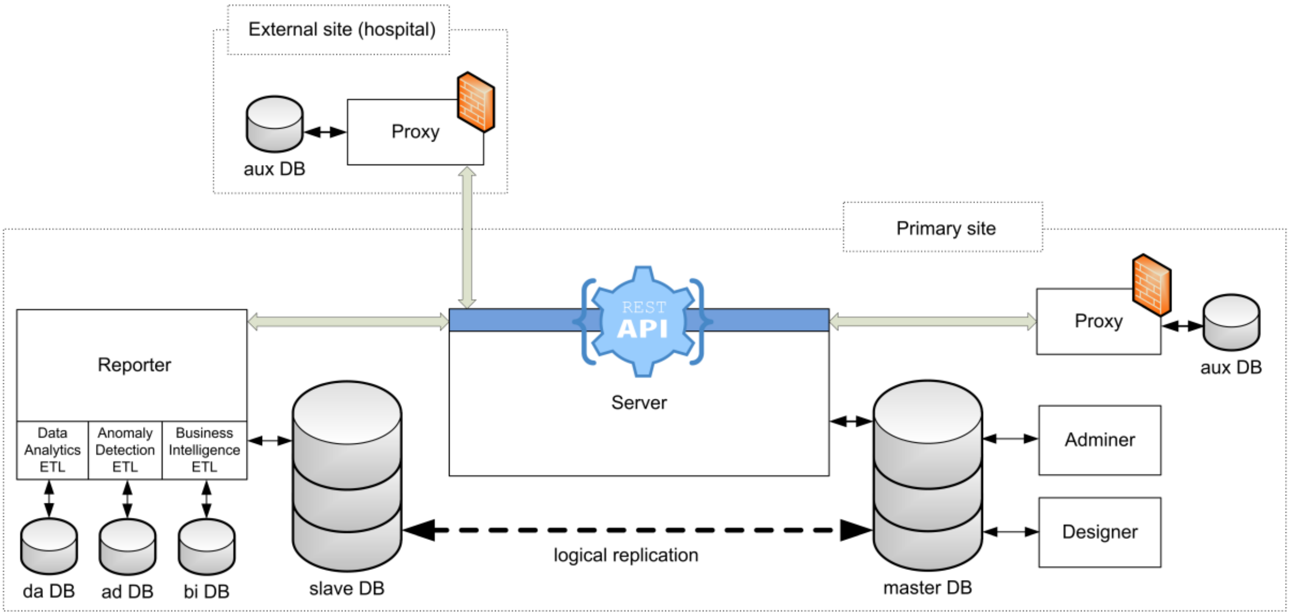
In order to fully implement an anomaly detection algorithm, which is data structure dependent, having an EDC system in place is essential. A thorough description representing such a system, including its architecture and data structure, is presented in this section.

EDC System and Its Data Structure

The EDC system utilized in this study referred to as the Clade-IS (Clinical Data Warehousing Information System) is a robust, modular, web-based software for data management and clinical trial management. It contains a huge amount of RWD from many clinical specialties, including neurology and psychiatry, that are readily available to be used for experimentation. The authors of this paper are engineers, data scientists, computer scientists, and data managers affiliated with the contract research institution where this EDC system has been developed and so they have a very good understanding of its data structures.

The system is composed of 5 mutually communicating components: proxy, server, adminer, designer, and reporter; see the architecture in Figure 1. The proxy, representing the user interface, propagates the user's activity, defined by requests made through a REST API (representational state transfer application interface) to the server, where the requests are processed. The server also stores and accesses registry data in a relational database and maintains data integrity. For example, the consistency and accuracy of data must be ensured throughout the transition between the components, as the format of the data varies according to its intended use, from its input, through storage, to extraction and reporting. It also ensures compliance with data access rules, which can be configured via users, groups, roles, and form statuses in the adminer. The next component, called the designer, represents a comprehensive form builder used by data managers to design eCRFs.

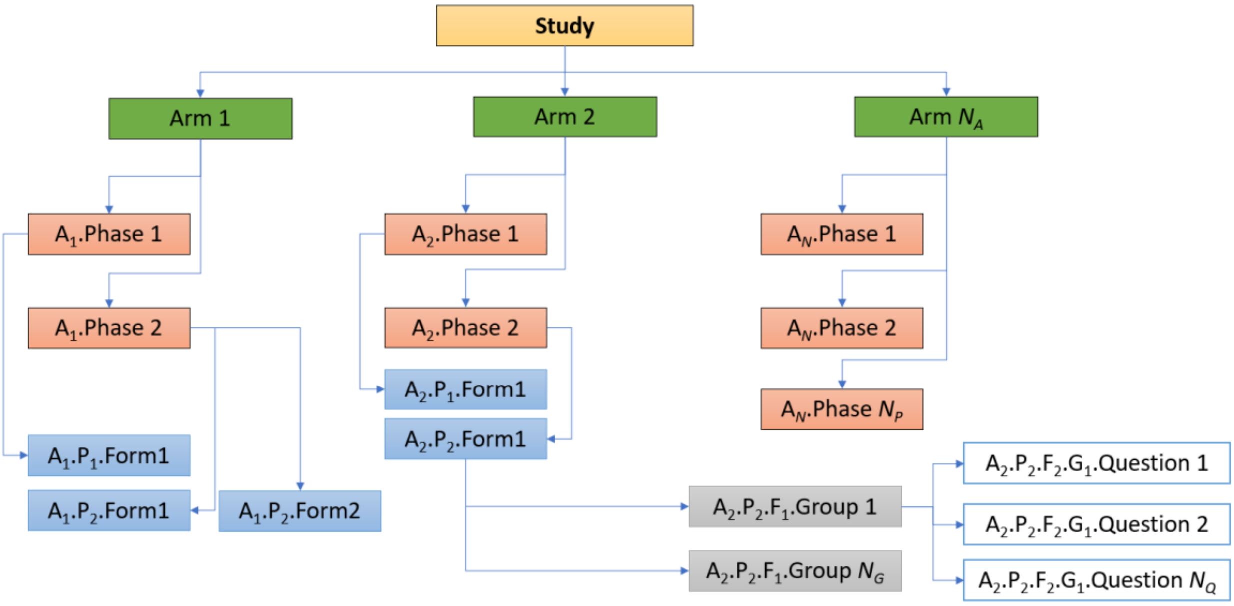
Figure 1. Architecture of the Clinical Data Warehousing Information System (Clade-IS) components and databases. The Server provides a representational state transfer application interface (REST API) for most operations including data storage. The Proxy represents a forwarded interface that transfers the user's activity to the Server. The Proxy can be optionally decentralized into a hospital or to another research facility. The Adminer and the Designer are used for configuring registry-specific permissions, designing electronic case report forms (eCRFs) and also for building and generating forms that are accessible to authenticated and authorized users. The Reporter is based on extract–transform–load (ETL) processes and serves for analytical and reporting purposes.



Finally, the reporter, serving as a toolkit for data analysts and data scientists, is a component based on the ETL (extract–transform–load) processes that facilitate data export and business intelligence. Besides the master database, which primarily serves for data storage operations, there are 4 other databases that the aforesaid components use for the following purposes: (1) the slave database is a logical replica of the master one performing all data extraction operations; (2) the proxy auxiliary database stores personal data gathered in research projects and studies outside the central repository in case that the centralized deployment of the Clade-IS is no longer possible under the General Data Protection Regulation (GDPR) on digital data; (3) the reporter-*ad* database serves for data export purposes; and (4) the reporter-*bi* database is used for business intelligence reporting.

The primary databases (master and slave) integrated into the Clade-IS are based on the entity–attribute–value (EAV) model, also known as the vertical database model, which is able to efficiently encode entities with sparse features. Such a functionality is directly applicable to clinical registries, as they typically contain plenty of available attributes describing an entity, but the number of attributes with assigned values is, once the data has been entered, rather low. The following data structures are used to build an eCRF: arm–phase–form–question group–question–answer, where a question–answer pair represents an attribute–value pair, respectively. The other structures represent entities in the EAV data model. Figure 2 serves to explain the meaning of the entities. The eCRF data are stored in JSON format; for instance, a single-answer question (Q10, Patient's age at diagnosis) is represented as *"Q10":{"value":63,"state":"done"}* and stored in a single cell in the database; see Multimedia Appendices 1 and 2 for data examples. The other database schemas differ depending on their specific purpose.

Figure 2. An example illustrating a structure of entities (arms, phases, forms, question groups) and attributes (questions) used for structuring electronic case report forms (eCRFs) in the Clinical Data Warehousing Information System (Clade-IS). Questions are logically grouped into question groups (eg, Demography question group, Comorbidities question group, etc), a form is composed of question groups (eg, Diagnosis form, Treatment form, etc), forms are grouped into phases (eg, Hospitalization forms phase, Follow-up forms phase, Quality of life forms phase, etc), and phases are grouped into arms which may represent different sub-populations of subjects in a study or a registry (eg, subjects diagnosed with affective disorders, schizophrenia, schizoaffective disorders and control subjects).



Anomaly Detection Algorithm

Anomalous data are identified by a scheduled script, built in the reporter component, that connects to the reporter-ad database, where it stores and accesses data in its own auxiliary tables. The main steps defining the detection algorithm are described in Figure 3.

The multidimensional nature of the detection algorithm requires that all eCRF questions be merged into 1 flat-wide table, where the rows represent the patients and the columns represent the individual variables (attributes) collected from all forms across the eCRF structure. In order for a single flat-wide table to be considered an appropriate analytical data set, each patient would need to be linked to any of the forms in a 1:1 relationship. In most registries, however, a patient is linked to his/her forms in a 1:*N* relationship, where *N* usually differs between patients. For instance, patient A records may comprise 1 patient form, 1 hospitalization form, 2 follow-up forms but, say, no quality-of-life investigation form, whereas patient B records may comprise 1 patient form, 1 hospitalization form, 3 follow-up forms, and 2 quality-of-life investigation forms. Merging all forms into a flat-wide table would result in misalignment of variables in columns. Even eCRFs with an extremely rigid structure and predefined number of form instances per patient may still produce meaningless column combinations in terms of temporal context of a patient's condition. To help overcome this problem, a concept of semiflattened tables is introduced here (Figure 4). The semiflattened tables consist of a "prefix" table, which is created by serializing all forms, allowing only a single instance to be run and 1 merged form that can be instantiated multiple times. This explains how N_{sw} semiflattened tables are created, where N_{sw} is the total number of all forms allowing multiple instances per patient. Therefore, the detection algorithm has to be run N_{sw} +1 times for the prefix table and for each semiflattened table independently. The rows in both the prefix and the semiflattened table contain variables of the following data types: string, text, integer, float, date, datetime, time, Boolean, and categorical variables. Because only numerical data are subjected to further analysis, data tables need to be preprocessed. There are 4 preprocessing steps in the algorithm: dropping, imputation, recoding, and normalization. First, variables for which the amount of missing data exceeds a preset percentage are dropped (excluded) from the table. Besides, all variables of string and text data types are dropped, as they represent only unimportant notes and comments irrelevant for this study. The remaining variables, which still have some missing values, are imputed using median that is calculated for each variable separately. In the next step, the non-numerical variables are recoded to numerical ones. The variables of date, time, and datetime data types are recoded to numerical values representing the number of seconds since 1.1.1600 00:00:01. The Boolean variables and the categorical variables are recoded in a way that each unique data item represents a different integer value (eg, "Female" 1 and "Male" 0). The ascending integer values are assigned according to a frequency of occurrence of unique data. The numerical variables holding integer and float data types do not require any recoding.

Figure 3. A scheme illustrating the anomaly detection algorithm and its links to the electronic data capture (EDC) system. The algorithm transforms registry raw data into semi-flattened tables which contain only meaningful combinations of variables in rows. The tables are preprocessed in four consecutive steps resulting in feature vectors from which one single centroid is computed. The distance between all data objects (feature vectors) and the centroid is measured using seven different distance metrics. The number of threshold-exceeding distances shows the strength of evidence of an anomaly. All anomalies are then subjected to post-hoc univariate tests to identify potentially problematic variables in the description of automatically generated electronic queries intended to be processed by data managers.

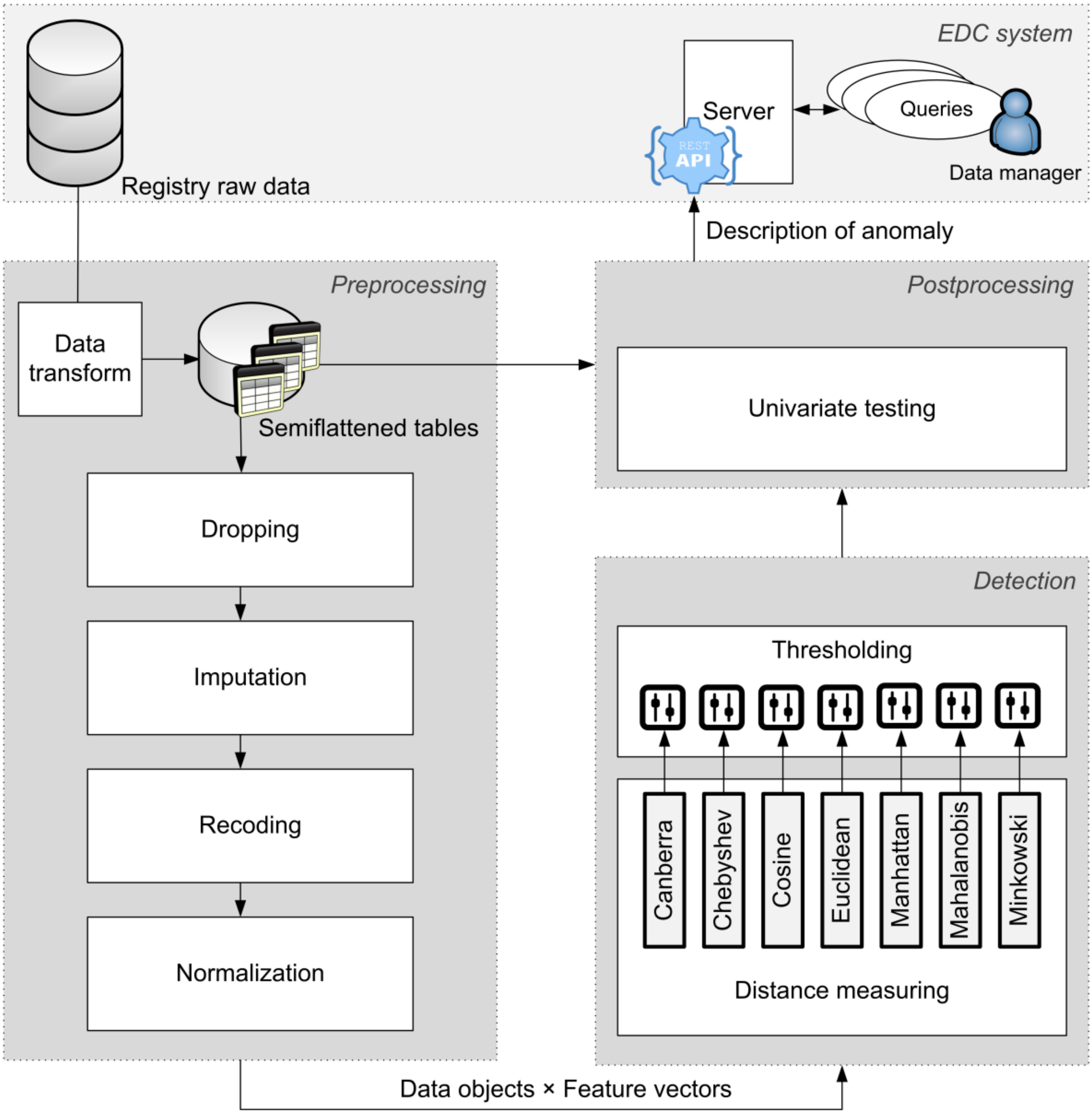
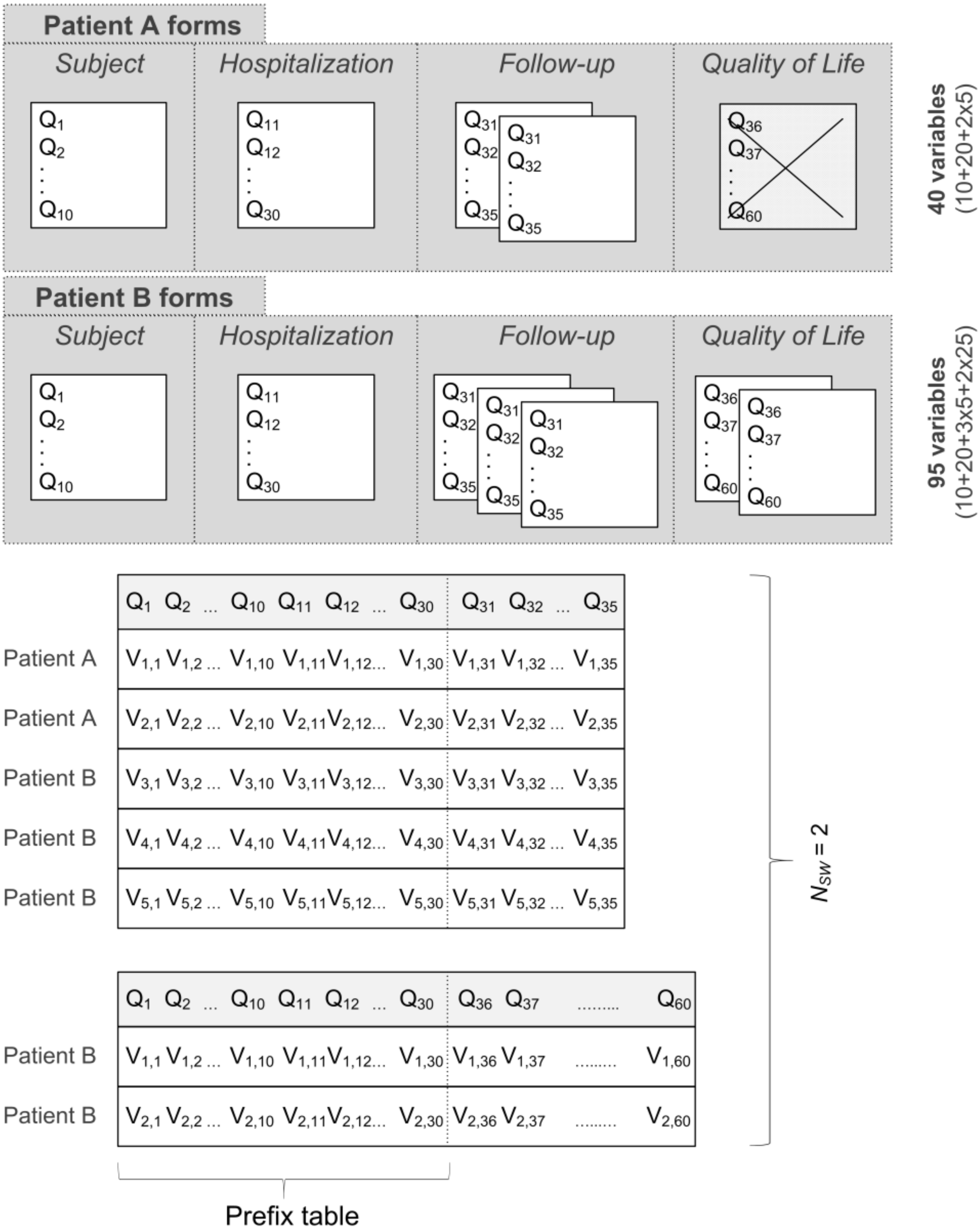


Figure 4. The concept of semi-flattened tables demonstrated on two patients‘ data. Two semi-flattened tables (N_{sw}=2) result here from two different repeating forms: Follow-up and Quality of Life. The other two forms: Subject and Hospitalization exist only in one single instance per patient, and thus all their variables (Q1, Q2,…,Q30) put together a prefix table. Multiple existence of the semi-flattened tables occurs with electronic case report forms (eCRFs) that allow multiple forms creation. Each instance of a repeating form is appended to the prefix table. This concept with semi-flattened tables assures that all values aligned in a column are related to the same variable.



	Q_1	Q_2	...	Q_{10}	Q_{11}	Q_{12}	...	Q_{30}	Q_{31}	Q_{32}	...	Q_{35}
Patient A	$V_{1,1}$	$V_{1,2}$	...	$V_{1,10}$	$V_{1,11}$	$V_{1,12}$	...	$V_{1,30}$	$V_{1,31}$	$V_{1,32}$	...	$V_{1,35}$
Patient A	$V_{2,1}$	$V_{2,2}$	...	$V_{2,10}$	$V_{2,11}$	$V_{2,12}$	...	$V_{2,30}$	$V_{2,31}$	$V_{2,32}$	...	$V_{2,35}$
Patient B	$V_{3,1}$	$V_{3,2}$	...	$V_{3,10}$	$V_{3,11}$	$V_{3,12}$	...	$V_{3,30}$	$V_{3,31}$	$V_{3,32}$	...	$V_{3,35}$
Patient B	$V_{4,1}$	$V_{4,2}$	...	$V_{4,10}$	$V_{4,11}$	$V_{4,12}$	...	$V_{4,30}$	$V_{4,31}$	$V_{4,32}$	...	$V_{4,35}$
Patient B	$V_{5,1}$	$V_{5,2}$	...	$V_{5,10}$	$V_{5,11}$	$V_{5,12}$	...	$V_{5,30}$	$V_{5,31}$	$V_{5,32}$	...	$V_{5,35}$

	Q_1	Q_2	...	Q_{10}	Q_{11}	Q_{12}	...	Q_{30}	Q_{36}	Q_{37}	....	Q_{60}
Patient B	$V_{1,1}$	$V_{1,2}$	...	$V_{1,10}$	$V_{1,11}$	$V_{1,12}$	...	$V_{1,30}$	$V_{1,36}$	$V_{1,37}$	....	$V_{1,60}$
Patient B	$V_{2,1}$	$V_{2,2}$	...	$V_{2,10}$	$V_{2,11}$	$V_{2,12}$	...	$V_{2,30}$	$V_{2,36}$	$V_{2,37}$	....	$V_{2,60}$

In the last preprocessing step, the data must be normalized because the variables may vary in orders of magnitude or units of measurement. At the very end of the preprocessing phase, the data table looks as follows: each row represents an observation with columns representing variables with acceptable proportion of missing data that are recoded to their numerical representations and subsequently (min–max) normalized to produce values between 0 and 1; see the example data before and after preprocessing in Table 1.

Once the fully automatic preprocessing phase is complete, the anomalous data are classified similarly to how it is done in

[15,17] using well-known clustering-based outlier detection techniques that also regard outliers as data objects not located in clusters of a data set. Here, only 1 cluster containing all data objects is created. Each object is described by a feature vector that takes the form of a row obtained from the preprocessed data table. The distance between a potential outlier and the cluster centroid is measured using 7 different distance metrics: Canberra (CAN), Chebyshev (CHEB), cosine (COS), Euclidean (EUC), Manhattan (MAN), Mahalanobis (MAH), and Minkowski (MINK). The aim of the proposed algorithm is not to perform a cluster analysis as the well-known k-means algorithm normally does. Instead, it seeks to find all data objects whose distance from a centroid is greater than a threshold differentiating anomalous records from the normal ones. The distance thresholds are calculated individually for each metric in 2 ways: (1) with a predefined percentile and (2) using the IQR rule, which sets the upper bound of the IQR multiplied by 1.5 and added to the third quartile. Data objects are identified as anomalous when at least one distance metric exceeds the minimum of both thresholds. With the predefined percentile, if the value is lower than the threshold specified by the IQR rule, the detection sensitivity can be increased, but usually at the expense of specificity. The strength of evidence of an anomaly is determined by the number of threshold-exceeding distance metrics.

The algorithm produces a table containing all detected anomalies represented by a patient identifier, the strength of evidence, and a list of potentially problematic variables identified using post-hoc univariate tests, which are different for normally and non-normally distributed variables. Afterward, a scheduled handler operating inside the reporter generates automatic queries which are of great concern to data managers, who are usually responsible for addressing them over the course of a study or a registry monitoring process.

Table 1. Example data before and after preprocessing. The index in rows represents a unique patient identifier. The column headings represent unique question identifiers—variable names encoding location in the study structure. The variables with a missing data rate of more than 20% were dropped. The other variables, which have an acceptable proportion of missing data, were imputed with median values. The data were subsequently recoded depending on the variable data type, and normalized to produce values in the interval [0, 1].[a]

Index	A0.P0.F2.G2.Q1	A0.P0.F2.G2.Q3	A1.P1.F7.G44.Q672	A1.P1.F7.G35.Q1443	A1.P1.F3.G3.Q6
Before preprocessing					
0001437	1947-01-23	Female	Yes	4	64
0001437	1947-01-23	Female	Yes	4	64
0001437	1947-01-23	Female	Yes	4	64
0001333	1941-06-24	Female	Yes	2	68
0001333	1941-06-24	Female	Yes	2	68
0001479	1948-11-03	Male	None	2	57
0001513	1950-03-26	Male	Yes	1	59
After preprocessing					
0001437	0.340432	1	0	1	0.657143
0001437	0.340432	1	0	1	0.657143
0001437	0.340432	1	0	1	0.657143
0001333	0.258807	1	0	0.5	0.714286
0001333	0.258807	1	0	0.5	0.714286
0001479	0.366453	0	1	0.5	0.557143
0001513	0.366453	0	0	0.25	0.585714

[a] A: study arm; P: study phase, where the related form is located at; F: form, where the question is located at; G: question group; Q: question.

Simulation of Data Anomalies and Performance Evaluation

In the context of this study, evaluation refers to an exploratory analysis designed to establish quantitative characteristics of anomaly detection performance of the algorithm built into the aforementioned EDC system. The performance evaluation is carried out using simulated anomalous data, which need to be artificially generated inside an anomaly-free data set, in order to obtain the ground-truth knowledge.

The simulated anomalies, that are generated in a wide format table, are subsequently preprocessed by dropping, imputation, and recoding, but not by normalization. First, a small percentage (1% by default) of all cells in the table being preprocessed is set as the number of values N_c intended to be changed. Second, a random number of patients is set as the number of anomalous data objects N_s intended to be generated. The ratio $N_v = N_c/N_s$ gives an approximate number of variables whose values need to be changed in order to transform a normal data object into an anomalous one. These changes are performed only on variables of the following data types: integer, float, date, time, and datetime. The values of normally distributed variables are transformed to a mean (6σ), whereas the values of non-normally distributed variables are transformed to random numbers from an interval formed by rather unusual values having a frequency of occurrence lower than 10% in a particular variable. Afterward, the Shapiro–Wilk test, able to discriminate between normal and non-normal distributions, is run. Every time an anomaly occurs, the automatic edit checks built into a given registry are triggered, assuring that the newly generated, anomalous data undergo the same validation procedures as if having been entered by a human investigator. At the end of the simulation, all the generated anomalous data objects are identified in terms of their position in the data table, either as original or as changed values.

Performance evaluation of the detection algorithm is carried out in 2 phases: (1) setting the best thresholds for each distance metric and (2) finding the best combination of the distance metrics. In the first phase, the receiver operating characteristic (ROC) curves are calculated for each individual distance metric by varying the threshold percentile value. The best threshold is then selected based on the C_1 criterion, which maximizes overall accuracy and Youden index [18], whereas the distance of the corresponding point on the ROC curve from the upper left corner *ULC_dist* is minimized:

$$C_1 = \textit{normalized (accuracy)}^2 + \textit{normalized (Youden index)}^2 - \textit{normalized (ULC\_dist)}^2 \quad \textbf{(1)}$$

where all 3 members of (1) are normalized to the interval (0, 1). All possible combinations of the distance metrics with the set threshold are then tested and the best one is determined by the C_2 criterion which is based on balanced accuracy but favors sensitivity over specificity:

$$C_2 = \textit{balanced\_accuracy} + \textit{sensitivity} = (TPR+TNR)/2 + TPR = (3TPR+TNR)/2 \quad \textbf{(2)}$$

where TPR and TNR stand for true positive rate and true negative rate, respectively.

Validation

In this study, validation refers primarily to repeatability verification which is performed as follows: all data from 2 different registries were subjected to expert review. As no problems were reported, these data sets could be used for evaluation and validation purposes. Once the detection algorithm is fully specified by the thresholds and the best combination of the distance metrics is identified by applying the 2-stage evaluation process to the first registry data, an independent data set from the second registry is used to validate the detection performance.

Results

EDC System Deployment

To date, the Clade-IS has been implemented in hundreds of clinical centers where it serves numerous research studies, mostly clinical registries and other RWD projects. Therefore, this EDC system contains millions of authentic records of different origin. Such a huge set of RWD made it possible to carry out anomaly detection using the designed detection algorithm whose performance was subsequently validated.

Five neuroscience-related registries were utilized here to investigate the possibility of deploying and using the aforementioned algorithm for automatic detection of anomalous data. The registries significantly differed in scope, that is, in research objectives, complexity of the eCRFs, duration, and also the number of patients involved (Table 2). While 2 out of 5 registries are sponsored by Masaryk University, 3 remaining registries belong to the neuromuscular section of the Czech Neurological Society, which did not allow their identification. For the sake of consistency, the names of all 5 registries are anonymized here.

Registry number 1 collects data on patients with myasthenia gravis, a rare, autoimmune disease affecting neuromuscular transmission. The registry serves to gather comprehensive information from as many patients as possible, covering the whole course of the disease and the response to treatment, in order to enhance development of new therapies and improve patient care. Registry number 2 collects data on patients diagnosed with any of the following neuromuscular diseases: Duchenne and Becker muscular dystrophy, spinal muscular atrophy, myotonic dystrophy, and facioscapulohumeral muscular dystrophy. The aim of the registry is to gather comprehensive information from as many patients with causal genetic defects as possible and thus to contribute to development of new treatments. Registry number 3 collects data on patients with spastic paresis caused by acquired brain injuries including a craniocerebral trauma, cerebral palsy, and central stroke. The aim of the registry is to develop visual analytics over the collected data to enhance decision-making processes related to physical and medical therapy at an individual patient level. Registry number 4 represents a longitudinal monitoring of patients with a cognitive impairment in the depressive phase of various affective disorders. The aim of the registry is to evaluate the diagnostic and prognostic potential of changes produced in brain morphology and function in patients with cognitive impairments and to investigate their impact on quality of a patient’s life and social functioning. Registry number 5 represents a 5-year, noninterventional, prospective follow-up study involving patients in the first episode of schizophrenia. The study aims to evaluate patients’ psychosocial needs in the early stages of the disease and also examine the effects of psychosocial interventions.

Table 2. Summary data presenting 5 neuroscience-related clinical registries powered by the Clade-IS[a] utilized to investigate the possibility of performing automatic detection of anomalous data.

Quantitative characteristic × registry characteristic	Registry number 1	Registry number 2	Registry number 3	Registry number 4	Registry number 5
Forms	4763	9372	13,711	214	67
Patients	1150	1649	405	33	29
Investigators	26	63	15	19	8
Sites	9	14	1	1	1
Years of study	5	9	1	4	5

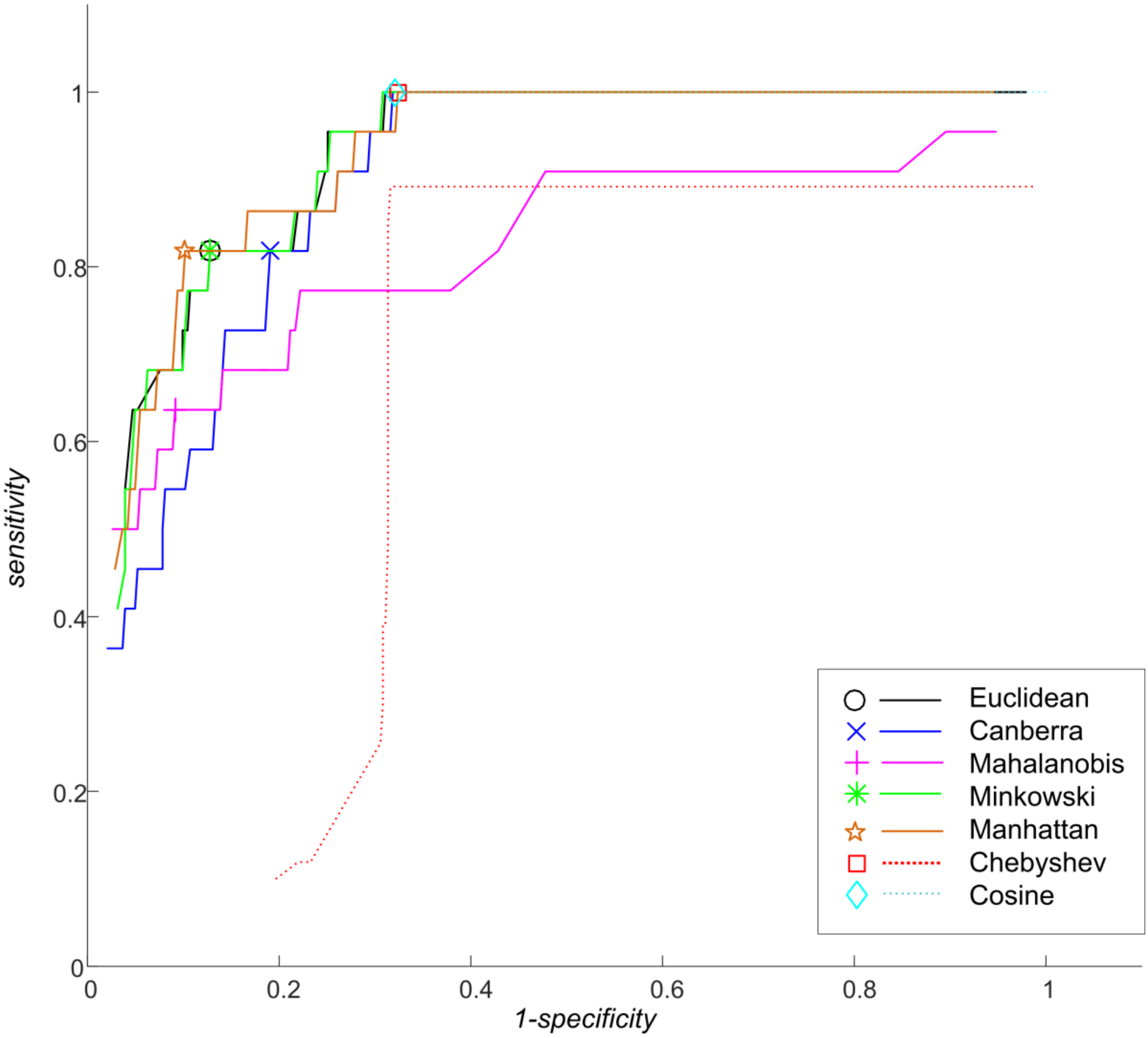
[a] Clade-IS: Clinical Data Warehousing Information System

Anomaly Detection Algorithm—Evaluation and Validation

The performance of the detection algorithm was evaluated using the data set extracted from Registry number 3 and then validated using the data set extracted from Registry number 5. The simulated anomalies were generated for each data set separately using the procedure described in the next section. The default number of cells to be changed was set to 1%, that is, 22 normal data objects were transformed to anomalous in the evaluation data set (Registry number 3) and 7 normal data objects were transformed to anomalous in the independent data set (Registry number 5).

Figure 5 shows the ROC curves calculated for single-distance metric–based classifiers whose function was to find the optimal thresholds. The worst detection performances achieved by individual metrics were those of the Chebyshev and cosine metrics. The results were consistent for both data sets (see the lowest values of C_1 in Table 3). These 2 distance metrics were, therefore, excluded from the subsequent ensemble classification. While a detailed ROC analysis was performed on the evaluation data set in order to find the best thresholds among 81 sampled and tested percentiles, in the case of the independent data set the performance characteristics were calculated only for 1 distance threshold setting.

Figure 5. ROC curves generated for single distance metric-based classifiers. The curves were created by connecting 81 points showing the true positive rate (sensitivity) and the false positive rate (1-specificity) calculated at various threshold settings ranging from 5th percentile distance to 95th percentile distance. The highlighted points indicate the thresholds with the best achieved detection performance as determined by the criterion C_1.



Once the thresholds were set, the best combination of the 5 remaining distance metrics was searched for. All possible ensembles were generated, first employing the distance metric–based classifiers individually, then combining 2, 3, 4 of them and, finally, all 5 classifiers were combined in 3 different scenarios: (1) the thresholds were set using the evaluation data set and so was the best combination of metrics (Table 4); (2) the thresholds were set using the evaluation data set whereas the combination of metrics was searched for using the independent data set (Table 5); (3) the thresholds and the combination of metrics were searched for using the independent data set only (Table 6).

The second scenario proved best in mimicking the real use of the detection algorithm, which would be required to detect anomalies in yet unseen data. Specifically, the best detection performance was achieved using the combination of Mahalanobis, Manhattan, and Canberra distance metrics, resulting in sensitivity of 85.7%, specificity of 72.7%, and balanced accuracy of 79.2%.

As anticipated, higher performance rates were achieved when the data sets were used separately for threshold setting and for searching the best combination of the distance metrics—as indicated in scenarios (1) and (2).

Table 3. The characteristics of detection performance achieved by individual single-distance metric–based classifiers using the evaluation data set and the independent data set.[a]

Distance metric	Percentile threshold	Sensitivity (%)	Specificity (%)	Accuracy (%)	Youden index	ULC_dist	C_1
Evaluation data set (Registry number 3)							
Canberra	77.5	81.82	80.94	80.99	0.628	0.263	1.481
Chebyshev	**64.0**	**100.00**	**67.62**	**69.38**	**0.676**	**0.324**	**1.384**
Cosine	**95.0**	**100.00**	**67.89**	**69.63**	**0.679**	**0.321**	**1.395**
Euclidean	86.0	81.82	87.21	86.91	0.690	0.222	1.760
Mahalanobis	88.0	63.64	90.86	89.38	0.545	0.375	1.423
Manhattan	86.0	81.82	89.82	89.38	0.716	0.208	1.882
Minkowski	83.5	81.82	87.21	86.91	0.690	0.222	1.760
Independent data set (Registry number 5)							
Canberra	77.5	57.14	86.36	79.31	0.435	0.450	1.177
Chebyshev	**64.0**	**28.57**	**50.00**	**44.83**	**–0.214**	**0.872**	**–0.595**
Cosine	**95.0**	**100.00**	**13.64**	**34.48**	**0.136**	**0.864**	**–0.517**
Euclidean	86.0	42.86	90.91	79.31	0.338	0.579	0.916
Mahalanobis	88.0	28.57	90.91	75.86	0.195	0.720	0.465
Manhattan	86.0	42.86	95.46	82.76	0.383	0.573	1.084
Minkowski	83.5	42.86	90.91	79.31	0.338	0.579	0.916

[a] The distance metrics with the lowest performance as determined by the criterion C_1 (highlighted in bold) were excluded from the subsequent classification.

Table 4. The characteristics of detection performance achieved by various ensembles of distance metric–based classifiers using the evaluation data set only. Ten combinations with the highest performance as determined by the criterion C^2 are displayed.[a]

Combination of distance metrics	Sensitivity (%)	Specificity (%)	Balanced accuracy (%)	Error (%)	Precision (%)	C_2
MAN[b] , CAN[c]	**95.46**	**82.51**	**88.98**	**16.79**	**23.86**	**1.844**
MAH[d], CAN	95.46	81.98	88.72	17.28	23.33	1.842
EUC[e], CAN	95.46	79.37	87.41	19.75	21.00	1.829
MINK[f], CAN	95.46	79.37	87.41	19.75	21.00	1.829
EUC, MINK, CAN	95.46	79.37	87.41	19.75	21.00	1.829
EUC, MAN, CAN	95.46	78.85	87.15	20.25	20.59	1.826
MAN, MINK, CAN	95.46	78.85	87.15	20.25	20.59	1.826
EUC, MAN, MINK, CAN	95.46	78.85	87.15	20.25	20.59	1.826
MAH, MAN, CAN	95.46	76.50	85.98	22.47	18.92	1.814
EUC, MAH, CAN	95.46	74.15	84.80	24.69	17.50	1.803

[a] The best performing combination of the distance metrics is highlighted in bold.

[b] MAN: Manhattan.

[c] CAN: Canberra.

[d] MAH: Mahalanobis.

[e] EUC: Euclidean.

[f] MINK: Minkowski.

Table 5. The characteristics of detection performance achieved by various ensembles of distance metric–based classifiers using the evaluation data set and the independent data set. Ten combinations with the highest performance as determined by the criterion C_2 are displayed.[a]

Combination of distance metrics	Sensitivity (%)	Specificity (%)	Accuracy (%)	Balanced accuracy (%)	Error (%)	Precision (%)	C_2
MAH[b] , MAN[c] , CAN[d]	**85.71**	**72.73**	**75.86**	**79.22**	**24.14**	**50.00**	**1.649**
CAN, EUC[e], MAH, MAN, MINK[f]	85.71	68.18	72.41	76.95	27.59	46.15	1.627
EUC, MAH, CAN	85.71	68.18	72.41	76.95	27.59	46.15	1.627
MAH, MINK, CAN	85.71	68.18	72.41	76.95	27.59	46.15	1.627
EUC, MAH, MAN, CAN	85.71	68.18	72.41	76.95	27.59	46.15	1.627
EUC, MAH, MINK, CAN	85.71	68.18	72.41	76.95	27.59	46.15	1.627
MAH, MAN, MINK, CAN	85.71	68.18	72.41	76.95	27.59	46.15	1.627
MAH, MAN	71.43	86.36	82.76	78.90	17.24	62.50	1.503
EUC, MAH	71.43	81.82	79.31	76.62	20.69	55.56	1.481
MAH, MINK	71.43	81.82	79.31	76.62	20.69	55.56	1.481

[a] The best performing combination of the distance metrics is highlighted in bold.

[b] MAH: Mahalanobis.

[c] MAN: Manhattan.

[d] CAN: Canberra.

[e] EUC: Euclidean.

[f] MINK: Minkowski.

Table 6. The characteristics of detection performance achieved by various ensembles of distance metric–based classifiers using the independent data set only. Ten combinations with the highest performance as determined by the criterion C_2 are displayed.[a]

Combination of distance metrics	Sensitivity (%)	Specificity (%)	Accuracy (%)	Balanced accuracy (%)	Error (%)	Precision (%)	C_2
CAN[b]	**85.71**	**86.36**	**86.21**	**86.04**	**13.79**	**66.67**	**1.718**
MAN[c], CAN	85.71	81.82	82.76	83.77	17.24	60.00	1.695
EUC[d], CAN	85.71	77.27	79.31	81.49	20.69	54.55	1.672
MINK[e], CAN	85.71	77.27	79.31	81.49	20.69	54.55	1.672
EUC, MAN, CAN	85.71	77.27	79.31	81.49	20.69	54.55	1.672
EUC, MINK, CAN	85.71	77.27	79.31	81.49	20.69	54.55	1.672
MAN, MINK, CAN	85.71	77.27	79.31	81.49	20.69	54.55	1.672
EUC, MAN, MINK, CAN	85.71	77.27	79.31	81.49	20.69	54.55	1.672
MAH[f], CAN	85.71	68.18	72.41	76.95	27.59	46.15	1.627
MAH, MAN, CAN	85.71	63.64	68.97	74.68	31.03	42.86	1.604

[a] The best performing combination of the distance metrics is highlighted in bold.

[b] CAN: Canberra.

[c] MAN: Manhattan.

[d] EUC: Euclidean.

[e] MINK: Minkowski.

[f] MAH: Mahalanobis.

Discussion

Anomaly Detection Context and Experiment Summary

In the era of EDC, it has become particularly difficult to process ever increasing data volumes in clinical registries. Data amount together with structural complexity of these databases make the task of anomaly detection, that may have a direct impact on the health care system, very demanding. Anomaly detection is an integral part of data analysis involving careful study of the identified anomalies and determination of their origin (data fraud, typing error, etc.), because it can significantly improve or negatively impact the subsequent analysis [19]. Even though anomalies tend to be misleading, they may carry valuable information [15,19]. For example, particular patient data may indicate that the patient has a different diagnosis than he/she is treated for, another anomalous pattern may indicate a new disease or reveal that investigators may have misinterpreted some questions. Therefore, detected anomalies need to be subjected to a careful assessment to mitigate the risk of losing valuable data by taking account of the unsuspicious ones, which may compromise the results and, as a consequence, lead to erroneous adjustments to clinical guidelines altering the current health care standards.

In this study, anomalous data were simulated and then detected. These operations were performed by a detection algorithm, whose detection performance was subsequently validated. The algorithm, running in a particular EDC system (Clade-IS), ends when automatic queries, whose function is to notify data managers and trial monitors of potentially anomalous data, have been generated. There are 2 key requirements which need to be met to implement such a detection algorithm in any EDC system successfully: (1) the ability of the system to create custom data views in the database and (2) the API able to react to data quality issues by its response (eg, a query generator). In the given settings, the accuracy was preferred over the algorithm execution time, so there was no need to optimize the algorithm for online use. A rapid online response is required when, for example, an intrusion activity is detected. This section presents a thorough description of (1) the detection algorithm running in the given EDC system and of (2) the actual validation experiments employing this algorithm together with the results interpretation. The findings are discussed here in terms of their validity and applicability (repeatability). The section is concluded with (3) a relevant literature review.

The tables loaded with raw data from 2 clinical registries were fed into the algorithm and a series of preprocessing steps, that is, dropping, imputation, recoding, and normalization, resulting in feature vectors were taken. These operations preceded the data simulation and the algorithm training. In the process, it was necessary to take account of data types which are supposed to be dropped as the given algorithm has not been devised to process all of them. Thus, some variables (texts, strings, and some raw JSON data) were excluded from further analysis. Although such an operation entails a significant information loss, it also represents a possible solution to the issues related to the "curse of dimensionality" (data reduction). One of the most difficult tasks was to handle the multiple instance forms supported by the Clade-IS. It means that the system allows not only forms limited to 1 instance per patient to be created, but also forms allowing more than 1 instance per patient. To tackle the problem of multiple forms filled in for a patient, the semiflattened tables have been introduced here. These tables aid in performing meaningful analysis and keep input data for each patient consistent, that is, with no blank attributes in places where data are expected. However, this approach has 2 limitations. First, the anomaly detection cannot be computed at the same time on all data available per patient. Instead, it is run separately on several semiflattened tables, each including data from 1 form structure instantiated multiple times. That said, anomalies resulting from a combination of forms with distinct form structures—the ones allowing multiple instances—could remain undetected. Second, information concerning data continuity (progression in time) that could possibly be filled in multiple forms created in a logical order was not investigated.

Principal Results

The anomaly analysis was performed by calculating the distance between a centroid and data points using several distance metrics. There were 2 aspects assessed and recorded: (1) the Boolean identifier able to identify whether a patient is anomalous or not, and (2) the strength of anomaly evidence as determined by the number of distance metrics that labeled a patient as anomalous. The presented procedure could be potentially further improved using the medoid instead of the centroid. Medoids are robust cluster members that tend to be less sensitive to distant observations than averaged centroids are. When an anomaly is detected, the patient is labeled using automatically generated queries, which enable a person in charge to check this anomaly directly in a web application. Thus, the individual query may serve as an opportunity to implement appropriate corrective and preventive actions enhancing data integrity on the part of data managers and may also notify trial monitors of incorrect data entry in the initial phases of the study. Here, 2 neuroscience-related data sets were used for the algorithm validation; the first one served for training, thus setting the appropriate values for the algorithm parameters (distance metric thresholds); the second data set was used to validate the algorithm detection capability. It means that the preset detection algorithm was applied to the test data and its repeatability and applicability were investigated in practice. The percentile-based threshold could be set in 2 ways: (1) based on expert knowledge in the field and (2) setting the thresholds based on data. When percentage is defined by an expert, the number of expected anomalies to be detected is rather predictable and as such assists project managers in budget and staffing allocations, making the anomaly checks procedure more effective. The second, from our perspective a more sophisticated approach, was proposed and carried out in this study. Specifically, each distance metric threshold was identified using a combination of overall accuracy (the ratio between correctly classified anomalies and normal data) and measurements based on ROC curves (Youden index and curve distance from the upper-left corner). The optimal percentile threshold defined for each metric then varied from 77.5% to 95.0%. Therefore, the optimal number of patients to be investigated ranges between 5.0% and 22.5%, in order to uncover as many potential anomalies as possible while no time is wasted on checking normal data.

The experiment was performed on data set number 3, where the optimal thresholds were found, and data set number 5, which served for set up testing. The best result for data set number 3 employing a single distance metric was achieved by the Manhattan metric that labeled 14.0% (57/405) of patients as suspected to be anomalous: C_1 (1.882), with sensitivity (81.8%) and specificity (89.8%) greater than 80.0%. When the thresholds (for each metric separately) were applied to the testing data set (number 5), the Canberra distance metric yielded the best results but sensitivity was very low compared with specificity: C_1 of 1.177, sensitivity of 57.1%, and specificity of 86.4%. This suggests that, despite the high number of patients labeled as suspected to be anomalous (meeting the low percentile threshold, 77.5% in the case of Canberra), it is still not guaranteed that anomalous data will be detected. The other metrics had sensitivity or specificity below 50.0% and so we conclude that a single metric is insufficient to detect an anomaly.

Significantly better results were obtained when the distance metrics were combined. In this scenario, a patient, whose data were labeled as anomalous by at least one metric, was considered as suspected to be anomalous. This suggests that the proposed method reveals more suspicious data than methods based on single metrics. Sensitivity results for data set number 3 (shown in Table 5) were better than those obtained by any single metric alone (shown in Table 3). These results further suggest that combining 2 metrics can significantly outperform sensitivity of any single metric. Because none of the single metrics had sensitivity greater than 82.0%, it also suggests that the distance metrics complement each other when combined because they label different patients as anomalous. As the best results achieved by combining the metrics yielded the same sensitivity (95.5%), specificity had the decisive power when assessing the results. The best combination observed was that of the Manhattan distance metric and the Canberra distance metric, with specificity of 82.5%, accuracy of 83.2%, and C_2 of 1.844. Combination of more than 2 metrics did not prove to be more efficient. In the case of validation data set number 5, the combination of 3 metrics (Mahalanobis, Manhattan, and Canberra) yielded the best results—sensitivity improved by almost 30% (85.7%), but specificity (approximately –14%; 72.7%) and overall accuracy (approximately –4%; 79.2%) were lower compared with the best single-metric performance (Canberra). These results also show that the threshold for the anomaly detection algorithm (method parameter), which has been set for 1 data set with a higher sample size (N), is possible to be applied to another data set, still producing satisfying results (Tables 4 and 5).

Limitations

It needs to be noted that the proposed algorithm for anomaly detection is limited by the following: (1) clinical registries are frequently incomplete, with large amounts of missing data (the data sets studied here are not an exception). Because a significant number of incomplete variables were removed in the data preprocessing phase (method parameter set by data manager), some valuable information could have been lost; (2) only quantitative data (or recoded qualitative data) can be further analyzed by the algorithm; (3) the detection algorithm is still computationally intensive and requires long detection times despite the fact that a large number of unfilled and unanalyzable variables had been removed, together with 2 distance metrics (Chebyshev and cosine). The most time-consuming part of the algorithm run is data preprocessing which lasts tens of minutes. The detection itself then takes less than 10 seconds per tested registry. The preprocessing and analysis are run at regular intervals and are not directly linked to the data entry action. The time required to detect an outlier since its onset is dependent on the interval, which is implementation dependent and usually set to 24 hours; and (4) the algorithm was validated on artificially simulated anomalies. Had the anomalous data been generated by field experts, such an approach could have proved effective in terms of expert-provided knowledge that would have ensured authenticity of the anomalies, making the validation more natural.

Comparison With Prior Work

There have been several research papers published on medical anomaly detection–related topics, as outlier detection has been widely applied in medical informatics for addressing different issues. According to the reviews, there are several detection techniques used in the field of medicine that can be divided into the following categories, listed in descending percentage order [15]: statistical (55.4%), clustering (15.2%), classification (12.5%), and nearest neighbor (ie, distance based, 8.9%), etc. As the numbers imply, statistic-based techniques tend to be used most frequently; however, it is well-known that the statistical assessment is not applicable to small sample sizes [20], therefore anomaly detection performed in small-scale studies or sites involving too few patients often leads to increasing the false-positive rate. For more reviews on anomaly detection in general, see [21] and for statistical monitoring process suggestions, we recommend [20]. That paper involved a multidisciplinary team of clinicians, statisticians, and data managers, who created a study-specific algorithm to flag the patients and sites with potentially fabricated data, which turned out to be fabricated and implanted in 7 sites, totaling 43 patients in 4 studies. Their algorithm for identifying sites with fabricated data achieved slightly lower results—except for 1 study, sensitivity and specificity were greater than 70%. In another research work [22], the authors combined k-means and isolation forest techniques, because the isolation forest–based methods are capable of finding anomalous patients that are not situated on the edge of a feature space. They, however, did not use ROC curves to define thresholds, but instead [23] split their data set into 2 subsets—first one consisted of only categorical variables and the second one of only continuous variables. This approach enabled them to work with each subset separately, searching (1) for infrequent category combinations in the subset with the categorical variables and (2) for distant objects defined by the cosine distance from the global mean in the subset with continuous variables. Then, they defined an anomaly score for each data object in both subsets. Adopting this approach, that is, splitting the data set into 2 subsets, could potentially improve our results. However, there would be many other parameters to be defined, such as the number of category combinations, that would complicate the setting of our anomaly detection algorithm. Estiri et al [24] used a different approach focusing on implausible rather than outlier data. The authors proposed a hierarchical k-means method to detect implausible observations, regardless of their values, that flag sparse clusters as anomalous, assuming no systematic errors. They also demonstrated that their clustering approach outperformed the conventional anomaly detection one that uses the standard deviation and Mahalanobis distance for identifying implausible laboratory data in the electronic health record. Although the authors consider the Mahalanobis distance to be standard, it did not work so well for us, especially in comparison with the other distance metrics (Table 3). To our knowledge, no paper presenting an EDC system with a built-in anomaly detection algorithm has been published to date.

Conclusions

We have proposed and described an algorithm for detection of anomalous data in clinical registries, which has been implemented in a particular EDC system. The algorithm has proved to be capable of detecting anomalous data with sensitivity greater than 85%. Besides, the detection results were satisfactory for preset parameter settings derived from a different data set which enabled the algorithm to be applied in practice. In future work, we will inspect queries in real-world settings in order to assess precision and usefulness of the proposed anomaly detector from the viewpoint of data managers and also users with other roles, such as site monitors and clinical investigators. Other ideas for further research include an investigation into expert-generated anomalies and finding ways to speed up the detection algorithm.

Acknowledgments

This research was funded by the Czech Ministry of Health, grant number 17-33136A. We thank Ms Katerina Žvaková for English language editing and also for all her comments that greatly improved the manuscript.

Authors' Contributions

DS was responsible for conceptualization of this study, acquisition of funding, and supervision of this work; RV carried out formal analysis; VC, RV, and DS were responsible for the methodology; VC, KM, and DK took part in the software implementation and analysis; KM, DK, and DS were responsible for visualization; VC, KM, and DS wrote the original draft; VC, RV, and DS reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The team of authors prepared this manuscript while being employed and working on 2 different projects at 2 institutions: (1) Neurominer: unveiling hidden patterns in neuroimaging data, a research project carried out at Masaryk University (MU), (2) Clade-IS: developing an original electronic data capture system for clinical research, a project focused on experimental development of software to be used for real-world-evidence projects at Institute of Biostatistics and Analyses Ltd (IBA), a spin-off company of MU.

Multimedia Appendix 1
Exemplification of the structure of patient A data (JSON).
[PDF File (Adobe PDF File), 83 KB - medinform_v9i5e27172_app1.pdf]

Multimedia Appendix 2
Exemplification of the structure of patient B data (JSON).
[PDF File (Adobe PDF File), 77 KB - medinform_v9i5e27172_app2.pdf]

References

1. Solomon DJ, Henry RC, Hogan JG, Van Amburg GH, Taylor J. Evaluation and implementation of public health registries. Public Health Rep 1991;106(2):142-150. [Medline: 1902306]
2. Hoque DME, Kumari V, Hoque M, Ruseckaite R, Romero L, Evans SM. Impact of clinical registries on quality of patient care and clinical outcomes: A systematic review. PLoS One 2017 Sep 8;12(9):e0183667 [FREE Full text] [doi: 10.1371/journal.pone.0183667] [Medline: 28886607]
3. Lu Z. Technical challenges in designing post-marketing eCRFs to address clinical safety and pharmacovigilance needs. Contemp Clin Trials 2010 Jan;31(1):108-118. [doi: 10.1016/j.cct.2009.11.004] [Medline: 19900576]
4. Arts DGT, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. J Am Med Inform Assoc 2002;9(6):600-611 [FREE Full text] [doi: 10.1197/jamia.m1087] [Medline: 12386111]
5. O'Reilly GM, Gabbe B, Moore L, Cameron PA. Classifying, measuring and improving the quality of data in trauma registries: A review of the literature. Injury 2016 Mar;47(3):559-567. [doi: 10.1016/j.injury.2016.01.007]
6. Houston L, Probst Y, Martin A. Assessing data quality and the variability of source data verification auditing methods in clinical research settings. Journal of Biomedical Informatics 2018 Jul;83:25-32. [doi: 10.1016/j.jbi.2018.05.010]
7. Timmermans C, Doffagne E, Venet D, Desmet L, Legrand C, Burzykowski T, et al. Statistical monitoring of data quality and consistency in the Stomach Cancer Adjuvant Multi-institutional Trial Group Trial. Gastric Cancer 2015 Aug 23;19(1):24-30. [doi: 10.1007/s10120-015-0533-9]
8. George SL, Buyse M. Data fraud in clinical trials. Clinical Investigation 2015 Feb;5(2):161-173. [doi: 10.4155/cli.14.116]

9. Walther B, Hossin S, Townend J, Abernethy N, Parker D, Jeffries D. Comparison of electronic data capture (EDC) with the standard data capture method for clinical trial data. PLoS One 2011;6(9):e25348 [FREE Full text] [doi: 10.1371/journal.pone.0025348] [Medline: 21966505]
10. van Dam J, Omondi Onyango K, Midamba B, Groosman N, Hooper N, Spector J, et al. Open-source mobile digital platform for clinical trial data collection in low-resource settings. BMJ Innov 2017 Jan 06;3(1):26-31. [doi: 10.1136/bmjinnov-2016-000164]
11. Gazali, Kaur S, Singh I. Artificial intelligence based clinical data management systems: A review. Informatics in Medicine Unlocked 2017;9:219-229. [doi: 10.1016/j.imu.2017.09.003]
12. Bruland P, Doods J, Brix T, Dugas M, Storck M. Connecting healthcare and clinical research: Workflow optimizations through seamless integration of EHR, pseudonymization services and EDC systems. International Journal of Medical Informatics 2018 Nov;119:103-108. [doi: 10.1016/j.ijmedinf.2018.09.007]
13. Zhengwu Lu. Electronic Data-Capturing Technology for Clinical Trials: Experience with a Global Postmarketing Study. IEEE Eng. Med. Biol. Mag 2010 Mar;29(2):95-102. [doi: 10.1109/memb.2009.935726]
14. Brandt CA, Argraves S, Money R, Ananth G, Trocky NM, Nadkarni PM. Informatics tools to improve clinical research study implementation. Contemporary Clinical Trials 2006 Apr;27(2):112-122. [doi: 10.1016/j.cct.2005.11.013]
15. Gaspar J, Catumbela E, Marques B, Freitas A. Systematic review of outliers detection techniques in medical data - preliminary study. In: Proceedings of the International Conference on Health Informatics - Volume 1: HEALTHINF, (BIOSTEC 2011). 2011 Presented at: HEALTHINF; 2011; Rome, Italy p. 575-582. [doi: 10.5220/0003168705750582]
16. Sakamoto J. A Hercule Poirot of clinical research. Gastric Cancer 2015 Oct 19;19(1):21-23. [doi: 10.1007/s10120-015-0555-3]
17. Lei D, Zhu Q, Chen J, Lin H, Yang P. Automatic K-Means Clustering Algorithm for Outlier Detection. Information Engineering and Applications. Lecture Notes in Electrical Engineering 2012;154:363-372. [doi: 10.1007/978-1-4471-2386-6_47]
18. Youden WJ. Index for rating diagnostic tests. Cancer 1950;3(1):32-35. [doi: 10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3]
19. Smiti A. When machine learning meets medical world: Current status and future challenges. Computer Science Review 2020 Aug;37:100280. [doi: 10.1016/j.cosrev.2020.100280]
20. Knepper D, Lindblad AS, Sharma G, Gensler GR, Manukyan Z, Matthews AG, et al. Statistical Monitoring in Clinical Trials: Best Practices for Detecting Data Anomalies Suggestive of Fabrication or Misconduct. Ther Innov Regul Sci 2016 Dec 30;50(2):144-154. [doi: 10.1177/2168479016630576]
21. Pimentel MA, Clifton DA, Clifton L, Tarassenko L. A review of novelty detection. Signal Processing 2014 Jun;99:215-249. [doi: 10.1016/j.sigpro.2013.12.026]
22. Karczmarek P, Kiersztyn A, Pedrycz W, Al E. K-Means-based isolation forest. Knowledge-Based Systems 2020 May;195:105659. [doi: 10.1016/j.knosys.2020.105659]
23. Koufakou A, Georgiopoulos M. A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes. Data Min Knowl Disc 2009 Nov 11;20(2):259-289. [doi: 10.1007/s10618-009-0148-z]
24. Estiri H, Klann JG, Murphy SN. A clustering approach for detecting implausible observation values in electronic health records data. BMC Med Inform Decis Mak 2019 Jul 23;19(1):1-16. [doi: 10.1186/s12911-019-0852-6]

Abbreviations

CAN: Canberra metric
CHEB: Chebyshev metric
Clade-IS: Clinical Data Warehousing Information System
COS: cosine metric
CSM: central statistical monitoring
eCRF: electronic case report form
EAV: entity–attribute–value
EDC: electronic data capture
ETL: extract–transform–load
EUC: Euclidean metric
GDPR: General Data Protection Regulation
JSON: JavaScript object notation
MAH: Mahalanobis metric
MAN: Manhattan metric
MINK: Minkowski metric
REST API: representational state transfer application interface
ROC: receiver operating characteristic
RWD: real-world data

Edited by G Eysenbach; submitted 14.01.21; peer-reviewed by J Majernik, D Surian, S Pranic; comments to author 04.02.21; revised version received 01.04.21; accepted 12.04.21; published 07.05.21.

Please cite as:
Churová V, Vyškovský R, Maršálová K, Kudláček D, Schwarz D
Anomaly Detection Algorithm for Real-World Data and Evidence in Clinical Research: Implementation, Evaluation, and Validation Study
JMIR Med Inform 2021;9(5):e27172
URL: https://medinform.jmir.org/2021/5/e27172
doi:10.2196/27172
PMID:33851576

Original Paper

An Algorithm (LaD) for Monitoring Childbirth in Settings Where Tracking All Parameters in the World Health Organization Partograph Is Not Feasible: Design and Expert Validation

Michael S Balikuddembe[1,2], MBChB, MMed; Peter K Wakholi[3], MSc, PhD; Nazarius M Tumwesigye[4], PhD, MSc, MA; Thorkild Tylleskar[1], MD, PhD, MA

[1]Center for International Health, University of Bergen, Bergen, Norway

[2]Division of Maternal and Foetal Medicine, Mulago Specialised Women and Newborn Hospital, Mulago Hospital, Kampala, Uganda

[3]School of Computing and Information Technology, Makerere University Kampala, Kampala, Uganda

[4]Department of Epidemiology and Biostatistics, Makerere University School of Public Health, Kampala, Uganda

Corresponding Author:
Thorkild Tylleskar, MD, PhD, MA
Center for International Health
University of Bergen
PO Box 7800
Bergen, 5020
Norway
Phone: 47 48074410
Email: Thorkild.Tylleskar@uib.no

Abstract

Background: After determining the key childbirth monitoring items from experts, we designed an algorithm (LaD) to represent the experts' suggestions and validated it. In this paper we describe an abridged algorithm for labor and delivery management and use theoretical case to compare its performance with human childbirth experts.

Objective: The objective of this study was to describe the LaD algorithm, its development, and its validation. In addition, in the validation phase we wanted to assess if the algorithm was inferior, equivalent, or superior to human experts in recommending the necessary clinical actions during childbirth decision making.

Methods: The LaD algorithm encompasses the tracking of 6 of the 12 childbirth parameters monitored using the World Health Organization (WHO) partograph. It has recommendations on how to manage a patient when parameters are outside the normal ranges. We validated the algorithm with purposively selected experts selecting actions for a stratified sample of patient case scenarios. The experts' selections were compared to obtain pairwise sensitivity and false-positive rates (FPRs) between them and the algorithm.

Results: The mean weighted pairwise sensitivity among experts was 68.2% (SD 6.95; 95% CI 59.6-76.8), whereas that between experts and the LaD algorithm was 69.4% (SD 17.95; 95% CI 47.1-91.7). The pairwise FPR among the experts ranged from 12% to 33% with a mean of 23.9% (SD 9.14; 95% CI 12.6-35.2), whereas that between experts and the algorithm ranged from 18% to 43% (mean 26.3%; SD 10.4; 95% CI 13.3-39.3). The was a correlation (mean 0.67 [SD 0.06]) in the actions selected by the expert pairs for the different patient cases with a reliability coefficient (α) of .91.

Conclusions: The LaD algorithm was more sensitive, but had a higher FPR than the childbirth experts, although the differences were not statistically significant. An electronic tool for childbirth monitoring with fewer WHO-recommended parameters may not be inferior to human experts in labor and delivery clinical decision support.

(JMIR Med Inform 2021;9(5):e17056) doi:10.2196/17056

KEYWORDS

algorithm; software validation; childbirth monitoring; WHO partograph

Introduction

From the late 20th century, there were concerted efforts to improve pregnancy outcomes, with the World Health Organization (WHO) partograph being the main labor monitoring tool used globally [1-3]. Increasing and easing of childbirth monitoring have been at the forefront of strategies for better maternal and newborn outcomes [4-6]. A spiraling increase in the number of caesarean sections due to prolonged labor led to research that challenged the cervical dilatation rates in the partograph [7-9]. Doubt arose on the validity of the partograph and intrapartum guidelines with calls for their re-evaluation [5,6,10]. Calls for more evidence-based care at birth led to increased research for more practical labor monitoring guidelines and tools [7,11-13].

In 2015, the American college of Obstetricians and the Society for Maternal-Fetal Medicine issued new guidelines on labor monitoring [14]. Later, the WHO released new recommendations on partograph use including calls for more research on the most appropriate paper-based or electronic tool to aid childbirth decision making [12]. Before any electronic decision support can be developed, an algorithm is needed outlining which decisions to take at each potential situation along the birth of a child. The algorithm is also preceded by a decision on which input variables to use is needed. Among the problems with the WHO partograph was a large number of variables to register and it was regarded as labor intensive and unpractical for low-resource settings [4,15]. We studied the labor monitoring tool expectations of childbirth experts in Africa to generate consensus on the most important parameters to monitor during birth in low-resource settings [16,17]. The findings included a reduction in the WHO-modified partograph items and several suggestions on changing the frequency of monitoring the labor items. The experts also expressed a need to adopt the recommendation for raising the starting point of the partograph from 4 cm of cervical dilatation.

In this paper, we describe the labor and delivery (LaD) algorithm, its development, and validation. In the validation we wanted to know if the algorithm is inferior, equivalent, or superior to human experts in recommending the necessary clinical actions during childbirth decision making.

Methods

Overview

We used the maternity experts' recommendations and literature findings to develop an alternative algorithm for labor and delivery monitoring (the LaD algorithm). We conducted a preliminary validation of its logic before fully implementing it. Because of lack of a gold standard against which to compare the logic, we compared it against opinions of experts in childbirth monitoring. Comparison of results from medical devices against experts is increasingly seen as the better alternative when no gold standard exists and decisions are highly dependent on opinions or anecdotal evidence [18-21].

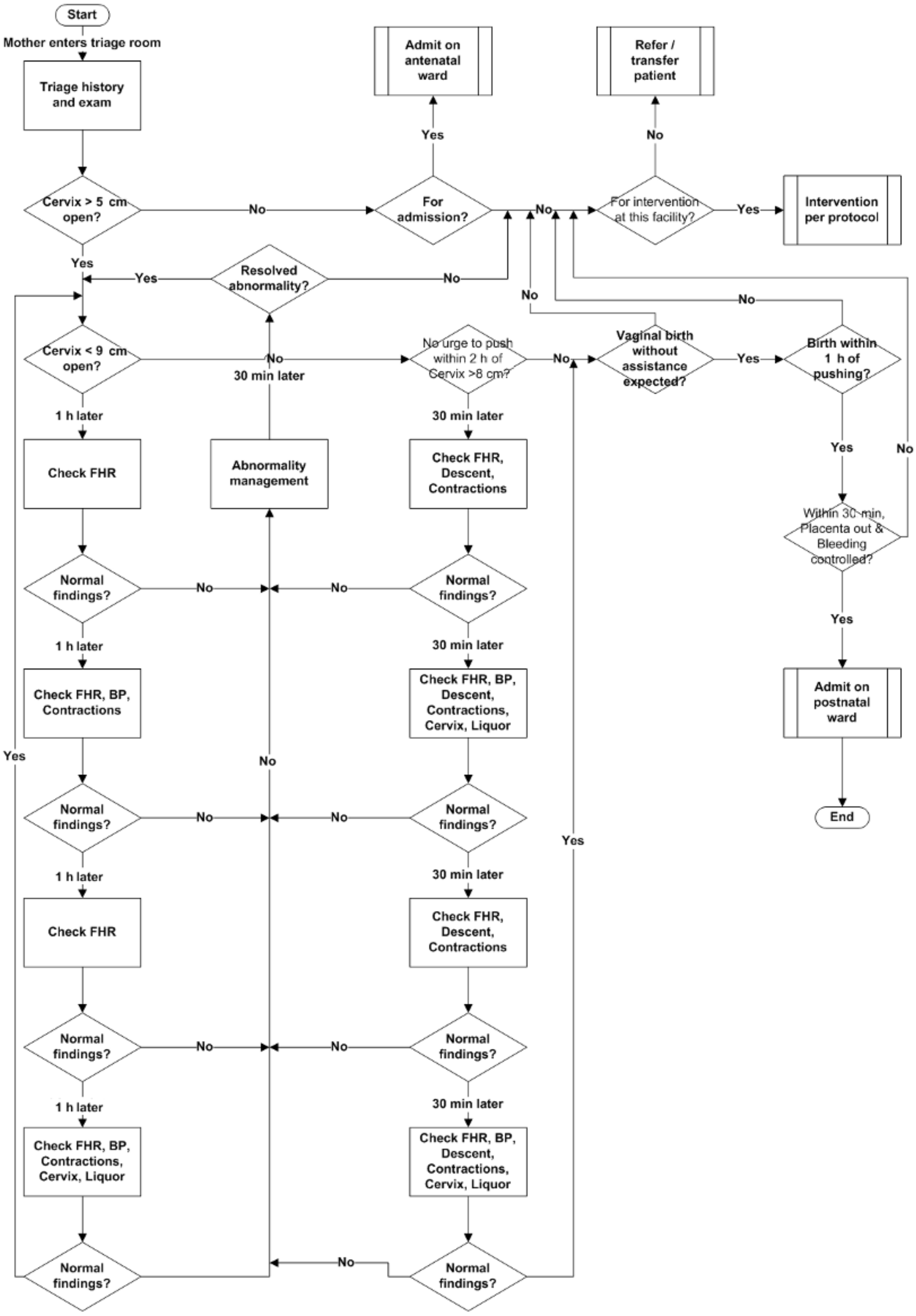
Development of the LaD Algorithm

From our earlier studies [16,17], the key parameters to monitor in childbirth were the fetal heart rate, amniotic fluid, cervical dilatation, uterine contractions, maternal blood pressure, and pulse rate. The suggested monitoring intervals ranged from 30 minutes to 4 hours. These are 6 of the 12 parameters in the WHO-modified partograph [22]. We used these recommendations and literature on the progress and outcomes of monitoring various childbirth items to generate a parameter list and monitoring intervals to include in the algorithm. Our main adjustment to the experts' suggestions was replacing the maternal pulse with second-stage tracking of the fetal station (a surrogate for fetal descent).

We used our acumen on labor progress and its monitoring process to draw the LaD algorithm using the Microsoft Visio 2013. It was revised to the layout shown in Figure 1. It shows the parameters to monitor at evidence-based time intervals.

For the algorithm to run on a computing device, we translated it into a recursive (ie, a problem is divided into subproblems of the same type. The solution to the problem is devised by combining the solutions obtained from the simpler parts of the problem) logic with 1152 possible patient scenarios and key decision support actions. Any abnormality in labor monitoring parameters is independently managed (as per local guidelines) and the final labor management decision is based on the success or failure in managing the subabnormalities. It is this logic that we validated with another group of childbirth experts.

Figure 1. The LaD algorithm for monitoring labor and delivery.



Validation of the LaD Algorithm

Between January and February 2019, 5 purposively selected childbirth care experts (E1, E2, E3, E4, and E5) independently answered a survey questionnaire covering 6 patient case scenarios (P1, P2, P3, P4, P5, and P6). The 5 experts had a mean experience of 17 years (SD 5.8 years) in medical practice and an obstetric career length ranging from 5 to 17 years (mean 10.6 years [SD 5.1 years]). Most worked in a teaching hospital, with their highest education level ranging from a master's degree to a Doctor of Philosophy (Table 1).

Table 1. Summary characteristics of experts who participated in validation.

Characteristics	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Mean (SD)
Experience as a doctor (years)	12	16	11	23	23	17 (5.8)
Experience as an obstetrician (years)	6	11	5	14	17	10.6 (5.1)
Number of times expert selected actions in the 5 scenarios (maximum 80)	36	34	44	35	46	39 (5.6)
Highest level of medical education	Master's degree	Master's degree	PhD candidate	Master's degree	PhD	—
Primary workplace	Military hospital	Medical school	Medical school	National hospital	Medical school	—

The case scenarios were taken from childbirth scenarios in the algorithm using stratified sampling. The cases were stratified using the amniotic fluid status into 3 strata: membranes intact, amniotic fluid clear, and amniotic fluid opaque or foul smelling. An online random number generator [23] was used to randomly select 2 cases from each stratum. The questionnaire had 15 labor-related conditions and 22 actions to consider. Each expert was allowed to select up to 16 of the 22 actions per case scenario, hence a maximum of 80 actions across 5 cases. The actions recommended by the algorithm for the study case scenarios were used to assess it.

We explained the survey procedure to the human experts before asking them to study the case scenarios and the accompanying set of possible actions to consider for managing each case. The expert would then recommend the most important actions for each case scenario given its conditions. The algorithm also recommended actions to the same cases based on results of an earlier study of a larger group of experts and literature. Experts in this study, however, were not aware of the algorithm nor other experts' action recommendations. They were invited to suggest possible modifications to the actions list for clarity and to provide better decision support for the case conditions.

We analyzed data to determine the unadjusted and weighted interexpert pairwise sensitivity [18,24], false-positive rates (FPRs), and reliability coefficients. Pairwise sensitivity was calculated for each pair of experts; for instance E3–E4 is the sensitivity of E4 with respect to E3 as reference. The sensitivity of the LaD algorithm versus each human reviewer (E–LaD) was also calculated to determine how LaD–human expert scores compare with interhuman expert pairwise (E–E) scores. FPRs were calculated for the unadjusted scores. The weight assigned to an action was determined by the number of experts that selected that action for a given case scenario. That is, an action weighed 1.0 if all 5 experts selected it as important, 0.6 if 3 selected it, and 0 if none selected it. Therefore, the weights were assigned after data entry. We compared the LaD algorithm scores with averages of the human pairwise scores for each case and across all scenarios. To rank the algorithm and human experts, we compared the lower border for the 95% CI of the mean sensitivity and the upper border of its mean FPR confidence interval with corresponding values for the experts. A larger number of the lower limit border for the sensitivity confidence interval and a smaller number of the upper limit for the FPR confidence interval meant a superior rank [21].

Results

Overview of Case Scenarios

A total of 5 of the 6 case scenarios were managed by all experts while the sixth was completed by 2 experts. The experts articulated that the noncompleted case was similar to another they had answered and saw no big difference in general management.

As indicated in Table 2, for the 5 case scenarios together, the experts selected an average of 39 actions of a possible 80. Across the experts, case scenarios 1 and 5 received most actions with an average of 11 each, whereas case scenario 2 needed the fewest actions at 5. From the unadjusted data, unlike the experts, the LaD algorithm had most actions for case scenario 3, but there was no difference in the weighted scores.

Table 2. Number of actions selected per patient (actual and adjusted values).

Evaluator	P1[a]		P2		P3		P4		P5	
	Action	Adjusted value	Action	Adjusted value	Action	Adjusted value	Action	Adjusted value	Action	Adjusted value
E1[b]	9	7.2	5	3.4	5	4	6	3.4	11	8
E2	8	6.8	5	3.4	8	4.8	5	3.6	8	6
E3	12	7.4	4	2.8	6	3.4	9	5.2	13	9.4
E4	11	7.8	5	2.6	7	3.6	4	1.4	8	5
E5	13	8.8	5	3.4	5	3.6	9	4.6	14	9.6
Mean (SD)	10.6 (2.1)	7.6 (0.8)	4.8 (0.4)	3.1 (0.4)	6.2 (1.3)	3.9 (0.6)	6.6 (2.3)	3.6 (1.5)	10.8 (2.8)	7.6 (2.0)
Labor and delivery algorithm (LaD)	8	4.6	7	3.4	12	5.4	7	3.8	8	5.6

[b]P: patient case scenario.

[a]E: expert.

Pairwise Sensitivity and FPRs for the Experts and the LaD Algorithm

The interrater pairwise sensitivity for the experts and the LaD algorithm is shown in Figure 2. The mean for unadjusted pairwise sensitivity among experts (E–E) for all cases was 57.2% (SD 7.86; 95% CI 47.4-67.0), whereas the weighted mean sensitivity was 68.2% (SD 6.95; 95% CI 59.6-76.8). The difference between these means was significant (SD 11.0; 95% CI 2.8-21.2, *P*=.01). With reference to the experts, the mean sensitivity scores of the LaD algorithm (E–LaD) were 62.6% (SD 17.01; 95% CI 41.5-83.7) and 69.4% (SD 17.95; 95% CI 47.1-91.7) before and after adjustment, respectively. The difference of 6.8 in E–LaD means the 95% CI of –14.9 to 28.5 was not statistically significant, *P*=.32). As shown in Figure 3, the weighted pairwise sensitivity for experts was significantly higher (*P*=.02) and closer to the LaD sensitivity than the unadjusted scores, especially when E4 was the reference expert. The algorithm was more sensitive than E1, E4, and E5, but less sensitive than E3.

For the 5 patient cases, the average FPR of experts ranged from 12% to 33% with a mean of 23.9% (SD 9.14; 95% CI 12.6-35.2), whereas that for the E–LaD ranged from 18% to 43% with a mean of 26.3% (SD 10.43; 95% CI 13.3-39.3). Table 3 shows that case 2 was an outlier (in left tail) for the expert-to-expert pairwise false-positive scores and case 3 was an outlier (right tail) for the expert-to-algorithm FPR scores.

Figure 2. The interrater pairwise sensitivity scores for the five cases.

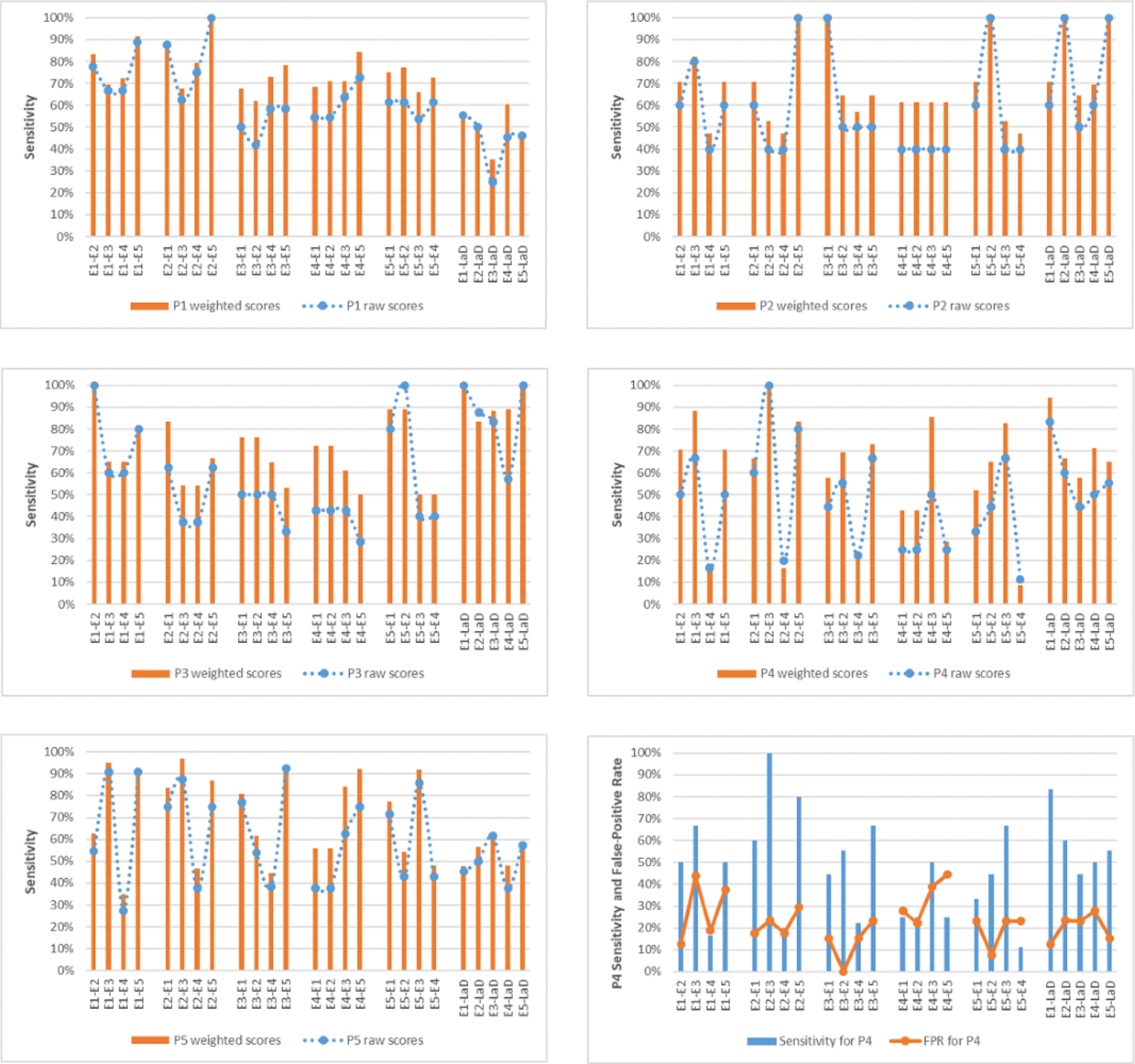


Figure 3. Comparison of the overall weighted and unadjusted pairwise sensitivity scores.

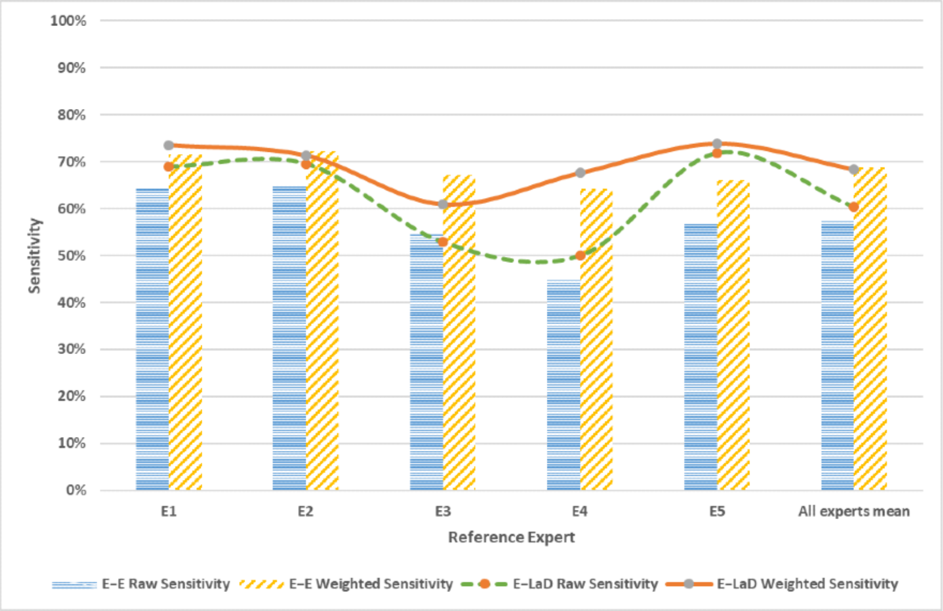


Table 3. Pairwise sensitivity and false-positive rates of experts and the labor and delivery (LaD) algorithm.

Comparisons	P1[a]	P2	P3	P4	P5	Mean (SD)	95% CI for mean	CI for difference of 2 means
E–E[b] pairwise sensitivity: unadjusted	65.9	56.5	55.0	45.6	62.8	57.2 (7.86)	47.4 to 67.0	–11.0 to 21.8
E–LaD pairwise sensitivity: unadjusted	44.4	74.0	85.6	58.7	50.3	62.6 (17.01)	41.5 to 83.7	
E–E pairwise sensitivity: weighted	75.9	67.2	68.6	57.3	71.9	68.2 (6.95)	59.6 to 76.8	–15.7 to 18.1
E–LaD pairwise sensitivity: weighted	49.3	80.8	92.1	71.0	54.0	69.4 (17.95)	47.1 to 91.7	
E–E pairwise FPR[c] for an action	33.1	12.2	18.3	23.2	32.9	23.9 (9.14)	12.6 to 35.2	–9.8 to 14.6
E–LaD pairwise FPR for an action	30.2	19.7	43.0	20.5	18.3	26.3 (10.43)	13.3 to 39.3	
E–E pairwise sensitivity for an action: unadjusted	65.9	56.5	55.0	45.6	62.8	57.2 (7.86)	47.4 to 67.0	2.8 to 21.2
E–E pairwise sensitivity for an action: weighted	75.9	67.2	68.6	57.3	71.9	68.2 (6.95)	59.6 to 76.8	
E–LaD pairwise agreement for an action: unadjusted	44.4	74.0	85.6	58.7	50.3	62.6 (17.01)	41.5 to 83.7	–14.9 to 28.5
E–LaD pairwise agreement for an action: weighted	49.3	80.8	92.1	71.0	54.0	69.4 (17.95)	47.1 to 91.7	

[a] P: patient case scenario.

[b] E: expert.

[c] FPR: false-positive rate.

Determining the Rank of LaD Algorithm Among Human Experts

The 95% CIs for the mean sensitivity scores of the algorithm and the human experts showed that the LaD algorithm had a higher upper limit before and after adjustment to the mean. By contrast, the lower limit of the confidence interval for the expert FPR mean was lower than that of the interval for the LaD algorithm mean. There was a positive correlation (mean $r^{\text{selection}}$ of 0.67 [SD 0.06]) in the actions that the expert pairs selected for the different patient cases (Table 4) with a reliability coefficient close to 1 (α=.91). This meant that the study experts agreed on most actions necessary for the cases and the same actions were likely to be recommended by these or other experts for the given patient scenarios.

Finally, we needed to know whether the differences in the mean sensitivity and FPR of the LaD algorithm and human experts were significant. The difference in mean sensitivity was 5.4 (95% CI –11.0 to 21.8) for the unadjusted means and 1.2 (95% CI –15.7 to 18.1, *P*=.57) for the weighted means. Because both intervals crossed the null, there was no statistical difference in the sensitivity of the experts and algorithm. In addition, the mean FPR of the experts and the algorithm was not significantly different with a 95% CI of –9.8 to 14.6 (*P*=.69).

On the basis of these sensitivity and false-positive scores, we found no statistical difference between the LaD algorithm and human experts recommending actions to childbirth monitoring health workers.

Table 4. Correlation and reliability coefficients of experts' choices of actions for the cases.

Comparisons	Selection correlation coefficient of actions selected by experts for each case, $r^{selection}$[a]					Reliability coefficient, α[b]
	P1[c]	P2	P3	P4	P5	
E1–E2[d]	0.857	0.706	0.913	0.686	0.722	
E1–E3	0.685	0.907	0.705	0.714	0.877	
E1–E4	0.703	0.538	0.685	0.275	0.443	
E1–E5	0.829	0.706	0.848	0.607	0.844	.925
E2–E1	0.857	0.706	0.913	0.686	0.722	
E2–E3	0.649	0.583	0.644	0.832	0.772	
E2–E4	0.751	0.538	0.625	0.267	0.511	
E2–E5	0.879	1.000	0.770	0.737	0.685	.923
E3–E1	0.685	0.908	0.705	0.713	0.876	
E3–E2	0.648	0.583	0.644	0.832	0.772	
E3–E4	0.720	0.593	0.629	0.445	0.613	
E3–E5	0.719	0.583	0.514	0.777	0.927	.919
E4–E1	0.703	0.538	0.760	0.275	0.443	
E4–E2	0.751	0.538	0.626	0.268	0.511	
E4–E3	0.720	0.593	0.629	0.445	0.613	
E4–E5	0.782	0.538	0.500	0.158	0.664	.861
E5–E1	0.829	0.706	0.843	0.607	0.844	
E5–E2	0.879	1.000	0.770	0.737	0.685	
E5–E3	0.719	0.583	0.514	0.777	0.926	
E5–E4	0.783	0.538	0.500	0.158	0.664	.922
Mean (SD)	0.757 (0.073)	0.669 (0.159)	0.687 (0.129)	0.550 (0.237)	0.706 (0.152)	.910 (0.027)

[a] $r^{selection}$ is an extension to Pearson r = square root of (sensitivity AB × selectivity AB), where selectivity RT = sensitivity TR. This is the selectivity for a test expert T against a reference expert R.

[b] α = kR/(1 + [k–1]R), where k is the number of experts and R is the average correlation of all expert pairs.

[c] P: patient case scenario.

[d] E: expert.

Discussion

Principal Findings

The search for an ideal labor and delivery monitoring decision support tool is ongoing and this study was one of many attempts to improve these tools. We have described the design of the LaD algorithm and validated it through comparison of its logic with human experts of childbirth monitoring. We found the algorithm to be equivalent in sensitivity and FPRs to experts with high reliability, that is, its action recommendations were close to the clinically "correct" ones. In clinical situations, lack of a gold standard against which to evaluate tools meant that traditional device validation tests were inappropriate and so childbirth experts had to act as the reference silver standard as in most types of clinical decision making [20,24]. Like Scheuer et al [21], we used the selection correlation coefficient $r^{selection}$ (an extension to Pearson r) because clinical experts often agree on many nonimportant actions for any patient case [18]. Most childbirth actions are not selected independent of one another, so our results would be less trustworthy if we used the kappa or pi statistics for measuring agreement. Likewise, we could not use Gwet AC statistic that necessitated assigning constant weights based on gold standards to parameters for all the patients, which would not be rational in our scenario [18,25]. The results of this study can be used to develop an abridged and more appropriate paper- or computer-based labor monitoring decision support tool that is less contentious than the WHO-modified partograph.

Limitations

The main limitations to this study are as follows: First, the low number of patient cases rated by the experts. Patient clinical scenarios have subtle or major differences that it would be virtually impossible to expect an exhaustive tool or validation. The cases were few, but each contained 22 actions to be considered; thus, the experts were not assessed on one case/condition per se, but on a sum of actions for each case and then the average of the 5. Therefore, the experts and algorithm were assessed on 110 instances summarized into 5 cases. This approach was similar to that used by Scheuer et al [21] who had over 5000 spike detections presented in under 40 scenarios [21]. Second, a total of 5 experts were not enough to tease out the effect of fast or slow actors when deciding to intervene in a clinical maternity setting. The fast actors tend to intervene too soon and so too much, whereas the slow actors intervene too late and so too late for good clinical outcomes, as expressed by Miller et al [26]. Third, the algorithm was based on suggestions from providers in low-income settings which are generally on the "too little, too late" side, and hence we expected the participants (E1, E4, and E5) to be more sensitive and E3 to be slower at acting. The strength of the pairwise sensitivity and the modified correlation we used is dampening the individual effect/biases of participants such that we still found no statistical differences between the group and the algorithm. Another limitation could have been our set of candidate actions from which experts selected. As was done by other researchers [18], we provided experts with candidate actions (from other studies) to encourage them to concentrate on relevant actions, but it could have hindered participants with divergent opinions from choosing their preferred actions. Following years of promoting the WHO partograph, some childbirth experts have got so engrained in it that any changes to its parameters could seem unfounded and unacceptable [27-29]. With these limitations in mind, we agreed that our validation results were preliminary and more assessments of the LaD algorithm would be done after its deployment and testing under more conditions.

Conclusions

The LaD algorithm was more sensitive but with a higher FPR than the childbirth experts, although the differences were not statistically significant. An electronic tool for childbirth monitoring with fewer parameters than those in the modified WHO partograph may not be inferior to human experts in labor and delivery clinical decision support.

Acknowledgments

We thank all participant experts for voluntarily sparing time to participate in the research.

Authors' Contributions

MB collected data and wrote the initial manuscript draft. All authors participated in proposal development and analysis of the data. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Mutebi A. Why are mothers and babies still dying? Voices from the community and service providers. BMJ Open. (Suppl 1) 2015;5:61. [doi: 10.1136/bmjopen-2015-forum2015abstracts.61]
2. Wall SN, Lee AC, Carlo W, Goldenberg R, Niermeyer S, Darmstadt GL, et al. Reducing intrapartum-related neonatal deaths in low- and middle-income countries-what works? Semin Perinatol 2010 Dec;34(6):395-407. [doi: 10.1053/j.semperi.2010.09.009] [Medline: 21094414]
3. World Health Organization. World Health Organization partograph in management of labour. The Lancet 1994 Jun;343(8910):1399-1404. [doi: 10.1016/s0140-6736(94)92528-3]
4. Ollerhead E, Osrin D. Barriers to and incentives for achieving partograph use in obstetric practice in low- and middle-income countries: a systematic review. BMC Pregnancy Childbirth 2014 Aug 16;14(1):281 [FREE Full text] [doi: 10.1186/1471-2393-14-281] [Medline: 25132124]
5. Jeffery J, Hewison A, Goodwin L, Kenyon S. Midwives' experiences of performing maternal observations and escalating concerns: a focus group study. BMC Pregnancy Childbirth 2017 Sep 02;17(1):282 [FREE Full text] [doi: 10.1186/s12884-017-1472-8] [Medline: 28865442]
6. Yang F, Bohren MA, Kyaddondo D, Titiloye MA, Olutayo AO, Oladapo OT, et al. Healthcare providers' perspectives on labor monitoring in Nigeria and Uganda: A qualitative study on challenges and opportunities. Int J Gynaecol Obstet 2017 Dec 07;139 Suppl 1:17-26. [doi: 10.1002/ijgo.12379] [Medline: 29218726]
7. Neal JL, Lowe NK. Physiologic partograph to improve birth safety and outcomes among low-risk, nulliparous women with spontaneous labor onset. Med Hypotheses 2012 Feb;78(2):319-326 [FREE Full text] [doi: 10.1016/j.mehy.2011.11.012] [Medline: 22138426]
8. Neal JL, Lowe NK, Patrick TE, Cabbage LA, Corwin EJ. What is the slowest-yet-normal cervical dilation rate among nulliparous women with spontaneous labor onset? J Obstet Gynecol Neonatal Nurs 2010 Jul;39(4):361-369 [FREE Full text] [doi: 10.1111/j.1552-6909.2010.01154.x] [Medline: 20629924]
9. Romijn MSc A, Muijtjens Dr Ir AMM, de Bruijne Dr MC, Donkers Dr HHLM, Wagner Prof Dr C, de Groot Prof Dr CJM, et al. What is normal progress in the first stage of labour? A vignette study of similarities and differences between midwives and obstetricians. Midwifery 2016 Oct;41:104-109. [doi: 10.1016/j.midw.2016.08.006] [Medline: 27586088]

10. Oladapo OT, Souza JP, Fawole B, Mugerwa K, Perdoná G, Alves D, et al. Progression of the first stage of spontaneous labour: A prospective cohort study in two sub-Saharan African countries. PLoS Med 2018 Jan 16;15(1):e1002492 [FREE Full text] [doi: 10.1371/journal.pmed.1002492] [Medline: 29338000]
11. Debilina R, Mandal D, Sahana R, Mandal A, Chowdhury P, Kundu T. ETD - expected ?time? of delivery: a new simple clinical tool for management of labour. Int J Med Appl Sci 2015;4(4):51-57.
12. World Health Organization. WHO Recommendations: Intrapartum Care for a Positive Childbirth Experience. Geneva, Switzerland: WHO; 2018.
13. Souza JP, Oladapo OT, Bohren MA, Mugerwa K, Fawole B, Moscovici L, WHO BOLD Research Group. The development of a Simplified, Effective, Labour Monitoring-to-Action (SELMA) tool for Better Outcomes in Labour Difficulty (BOLD): study protocol. Reprod Health 2015 May 26;12(1):49 [FREE Full text] [doi: 10.1186/s12978-015-0029-4] [Medline: 26006758]
14. No authors listed. Obstetric care consensus no. 1: safe prevention of the primary cesarean delivery. Obstet Gynecol 2014 Mar;123(3):693-711. [doi: 10.1097/01.AOG.0000444441.04111.1d] [Medline: 24553167]
15. Bedwell C, Levin K, Pett C, Lavender DT. A realist review of the partograph: when and how does it work for labour monitoring? BMC Pregnancy Childbirth 2017 Jan 13;17(1):31 [FREE Full text] [doi: 10.1186/s12884-016-1213-4] [Medline: 28086823]
16. Balikuddembe M, Wakholi P, Tumwesigye N, Tylleskär T. Midwifery providers' preferences for a childbirth monitoring tool in low-income health units in Uganda. In: Ugon A, editor. Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth. Amsterdam, The Netherlands: IOS Press; 2018:456-460.
17. Balikuddembe MS, Tumwesigye NM, Wakholi PK, Tylleskär T. Expert perspectives on essential parameters to monitor during childbirth in low resource settings: a Delphi study in sub-Saharan Africa. Reprod Health 2019 Aug 05;16(1):119 [FREE Full text] [doi: 10.1186/s12978-019-0786-6] [Medline: 31382989]
18. Wilson S, Harner R, Duffy F, Tharp B, Nuwer M, Sperling M. Spike detection. I. Correlation and reliability of human experts. Electroencephalography and Clinical Neurophysiology 1996 Mar;98(3):186-198. [doi: 10.1016/0013-4694(95)00221-9]
19. Trautner BW, Bhimani RD, Amspoker AB, Hysong SJ, Garza A, Kelly PA, et al. Development and validation of an algorithm to recalibrate mental models and reduce diagnostic errors associated with catheter-associated bacteriuria. BMC Med Inform Decis Mak 2013 Apr 15;13(1):48 [FREE Full text] [doi: 10.1186/1472-6947-13-48] [Medline: 23587259]
20. Sharma NK, Pedreira C, Centeno M, Chaudhary UJ, Wehner T, França LGS, et al. A novel scheme for the validation of an automated classification method for epileptic spikes by comparison with multiple observers. Clin Neurophysiol 2017 Jul;128(7):1246-1254 [FREE Full text] [doi: 10.1016/j.clinph.2017.04.016] [Medline: 28531810]
21. Scheuer ML, Bagic A, Wilson SB. Spike detection: Inter-reader agreement and a statistical Turing test on a large data set. Clin Neurophysiol 2017 Jan;128(1):243-250 [FREE Full text] [doi: 10.1016/j.clinph.2016.11.005] [Medline: 27913148]
22. Vidyashri Kamath C, Nagarathna G, Sharanya S. Documentation of the modified WHO partograph during labour in a South Indian tertiary care hospital. jemds 2015 Oct 12;4(82):14415-14421. [doi: 10.14260/jemds/2015/2050]
23. Furey E. Random number and letter set generator. URL: www.calculatorsoup.com/calculators/statistics/number-generator.php [accessed 2021-04-11]
24. Teal CR, Haidet P, Balasubramanyam AS, Rodriguez E, Naik AD. Measuring the quality of patients' goals and action plans: development and validation of a novel tool. BMC Med Inform Decis Mak 2012 Dec 27;12(1):152 [FREE Full text] [doi: 10.1186/1472-6947-12-152] [Medline: 23270422]
25. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol 2013 Apr 29;13(1):61 [FREE Full text] [doi: 10.1186/1471-2288-13-61] [Medline: 23627889]
26. Miller S, Abalos E, Chamillard M, Ciapponi A, Colaci D, Comandé D, et al. Beyond too little, too late and too much, too soon: a pathway towards evidence-based, respectful maternity care worldwide. The Lancet 2016 Oct;388(10056):2176-2192. [doi: 10.1016/s0140-6736(16)31472-6]
27. Fistula Care and Maternal Health Task Force: Revitalizing the Partograph: Does the Evidence Support a Global Call to Action?. URL: http://www.bettercaretogether.org/sites/default/files/resources/EngenderHealth-Fistula-Care-Partograph-Meeting-Report-9-April-12.pdf [accessed 2016-11-04]
28. Kushwah B, Singh AP, Singh S. The partograph: an essential yet underutilized tool. jemds 2013 Jun 14;2(24):4373-4379. [doi: 10.14260/jemds/849]
29. Cohen WR, Friedman EA. Perils of the new labor management guidelines. Am J Obstet Gynecol 2015 Apr;212(4):420-427. [doi: 10.1016/j.ajog.2014.09.008] [Medline: 25218127]

Abbreviations

FPR: false-positive rate
WHO: World Health Organization

Edited by C Lovis; submitted 18.11.19; peer-reviewed by J Shen, O Abiodun-Ojo; comments to author 19.01.20; revised version received 29.04.20; accepted 04.05.20; published 27.05.21.

Please cite as:
Balikuddembe MS, Wakholi PK, Tumwesigye NM, Tylleskar T
An Algorithm (LaD) for Monitoring Childbirth in Settings Where Tracking All Parameters in the World Health Organization Partograph Is Not Feasible: Design and Expert Validation
JMIR Med Inform 2021;9(5):e17056
URL: https://medinform.jmir.org/2021/5/e17056
doi:10.2196/17056
PMID:34042599

Original Paper

Automating Stroke Data Extraction From Free-Text Radiology Reports Using Natural Language Processing: Instrument Validation Study

Amy Y X Yu[1], MD; Zhongyu A Liu[1], MD; Chloe Pou-Prom[2], MSc; Kaitlyn Lopes[1], BSc; Moira K Kapral[3], MD; Richard I Aviv[4], MBChB; Muhammad Mamdani[5], MA, MPH, PharmD

[1]Department of Medicine (Neurology), University of Toronto – Sunnybrook Health Sciences Centre, Toronto, ON, Canada

[2]Unity Health Toronto, Toronto, ON, Canada

[3]Department of Medicine (General Internal Medicine), University of Toronto – University Health Network, Toronto, ON, Canada

[4]Department of Radiology, Division of Neuroradiology, University of Ottawa, Ottawa, ON, Canada

[5]Department of Medicine, Unity Health Toronto, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Amy Y X Yu, MD
Department of Medicine (Neurology)
University of Toronto – Sunnybrook Health Sciences Centre
2075 Bayview Avenue
Toronto, ON, M4N 3M5
Canada
Phone: 1 416 480 6100 ext 4866
Fax: 1 416 480 5753
Email: amyyx.yu@utoronto.ca

Abstract

Background: Diagnostic neurovascular imaging data are important in stroke research, but obtaining these data typically requires laborious manual chart reviews.

Objective: We aimed to determine the accuracy of a natural language processing (NLP) approach to extract information on the presence and location of vascular occlusions as well as other stroke-related attributes based on free-text reports.

Methods: From the full reports of 1320 consecutive computed tomography (CT), CT angiography, and CT perfusion scans of the head and neck performed at a tertiary stroke center between October 2017 and January 2019, we manually extracted data on the presence of proximal large vessel occlusion (primary outcome), as well as distal vessel occlusion, ischemia, hemorrhage, Alberta stroke program early CT score (ASPECTS), and collateral status (secondary outcomes). Reports were randomly split into training (n=921) and validation (n=399) sets, and attributes were extracted using rule-based NLP. We reported the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and the overall accuracy of the NLP approach relative to the manually extracted data.

Results: The overall prevalence of large vessel occlusion was 12.2%. In the training sample, the NLP approach identified this attribute with an overall accuracy of 97.3% (95.5% sensitivity, 98.1% specificity, 84.1% PPV, and 99.4% NPV). In the validation set, the overall accuracy was 95.2% (90.0% sensitivity, 97.4% specificity, 76.3% PPV, and 98.5% NPV). The accuracy of identifying distal or basilar occlusion as well as hemorrhage was also high, but there were limitations in identifying cerebral ischemia, ASPECTS, and collateral status.

Conclusions: NLP may improve the efficiency of large-scale imaging data collection for stroke surveillance and research.

(JMIR Med Inform 2021;9(5):e24381) doi:10.2196/24381

KEYWORDS

stroke; diagnostic imaging; data extraction; natural language processing; neurovascular; imaging; stroke surveillance; surveillance

Introduction

Stroke is a leading cause of death and disability [1]. Neuroimaging study findings inform treatment and prognosis. For example, recent clinical trials have demonstrated the efficacy of endovascular thrombectomy, a mechanical clot-retrieval procedure, in improving functional outcomes in patients with acute ischemic stroke and proximal large vessel occlusion [2-5]. Data on efficacy of this procedure in patients with distal or smaller vessel occlusion are currently lacking. Although large health administrative databases have information on whether a stroke was ischemic or hemorrhagic, detailed neuroimaging findings are usually found in narrative diagnostic imaging reports and obtained through resource-intensive manual chart abstractions [6,7].

The lack of population-based neuroimaging data limits the ability to characterize the prevalence of large vessel occlusion. A recent meta-analysis of cohort studies of patients with ischemic stroke found that the prevalence of large vessel occlusion ranged widely, from 13% to 52% [8], suggesting that smaller cohort studies can be vulnerable to selection bias. Therefore, automating the extraction of information on vessel occlusion from diagnostic imaging reports is needed for population-based disease surveillance and clinical research.

Natural language processing (NLP) can convert large amounts of free-text data into structured data and has been used to extract information on stroke type and location from diagnostic imaging reports [9-11]. However, its ability to characterize vascular occlusions is not well understood. We aimed to determine the accuracy of an NLP tool [12] in identifying the presence and location of vascular occlusions and other stroke-related attributes from neuroimaging reports of computed tomography (CT), CT angiography (CTA), and CT perfusion (CTP) scans. We hypothesized that an NLP tool can identify large vessel occlusion with high accuracy.

Methods

Manual Chart Abstraction

We obtained full free-text reports of 1320 consecutive stroke protocol imaging studies comprising CT, CTA, and CTP imaging of the head and neck performed between October 2017 and January 2019 at a university-affiliated comprehensive stroke center that provides consultation for endovascular thrombectomy to a catchment area of 2.5 million people. A stroke specialist and a trained research assistant manually extracted stroke-related attributes from the reports. The primary outcome was the presence of large vessel occlusion defined as occlusion in the M1 segment of the middle cerebral artery (MCA-M1) or A1 segment of the anterior cerebral artery (ACA-A1) with or without involvement of the carotid terminus because occlusion at these sites is treatable with endovascular thrombectomy. We chose this as the primary outcome because patients with this type of occlusion can be treated with endovascular thrombectomy. Isolated intracranial internal carotid artery occlusion was not categorized as large vessel occlusion in this study because the effectiveness of endovascular thrombectomy has not been shown in this population [13].

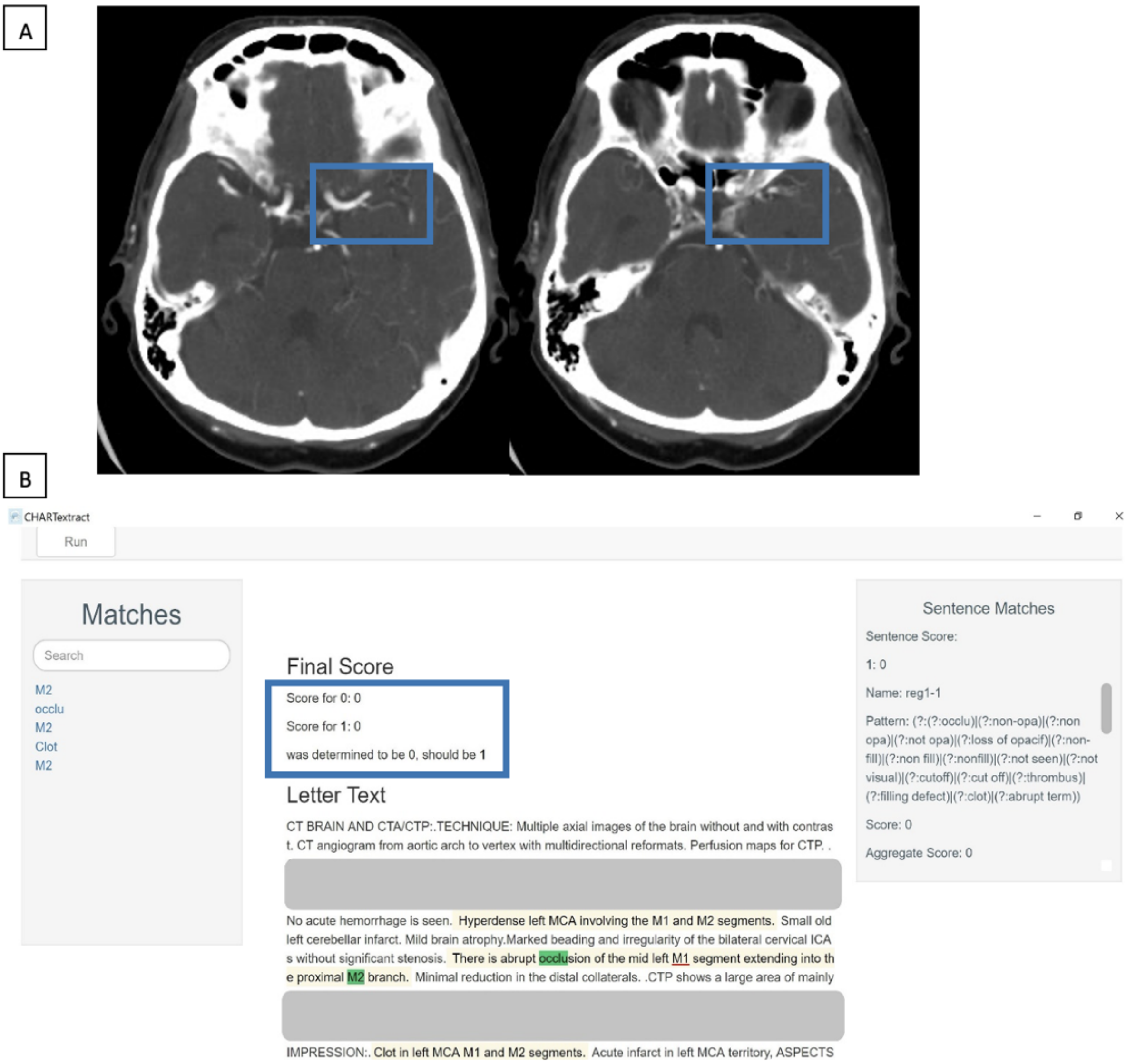
Secondary outcomes included (1) the presence of cerebral ischemia, (2) Alberta stroke program early CT score (ASPECTS) [14], (3) the presence of any intracranial hemorrhage, (4) distal anterior circulation occlusion defined as occlusion in the middle or anterior cerebral arteries in the M2 or A2 segments or beyond, (5) basilar occlusion, and (6) qualitative measure of collateral status (ie, good, intermediate, or poor). The manually extracted data were considered the reference standard. Duplicate chart abstraction on 200 charts showed that the inter-rater reliability was >96% for all attributes except for the presence of cerebral ischemia for which it was 80%. We randomly split the reports into training (n=921) and validation (n 399) sets.

CHARTextract NLP Tool

NLP rule sets for stroke attribute extraction from free-text diagnostic imaging reports were created using CHARTextract version 0.3.2, freely available online [12]. CHARTextract is a rule-based information extraction tool that relies on regular expressions and works at the sentence level to identify word patterns. We opted to use a rule-based approach due to the small sample size and the availability of domain experts to develop and refine the rules.

We created information extraction pipelines by using an iterative process where each rule was assigned a weight by the end-user in the training set. For example, if a report contains the text "presence of middle cerebral artery occlusion…," the system's estimate of the probability of a large vessel occlusion increases; however, if a report contains the text "no evidence of…," it will lower the system's estimate of the probability. As shown in Figure 1, the tool displays the discrepancies between the chart abstractor label and the tool's prediction, thus allowing for rapid iterative refinement of the rules by the end user. Rules were developed for each attribute through an iterative process by the end-user (ZL, AY, and CP) by using the training set that was validated in the validation set. For the presence of large vessel occlusion (our primary outcome), we also recorded whether the discrepancy between the chart abstractor and the NLP tool was due to abstractor or tool error. The rules thus developed are shown in Multimedia Appendix 1.

Figure 1. Example 1 of a discrepancy between the chart abstractor and CHARTextract tool output. (A) Computed tomography angiography scan showing loss of opacification in the left middle cerebral artery, involving the left M1 segment and extending into the M2 segment. (B) CHARTextract tool output: the chart abstractor labeled that large vessel occlusion was present, but the CHARTextract tool determined this attribute to be absent. The rules were revised to reflect that occlusion involving the "M1 segment" should be considered a large vessel occlusion even if the terms "MCA" or "middle cerebral artery" were absent.



Final Score

Score for 0: 0

Score for 1: 0

was determined to be 0, should be **1**

Letter Text

CT BRAIN AND CTA/CTP:.TECHNIQUE: Multiple axial images of the brain without and with contras t. CT angiogram from aortic arch to vertex with multidirectional reformats. Perfusion maps for CTP. .

[redacted]

No acute hemorrhage is seen. Hyperdense left MCA involving the M1 and M2 segments. Small old left cerebellar infarct. Mild brain atrophy.Marked beading and irregularity of the bilateral cervical ICA s without significant stenosis. There is abrupt occlusion of the mid left M1 segment extending into th e proximal M2 branch. Minimal reduction in the distal collaterals. .CTP shows a large area of mainly

[redacted]

IMPRESSION:. Clot in left MCA M1 and M2 segments. Acute infarct in left MCA territory, ASPECTS

Sentence Matches

Sentence Score:

1: 0

Name: reg1-1

Pattern: (?:(?:occlu)|(?:non-opa)|(?:non opa)|(?:not opa)|(?:loss of opacif)|(?:non-fill)|(?:non fill)|(?:nonfill)|(?:not seen)|(?:not visual)|(?:cutoff)|(?:cut off)|(?:thrombus)|(?:filling defect)|(?:clot)|(?:abrupt term))

Score: 0

Aggregate Score: 0

Statistical Methods

The stroke-related attributes identified by the NLP tool, CHARTextract version 0.3.2, were compared to the reference standard. The sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated using this tool.

Ethics Approval

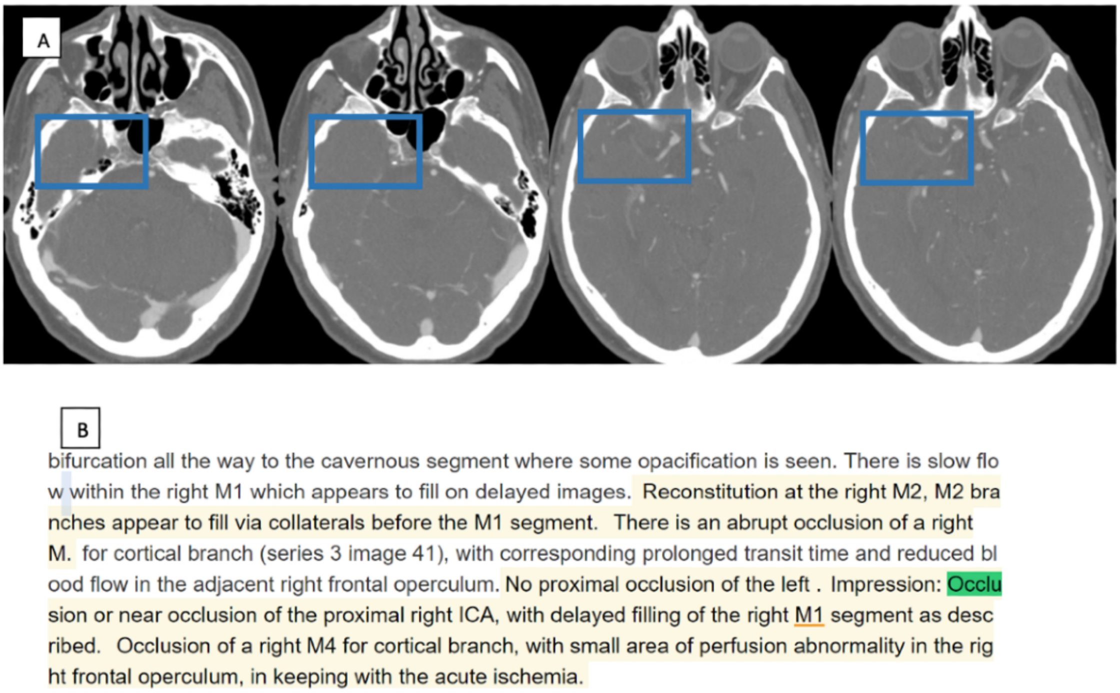
The study was approved by the Sunnybrook Health Sciences Centre and Unity Health Toronto Research Ethics Boards with a waiver of individual patient consent prior to data collection.

Results

Among the 1320 consecutive diagnostic imaging reports manually reviewed, chart abstractors identified 184 large vessel occlusions (MCA-M1, n=157; ACA-A1, n=27) in 161 (12.2%) reports. Distal anterior circulation occlusion was reported in 188 (14.2%) scans, basilar artery occlusion in 26 (2.0%) scans, established ischemia in 391 (29.6%) scans, and intracranial hemorrhage in 139 (10.5%) scans. ASPECTS was reported only in 384 (29.1%) reports (ASPECTS <5, n=40; ASPECTS ≥5, n=344), and collateral status was described in 216 (16.4%) reports (good, n=141; intermediate, n=26; poor, n=49).

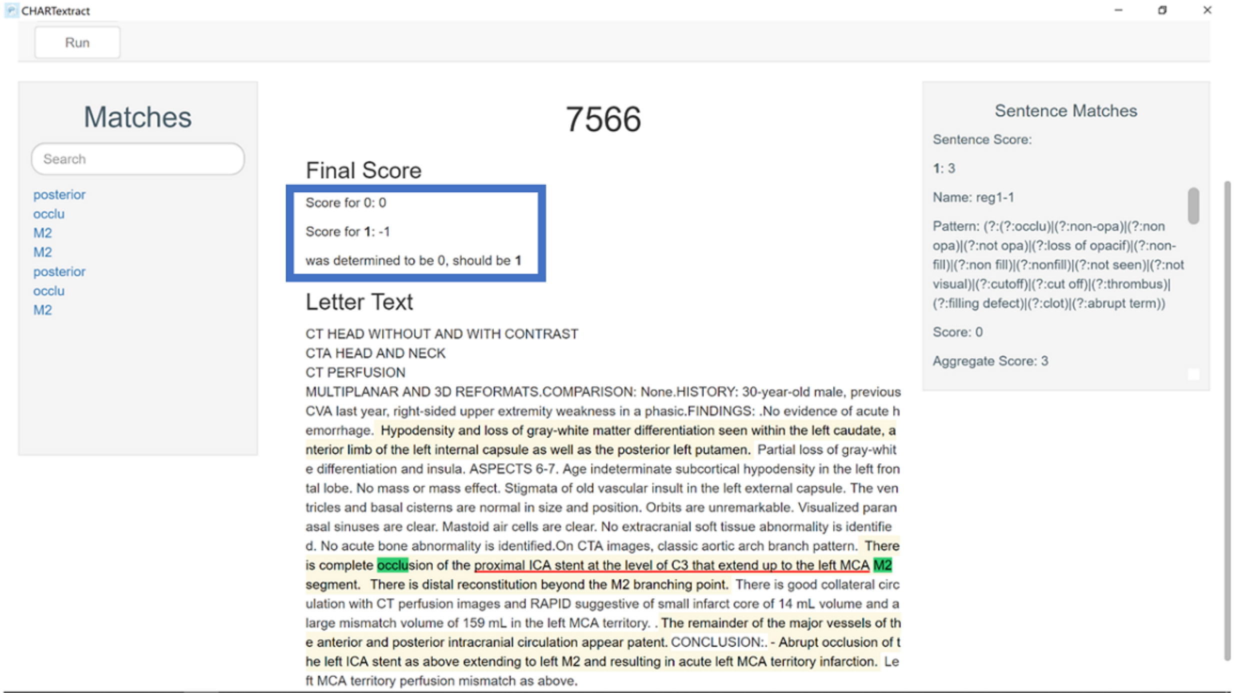
Compared to the reference standard, the NLP tool identified large vessel occlusion with an overall accuracy of 97.3% (95.5% sensitivity, 98.1% specificity, 84.1% PPV, and 99.4% NPV). Despite an iterative process to refine rules, some scenarios remained challenging to translate into rules. Figure 2 illustrates an example wherein the CHARTextract tool determined large vessel occlusion to be present because the words "occlusion" and "M1 segment" were detected in the same sentence, but the report indicated that the occlusion was in the cavernous portion of the internal carotid artery with reconstitution of blood flow in the M1 segment. In another example illustrated in Figure 3, the CHARTextract tool determined that large vessel occlusion was absent because the report indicated the presence of an occlusion extending from the internal carotid artery to the M2 segment. Here, the tool only detected "internal carotid artery" and "M2" as keywords and could not interpret the vascular anatomy described in the report. Nevertheless, in the validation set, the overall accuracy for large vessel occlusion was still high at 95.2% (90.0% sensitivity, 97.4% specificity, 76.3% PPV, and 98.5% NPV). We also found that two of the 25 discrepancies between the abstractors and the NLP tool were due to chart abstractor error.

Figure 2. Example 2 of a discrepancy between the chart abstractor and CHARTextract tool output. (A) Computed tomography angiography scan showing near-occlusion of the cavernous internal carotid artery with reconstitution of the middle cerebral artery. (B) CHARTextract output: the abstractor labeled that large vessel occlusion was absent, but the CHARTextract tool determined this attribute to be present because the words "occlusion" and "M1 segment" were detected in the same sentence.



bifurcation all the way to the cavernous segment where some opacification is seen. There is slow flo
w within the right M1 which appears to fill on delayed images. Reconstitution at the right M2, M2 bra
nches appear to fill via collaterals before the M1 segment. There is an abrupt occlusion of a right
M. for cortical branch (series 3 image 41), with corresponding prolonged transit time and reduced bl
ood flow in the adjacent right frontal operculum. No proximal occlusion of the left . Impression: Occlu
sion or near occlusion of the proximal right ICA, with delayed filling of the right M1 segment as desc
ribed. Occlusion of a right M4 for cortical branch, with small area of perfusion abnormality in the rig
ht frontal operculum, in keeping with the acute ischemia.

Figure 3. Example 3 of a discrepancy between the chart abstractor and CHARTextract tool output. The abstractor labeled that large vessel occlusion was present because the abstractor was able to interpret that an occlusion from the internal carotid artery and extending to the M2 segment of the middle cerebral artery involves the M1 segment, but the CHARTextract tool determined this attribute to be absent because the tool detects key words without knowledge of vascular anatomy.



7566

Final Score

Score for 0: 0

Score for **1**: -1

was determined to be 0, should be **1**

Letter Text

CT HEAD WITHOUT AND WITH CONTRAST
CTA HEAD AND NECK
CT PERFUSION
MULTIPLANAR AND 3D REFORMATS.COMPARISON: None.HISTORY: 30-year-old male, previous
CVA last year, right-sided upper extremity weakness in a phasic.FINDINGS: .No evidence of acute h
emorrhage. Hypodensity and loss of gray-white matter differentiation seen within the left caudate, a
nterior limb of the left internal capsule as well as the posterior left putamen. Partial loss of gray-whit
e differentiation and insula. ASPECTS 6-7. Age indeterminate subcortical hypodensity in the left fron
tal lobe. No mass or mass effect. Stigmata of old vascular insult in the left external capsule. The ven
tricles and basal cisterns are normal in size and position. Orbits are unremarkable. Visualized paran
asal sinuses are clear. Mastoid air cells are clear. No extracranial soft tissue abnormality is identifie
d. No acute bone abnormality is identified.On CTA images, classic aortic arch branch pattern. There
is complete occlusion of the proximal ICA stent at the level of C3 that extend up to the left MCA M2
segment. There is distal reconstitution beyond the M2 branching point. There is good collateral circ
ulation with CT perfusion images and RAPID suggestive of small infarct core of 14 mL volume and a
large mismatch volume of 159 mL in the left MCA territory. . The remainder of the major vessels of th
e anterior and posterior intracranial circulation appear patent. CONCLUSION:. - Abrupt occlusion of t
he left ICA stent as above extending to left M2 and resulting in acute left MCA territory infarction. Le
ft MCA territory perfusion mismatch as above.

Sentence Matches

Sentence Score:

1: 3

Name: reg1-1

Pattern: (?:(?:occlu)|(?:non-opa)|(?:non opa)|(?:not opa)|(?:loss of opacif)|(?:non-fill)|(?:non fill)|(?:nonfill)|(?:not seen)|(?:not visual)|(?:cutoff)|(?:cut off)|(?:thrombus)|(?:filling defect)|(?:clot)|(?:abrupt term))

Score: 0

Aggregate Score: 3

The accuracy of the CHARTextract tool for the other stroke attributes is presented in Table 1. The tool identified these other attributes with moderately high accuracy except for presence of established ischemia, which had a lower sensitivity and PPV of 82.2% and 80.5%, respectively, in the derivation cohort and 80.8% and 64.1%, respectively, in the validation cohort. The other exception was basilar occlusion, which was only present in 2.0% (26/1320) of the reports. Although the sensitivity and PPV for basilar occlusion were 100% and 95.0%, respectively, in the derivation cohort, the corresponding values were lower in the validation cohort (ie, 71.4% and 41.7%)

Table 1. Accuracy of the natural language processing tool CHARTextract to identify stroke-related attributes in diagnostic imaging reports.

Cohort and stroke-related attribute	Attribute prevalence, n (%)	Sensitivity (%)	Specificity(%)	PPV[a] (%)	NPV[b] (%)	Overall accuracy (%)
Derivation cohort (n=921)						
Anterior proximal occlusion	111 (12.1)	95.5	98.1	84.1	99.4	97.3
Anterior distal occlusion	127 (13.8)	92.9	98.0	88.1	98.9	97.3
Basilar occlusion	19 (2.1)	100	99.9	95.0	100	99.9
Presence of established ischemia	287 (31.2)	82.2	91.7	80.5	91.9	88.3
Presence of any hemorrhage	114 (12.4)	93.0	98.2	87.6	99.0	97.5
Validation cohort (n=399)						
Anterior proximal occlusion	50 (12.5)	90.0	97.4	76.3	98.5	95.2
Anterior distal occlusion	61 (15.3)	83.6	97.7	86.4	97.1	95.5
Basilar occlusion	7 (1.8)	71.4	98.2	41.7	99.5	97.7
Presence of established ischemia	104 (26.1)	80.8	85.1	64.1	92.5	83.2
Presence of any hemorrhage	25 (6.3)	88.0	96.0	59.5	99.2	95.5

[a] PPV: positive predictive value.

[b] NPV: negative predictive value.

The metrics for ASPECTS and collateral status are shown separately because data were incomplete (Table 2). Importantly, we found that the NLP tool was able to identify the reports with missing data with high accuracy. For example, information on ASPECTS was absent in 71.8% (661/921) of the reports in the derivation cohort and 68.99% (275/399) for the validation cohort. The tool accurately identified that this attribute was missing with a sensitivity and PPV of 99.7% and 99.7%, respectively, in the derivation cohort and 99.3% and 98.6%, respectively, in the validation cohort.

Table 2. Accuracy of the natural language processing tool CHARTextract to identify Alberta stroke program early CT score (ASPECTS) and collateral vascular status based on diagnostic imaging reports.

Cohort and stroke-related attributes	Attribute prevalence, n (%)	Sensitivity (%)	Specificity (%)	PPV[a] (%)	NPV[b] (%)	Overall accuracy (%)
Derivation cohort (n=921)						
ASPECTS						98.8
Not reported	661 (71.8)	99.7	99.2	99.7	99.2	
<5	30 (3.3)	96.7	99.2	80.6	99.9	
≥5	230 (25.0)	96.5	99.7	99.1	98.9	
Collateral status						98.4
Not reported	774 (84.0)	99.2	96.6	99.4	95.9	
Poor	34 (3.7)	94.1	100	100	99.8	
Intermediate	19 (2.1)	78.9	100	100	99.6	
Good	94 (10.2)	96.8	98.8	90.1	99.6	
Validation cohort (n=399)						
ASPECTS						98.5
Not reported	275 (68.9)	99.3	96.8	98.6	98.4	
<5	10 (2.5)	70.0	100	100.0	99.2	
≥5	114 (28.6)	99.1	99.3	98.3	99.6	
Collateral status						98.2
Not reported	330 (82.7)	99.7	91.3	98.2	98.4	
Poor	15 (3.8)	93.3	99.7	93.3	99.7	
Intermediate	7 (1.8)	71.4	100	100	99.5	
Good	47 (11.8)	93.6	100	100	99.2	

[a] PPV: positive predictive value.

[b] NPV: negative predictive value.

Discussion

Principal Findings

We showed that an NLP approach can automate data extraction from neuroimaging reports with moderately high accuracy, supporting its potential application for stroke surveillance, health system planning, and population-based clinical research. The PPV of CHARTextract to identify large vessel occlusion was 76.3%, meaning that of 100 reports identified to have a large vessel occlusion, there were 24 false-positive cases, but the sensitivity, specificity, and NPV were over 90%, indicating the prevalence of fewer false-negative cases. Thus, NLP may be a helpful screening tool for case finding purposed when using a large dataset.

Although we did not formally record the time required for data abstraction, the abstractors estimate an average review time of 5 minutes per chart, which adds to 110 hours of sustained attention to review a total of 1320 charts. On the other hand, once the rule sets have been developed, the NLP tool can extract the requested variables within seconds.

Limitations

There are several limitations of NLP that are worth discussing. First, the NLP approach can only extract information from the radiologist's reported interpretation of diagnostic images, and it is not designed to be directly used for imaging interpretation [4]. Although the tool was accurate in identifying which reports had missing data on ASPECTS and collateral status, information on these attributes was simply not obtainable without the direct assessment of the images. Second, each rule is applied at a sentence level so that the tool will not be able to capture attributes if keywords occur across different sentences. Third, the tool does not distinguish between homonyms in the English language. For instance, we experienced challenges with the word "ASPECT" used to describe the score and "aspect" used to describe a facet of the brain or a component of a blood vessel. Finally, the NLP approach is influenced by variations in reporting practices to describe imaging findings. This was most apparent in the evaluation of the presence of cerebral ischemia. The terms used to describe this attribute were less predictable and frequently contained ambiguous language such as "possible subtle hypodensity" or "cannot rule out early ischemia." Interestingly, the cerebral ischemia attribute also had a lower inter-rater reliability between the chart abstractors compared to the other attributes evaluated. We noticed that the nonclinical research assistant, who has extensive experience with chart abstraction for stroke research, was more liberal in recording ischemia, whereas the stroke specialist was more selective in recording ischemia depending on the language used by the radiologist. In this situation, the application of NLP rule sets may improve the standardization of data collection. Finally, the current proof-of-concept study has a small sample size. External validation of our methods with a larger sample of radiology reports is needed to address the limitations arising from variation in reporting practices.

Conclusions

NLP approaches can identify the presence of large vessel occlusion with high accuracy and have the potential to improve the efficiency of large-scale data collection from imaging reports. External validation of our approach is needed.

Acknowledgments

This study was supported by the Sunnybrook Alternative Funding Plan Association through the Innovation Fund of the Alternative Funding Plan from the Academic Health Sciences Centres, Ontario, Canada. AY is supported by a national new investigator award from the Heart and Stroke Foundation of Canada. MKK is supported by a mid-career investigator award from the Heart and Stroke Foundation of Canada and holds the Lillian Love Chair in Women's Health at the University Health Network.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CHARTextract tool rules.
[PDF File (Adobe PDF File), 1217 KB - medinform_v9i5e24381_app1.pdf]

References

1. Krueger H, Koot J, Hall RE, O'Callaghan C, Bayley M, Corbett D. Prevalence of individuals experiencing the effects of stroke in Canada: Trends and projections. Stroke 2015 Aug;46(8):2226-2231. [doi: 10.1161/STROKEAHA.115.009616] [Medline: 26205371]
2. Goyal M, Menon BK, van Zwam WH, Dippel DWJ, Mitchell PJ, Demchuk AM, HERMES collaborators. Endovascular thrombectomy after large-vessel ischaemic stroke: A meta-analysis of individual patient data from five randomised trials. Lancet 2016 Apr 23;387(10029):1723-1731. [doi: 10.1016/S0140-6736(16)00163-X] [Medline: 26898852]
3. Nogueira RG, Jadhav AP, Haussen DC, Bonafe A, Budzik RF, Bhuva P, DAWN Trial Investigators. Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct. N Engl J Med 2018 Jan 04;378(1):11-21. [doi: 10.1056/NEJMoa1706442] [Medline: 29129157]
4. Albers GW, Marks MP, Kemp S, Christensen S, Tsai JP, Ortega-Gutierrez S, DEFUSE 3 Investigators. Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging. N Engl J Med 2018 Feb 22;378(8):708-718 [FREE Full text] [doi: 10.1056/NEJMoa1713973] [Medline: 29364767]
5. Thomalla G, Simonsen CZ, Boutitie F, Andersen G, Berthezene Y, Cheng B, WAKE-UP Investigators. MRI-guided thrombolysis for stroke with unknown time of onset. N Engl J Med 2018 Aug 16;379(7):611-622. [doi: 10.1056/NEJMoa1804355] [Medline: 29766770]
6. Ung D, Kim J, Thrift AG, Cadilhac DA, Andrew NE, Sundararajan V, et al. Promising use of big data to increase the efficiency and comprehensiveness of stroke outcomes research. Stroke 2019 May;50(5):1302-1309. [doi: 10.1161/STROKEAHA.118.020372] [Medline: 31009352]
7. Yu AY, Holodinsky JK, Zerna C, Svenson LW, Jetté N, Quan H, et al. Use and utility of administrative health data for stroke research and surveillance. Stroke 2016 Jul;47(7):1946-1952. [doi: 10.1161/strokeaha.116.012390]
8. Waqas M, Rai AT, Vakharia K, Chin F, Siddiqui AH. Effect of definition and methods on estimates of prevalence of large vessel occlusion in acute ischemic stroke: a systematic review and meta-analysis. J Neurointerv Surg 2020 Mar;12(3):260-265. [doi: 10.1136/neurintsurg-2019-015172] [Medline: 31444289]
9. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural language processing in radiology: A systematic review. Radiology 2016 May;279(2):329-343. [doi: 10.1148/radiol.16142770] [Medline: 27089187]
10. Kim C, Zhu V, Obeid J, Lenert L. Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke. PLoS One 2019;14(2):e0212778 [FREE Full text] [doi: 10.1371/journal.pone.0212778] [Medline: 30818342]
11. Ong CJ, Orfanoudaki A, Zhang R, Caprasse FPM, Hutch M, Ma L, et al. Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports. PLoS One 2020;15(6):e0234908 [FREE Full text] [doi: 10.1371/journal.pone.0234908] [Medline: 32559211]
12. CHARTextract - Li Ka Shing Centre for Healthcare Analytics Research & Training (LKS-CHART). 2019. URL: https://lks-chart.github.io/CHARTextract-docs/ [accessed 2019-08-02]

13. Lakomkin N, Dhamoon M, Carroll K, Singh IP, Tuhrim S, Lee J, et al. Prevalence of large vessel occlusion in patients presenting with acute ischemic stroke: a 10-year systematic review of the literature. J Neurointerv Surg 2019 Mar;11(3):241-245. [doi: 10.1136/neurintsurg-2018-014239] [Medline: 30415226]
14. Barber PA, Demchuk AM, Zhang J, Buchan AM. Validity and reliability of a quantitative computed tomography score in predicting outcome of hyperacute stroke before thrombolytic therapy. ASPECTS Study Group. Alberta Stroke Programme Early CT Score. Lancet 2000 May 13;355(9216):1670-1674. [doi: 10.1016/s0140-6736(00)02237-6] [Medline: 10905241]

Abbreviations

ACA-A1: A1 segment of the anterior cerebral artery
ASPECTS: Alberta stroke program early CT score
CT: computed tomography
CTA: computed tomography angiography
CTP: computed tomography perfusion
MCA-M1: M1 segment of the middle cerebral artery
NLP: natural language processing
NPV: negative predictive value
PPV: positive predictive value

Edited by G Eysenbach; submitted 16.09.20; peer-reviewed by R Dewey, M Aksoy; comments to author 28.10.20; revised version received 10.11.20; accepted 16.04.21; published 04.05.21.

Please cite as:
Yu AYX, Liu ZA, Pou-Prom C, Lopes K, Kapral MK, Aviv RI, Mamdani M
Automating Stroke Data Extraction From Free-Text Radiology Reports Using Natural Language Processing: Instrument Validation Study
JMIR Med Inform 2021;9(5):e24381
URL: https://medinform.jmir.org/2021/5/e24381
doi:10.2196/24381
PMID:33944791

Original Paper

Extracting Drug Names and Associated Attributes From Discharge Summaries: Text Mining Study

Ghada Alfattni[1,2], BSc (Hons), MSc; Maksim Belousov[1], BSc (Hons), PhD; Niels Peek[3,4,5], PhD; Goran Nenadic[1,5], BSc (Hons), MSc, PhD

[1]Department of Computer Science, University of Manchester, Manchester, United Kingdom

[2]Department of Computer Science, Jamoum University College, Umm Al-Qura University, Makkah, Saudi Arabia

[3]Centre for Health Informatics, Division of Informatics, Imaging and Data Sciences, University of Manchester, Manchester, United Kingdom

[4]National Institute of Health Research Manchester Biomedical Research Centre, Manchester Academic Health Science Centre, University of Manchester, Manchester, United Kingdom

[5]The Alan Turing Institute, Manchester, United Kingdom

Corresponding Author:
Ghada Alfattni, BSc (Hons), MSc
Department of Computer Science
University of Manchester
Oxford Road
Manchester,
United Kingdom
Phone: 44 7478375084
Email: gafattni@uqu.edu.sa

Abstract

Background: Drug prescriptions are often recorded in free-text clinical narratives; making this information available in a structured form is important to support many health-related tasks. Although several natural language processing (NLP) methods have been proposed to extract such information, many challenges remain.

Objective: This study evaluates the feasibility of using NLP and deep learning approaches for extracting and linking drug names and associated attributes identified in clinical free-text notes and presents an extensive error analysis of different methods. This study initiated with the participation in the 2018 National NLP Clinical Challenges (n2c2) shared task on adverse drug events and medication extraction.

Methods: The proposed system (DrugEx) consists of a named entity recognizer (NER) to identify drugs and associated attributes and a relation extraction (RE) method to identify the relations between them. For NER, we explored deep learning-based approaches (ie, bidirectional long-short term memory with conditional random fields [BiLSTM-CRFs]) with various embeddings (ie, word embedding, character embedding [CE], and semantic-feature embedding) to investigate how different embeddings influence the performance. A rule-based method was implemented for RE and compared with a context-aware long-short term memory (LSTM) model. The methods were trained and evaluated using the 2018 n2c2 shared task data.

Results: The experiments showed that the best model (BiLSTM-CRFs with pretrained word embeddings [PWE] and CE) achieved lenient micro F-scores of 0.921 for NER, 0.927 for RE, and 0.855 for the end-to-end system. NER, which relies on the pretrained word and semantic embeddings, performed better on most individual entity types, but NER with PWE and CE had the highest classification efficiency among the proposed approaches. Extracting relations using the rule-based method achieved higher accuracy than the context-aware LSTM for most relations. Interestingly, the LSTM model performed notably better in the reason-drug relations, the most challenging relation type.

Conclusions: The proposed end-to-end system achieved encouraging results and demonstrated the feasibility of using deep learning methods to extract medication information from free-text data.

(JMIR Med Inform 2021;9(5):e24678) doi:10.2196/24678

KEYWORDS

information extraction; electronic health records; discharge summaries; natural language processing; medication prescriptions

Introduction

Background

Electronic health records (EHRs) are a valuable source of routinely collected health data that can be used for secondary purposes, including clinical and epidemiological research [1]. They typically contain information on consultations, admissions, symptoms, clinical examinations, test results, diagnoses, treatments, and outcomes. Medication prescriptions are a key source for understanding the effects of patient treatment. In some settings (eg, general practitioners' practices), they might be recorded in a structured fashion through prescribing software and would comprise, apart from drug names, medication attributes such as dosage, frequency, and duration. Still, there are often additional, free-text sources of prescription information, such as clinic letters or discharge summaries, particularly in secondary care. Extracting information from free-text is challenging because much of the information is provided in a narrative manner, and the text is often written in haste and under considerable time pressure. There has been strong interest among researchers in the use of natural language processing (NLP) to extract information from clinical free-text notes on a large scale [2-9], including a number of shared tasks and benchmark data sets to assess and advance the state-of-the-art in this domain, such as challenges in medication extraction [7]; chemical and drug named entity recognition (NER) [10]; drug-drug interaction extraction [11]; and extraction of medications, indications, and adverse drug events (ADEs) [12,13].

Medication prescription instructions are a specific clinical sublanguage, where expressions are often abbreviated (eg, *od* for *once a day*) and may contain spelling errors (eg, *20 mcg evry othr wk*) [14,15]. Existing approaches for extracting drugs and associated attributes from the clinical text are diverse in their methods, using various approaches including dictionary lookup (ie, searching for matches from existing drug dictionaries) [16-18], rule-based approaches (manually design patterns, eg, regular expressions that can be searched in free-text) [2-4,8,14,16,19-22], machine learning approaches (training models on example data) [23-28], and hybrid approaches that combine different methods [29-32]. Recently, methods based on deep learning and neural networks, such as convolutional neural networks and recurrent neural networks, have been shown to be state-of-the-art in drug attribute extraction tasks [33-41]. Deep learning methods take relevant features (eg, orthographic and lexical features) as inputs and produce labels as outputs. These manually constructed feature vectors can then be replaced with, for example, word embeddings (WE), character embeddings (CEs), and feature embeddings. Embeddings are representations of tokens in an n-dimensional space, typically learned over large collections of unlabeled data through an unsupervised process (eg, word2vec [42], Global Vectors for Word (GloVe) [43], and fastText [44]). Recently, more advanced embedding methods and representations (eg, Embeddings from Language Models [ELMo] [45] and Bidirectional Encoder Representations from Transformers [BERT] [46]) have further advanced state-of-the-art clinical NLP.

Objectives

Although deep learning methods have been extensively used in medication information extraction [13], the effects of various architectures and token representations have not been widely discussed. The purpose of this study is to provide a comprehensive comparison of various representations used for drug information extraction within the same settings. The main contributions of our work are as follows:

- An investigation of the effect of various token representations (ie, CE, WE, and semantic-feature embeddings [SFEs]) on extracting medication information
- The comparison between a rule-based method and deep learning approaches for identifying relations between drugs and associated attributes.

Methods

Overview

The DrugEx system proposed here is composed of (1) an NER method for extracting mentions of drug names and drug-associated attributes and (2) a relation extraction (RE) method for identifying relations between drugs and their associated attributes. The NER task involves extracting 8 types of entities: drug, strength, duration, route, form, dosage, frequency, and reason of administration (see Textbox 1 for definitions and examples of the extracted entities).

Textbox 1. Definitions and examples of entity types extracted by the DrugEx system.

- Drug: The chemical name of a drug or the advertised brand name under which a drug is sold (eg, aspirin)
- Dosage: The amount of medicine that the patient takes or should take (eg, 2 tablets, 5 mL)
- Strength: The amount of drug in a given dosage (eg, 200 mg)
- Frequency: The rate at which medication was taken or is repeated over a particular period (eg, daily, every 4 hours)
- Duration: The period of continuous medication taking (eg, *pro re nata*, for 5 days)
- Route: The path by which medication is taken into the body or the location at which it is applied (eg, topical, *per os*)
- Form: The form in which a medication is marketed for use (eg, tablet)
- Reason: The reason for medication administration (eg, for pain)

The scope of these entity types and the data sets that were used for training and evaluation were provided as part of the 2018 National NLP Clinical Challenges (n2c2) shared task track 2: ADEs and medication extraction in EHR challenge [13,47]. The data set consists of discharge summaries drawn from the Medical Information Mart for Intensive Care III (MIMIC-III) clinical care database [48]. It comprises 505 documents, of which 303 documents were used as the training set, and the remaining 202 documents were used as the test set. These data were annotated by 7 domain experts, consisting of 4 physician-assistant students and 3 nurses. Annotations included drug, strength, dosage, frequency, duration, form, route, reason, and ADEs; ADEs annotations have been omitted here as they are beyond the scope of this study.

The annotations also included relations between drugs and other attributes. Table 1 shows the descriptive statistics for the associated drug attributes in the n2c2 data set and how often each of them was linked to more than 1 drug. Noticeably, 17% (1412/8579) of the reason entities were associated with more than one drug; the maximum number of drugs associated with a single reason was 10. For example, in “START: Guaifensin with codeine QHS and Benzonatate as needed for cough,” the reason *cough* is associated with 2 drugs: guaifenesin (with codeine) and benzonatate. Table 2 shows the number of drug entities participating in each link and the ratio of drugs with more than one link. From a total of 11,028 form-drug relations, 4517 (41%) drugs that have been associated with the form attribute has more than one association (ie, multiple forms reported for a single drug entity), for example, “Bisacodyl 5 mg Tablet Sig: 1-2 Tablets PO once a day as needed for constipation;” both mentions of tablets were annotated as form, and they both associated to the bisacodyl drug.

Table 1. Descriptive statistics of entity types in the National NLP Clinical Challenges (n2c2) data set.

Entity types	Entities, n (%)	Links to 1 drug, n (%)	Links to multiple drugs, n (%)	Maximum number of drug associations
Drug	26,800 (32.57)	—[a]	—	—
Form	11,010 (13.38)	10,980 (99.56)	48 (<1)	2
Strength	10,921 (13.27)	10,913 (99.70)	33 (<1)	3
Frequency	10,293 (12.51)	10,281 (99.39)	63 (1)	4
Route	8989 (10.92)	9000 (99.08)	84 (1)	4
Dosage	6902 (8.39)	6877 (99.38)	43 (1)	4
Reason	6400 (7.78)	7158 (83.44)	1421 (16.56)	10
Duration	970 (1.2)	991 (92.7)	78 (7)	4

[a] Not applicable.

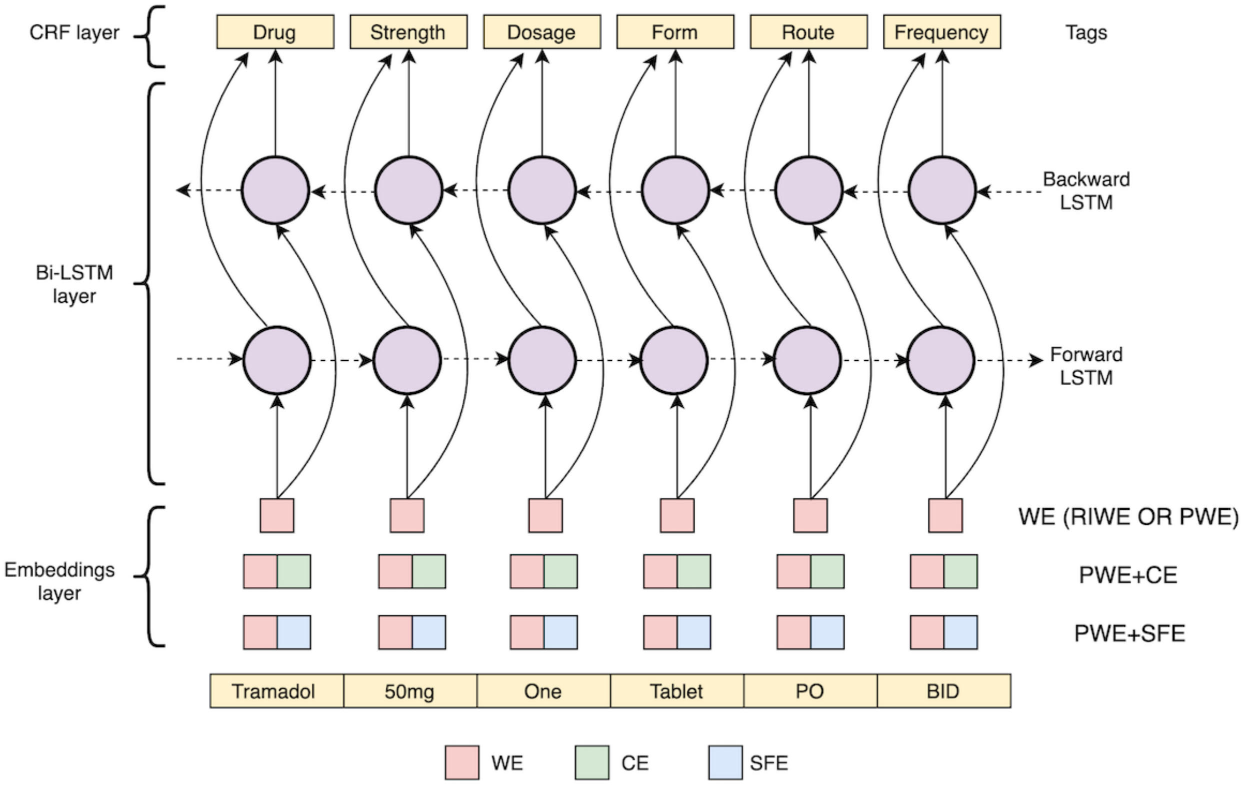
Table 2. Descriptive statistics of relations between drugs and their associated attributes in the National NLP Clinical Challenges (n2c2) data set.

Relation type	Relations, n (%)	Drugs with 1 link, n (%)	Drugs with more than 1 link, n (%)
Strength-drug	10,946 (18.88)	10,639 (97.20)	307 (2.8)
Frequency-drug	10,344 (17.84)	10,054(97.20)	290 (2.8)
Route-drug	9084 (15.67)	8903 (98.01)	181 (1.99)
Reason-drug	8579 (14.80)	7704 (89.80)	875 (10.2)
Dosage-drug	6920 (11.94)	6765 (97.76)	155 (2.2)
Form-drug	11,028 (19.02)	6511 (59.04)	4517 (40.96)
Duration-drug	1069 (1.84)	1021 (95.51)	48 (5)

NER Method

All NER models rely on bidirectional long-short term memory with conditional random fields (BiLSTM-CRF) architecture (Figure 1), which is composed of 3 different layers: embedding layer, bidirectional long-short term memory (BiLSTM) layer, and conditional random fields (CRFs) layer.

Figure 1. The architecture of bidirectional long-short term memory with conditional random field for the named entity recognition models. BiLSTM-CRF: bidirectional long-short term memory with conditional random field; PWE+CE: pretrained word embeddings and character embeddings; PWE: pretrained word embeddings; PWE+SFE: pretrained word embeddings and semantic-feature embeddings; RIWE: randomly initialized word embeddings; WE: word embeddings.



Preprocessing

The data were first tokenized using spaCy, an open-source library for NLP, with support for various languages. Then, as target entities differ in length and may contain more than one token, each token was annotated using the BIOES (Begin, Inside, Outside, End, Single) tagging scheme to capture information about the sequence of tokens. We further processed the discharge summaries using the Clinical Language Annotation, Modeling, and Processing Toolkit (CLAMP) [49] and the Clinical Text Analysis and Knowledge Extraction System (cTAKES) [50] to extract token-level clinical semantic tags (eg, medication, disease disorder, and procedure; see the section *Embedding Layer* for details), which were used for SFEs.

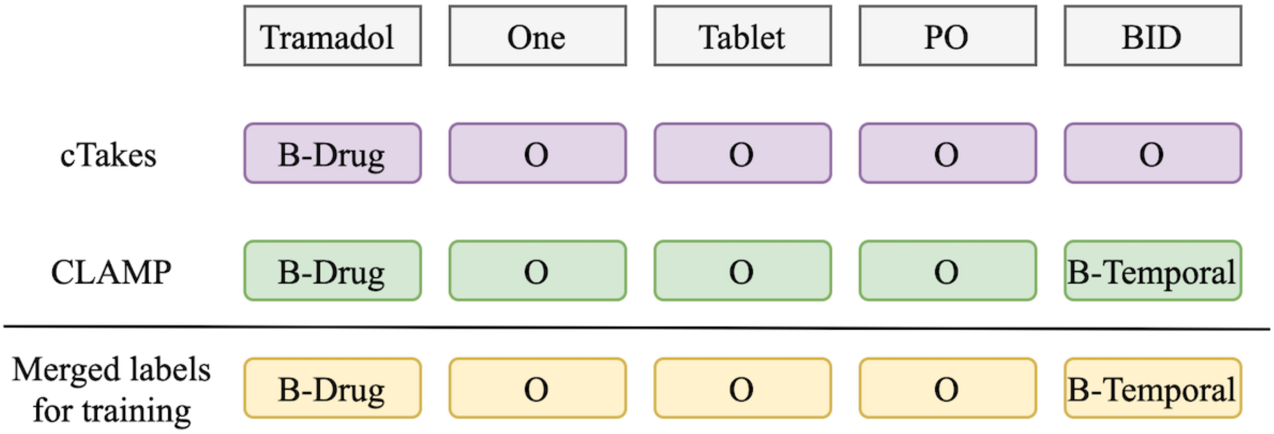
Embedding Layer

The embedding layer maps tokens into vectors of numbers that represent their meanings. WEs provide dense representations that make them capable of representing many aspects of similarities between words, such as semantic relations and morphological properties [51,52]. Several methods can be used to initialize the values in WEs at the beginning of neural network training. We examined the randomly initialized word embeddings (RIWE) and the pretrained word embeddings (PWE), where the latter has been pretrained on data from the clinical (ie, target) domain.

Although WEs can capture tokens' semantics, they might still be affected by data sparsity and, therefore, cannot remediate synonyms, out-of-vocabulary tokens, and misspellings. WE may not be able to capture morphemes (such as prefixes and suffixes) derived from classic Latin and ancient Greek roots, which are often included in drug names and drug attributes. Thus, we addressed these issues by using character feature embeddings in addition to WEs. The concatenation of the PWE with the CEs allows the model to learn subtoken patterns such as morphemes and roots, thereby aiming to capture out-of-vocabulary tokens, different forms, and any other information not captured by WEs [53].

We also considered representations beyond tokens, aiming to add clinical semantics to words. Specifically, the concatenation of the PWE and SFEs was used to represent the clinical categories of entities identified in the text, such as medical problems, tests, or temporal information. Note that in this study, we did not evaluate SFE without PWEs. Some entity types (such as frequency or route) are not present among the semantic tags we used, whereas other semantic tags (such as signs, symptoms, disease, and disorder) are more frequent. Therefore, the representations of semantic tags were learned simultaneously with word representations and concatenated together to form the final token representations. We used CLAMP [49] to extract semantic tags (ie, problem, treatment, and temporal entities) with associated assertion tag attributes (ie, present or absent). We also used the default clinical pipelines in cTAKES [50] to tag tokens with other semantic categories (ie, Medication, DiseaseDisorder, and SignSymptom). In each pipeline, tokens were tagged with the corresponding semantic features and attributes (if available); otherwise, they were tagged with the outside (ie, O) tag. Token-level semantic tags from both pipelines were then mapped and merged based on their types to create a set of semantic features (Figure 2).

Figure 2. Semantic-feature token embeddings. B-Drug: begin-drug; B-Temporal: begin-temporal; CLAMP: Clinical Language Annotation, Modeling, and Processing Toolkit; cTakes: Clinical Text Analysis and Knowledge Extraction System; O: outside.



BiLSTM Layer

The BiLSTM layer takes the sequence of vectors (ie, token representations) corresponding to a sequence of tokens (the output from the embedding layer) and calculates the hidden states by processing the sequence of token representations forward and backward (ie, left-to-right and right-to-left) to learn important token-level features. It then outputs the sequence of vectors, including the probability of each label for each corresponding token. The labels were either 1 of the 8 entity types (Textbox 1) or none. The label assigned to the token is the label with the highest probability from the predicted labels' sequence (output from the BiLSTM layer).

CRF Layer

The BiLSTM output does not consider the dependencies between neighboring labels when predicting the current label. For example, it may be more likely to have a token labeled as a drug name followed by a token labeled as *strength* than any other entity type. Thus, to learn these dependencies, we added a CRF layer that uses past and future labels to optimize predictions and obtain the most probable sequence of predicted labels. Finally, the labels (BIOES tags) were combined into named entities by merging consecutive labeled B-, I-, E-, or S-tags of the same class.

NER Models Training and Tuning of Hyperparameters

We used the standard data split established by the n2c2 organizers, using the training set for fitting models, tuning the model parameters, and evaluating our best models on the test set. As there is no official development set, we randomly selected 9.9% (30/303) of the training documents for validation. This data set was used to optimize the models' hyperparameters.

We trained all neural network models using stochastic gradient descent, with a learning rate of 0.005. In the baseline model (RIWE), we randomly selected 100-dimensional WEs. In other models, we used pretrained 600-dimensional WEs [54], which were trained on approximately 2 million discharge summaries drawn from the MIMIC-III data [48] using the word2vec continuous bag-of-words method [42]. CEs were 25-dimensional vectors, whereas SFEs were 50-dimensional vectors. The number of hidden states was set to 300 dimensions for running the BiLSTM WEs and to 25-dimensions for running the BiLSTM for learning CE. We also applied dropout to the token embeddings at a rate of 0.5 to avoid overfitting. The number of epochs was determined by an early stopping criterion (ie, after 10 epochs with no improvement) on the validation set, with the maximum number of epochs set to 100. Finally, the batch size was set to 32. These hyperparameters were optimized through a random search of the validation set [55]. We tested WEs with dimensions ranging from 100 to 600, CE and SFEs with 25, 50, and 100 dimensions, and the dropout rate with values in the range between 0 and 0.75.

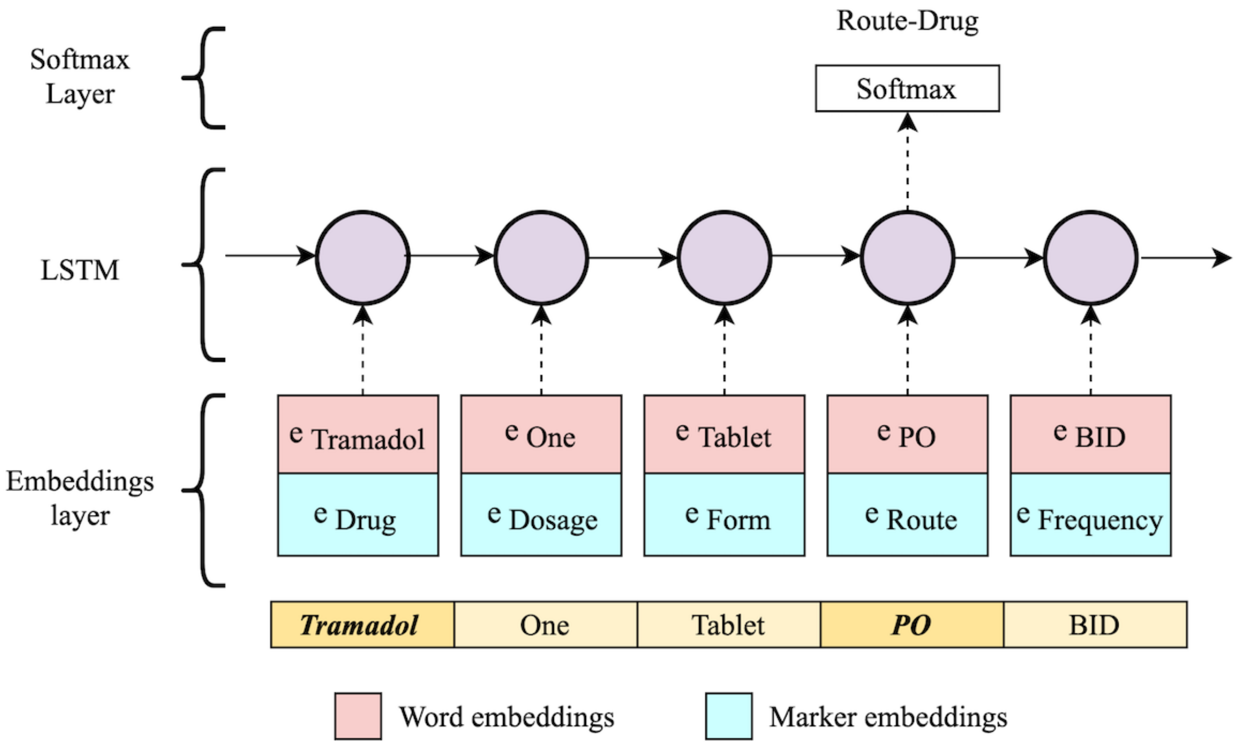
RE Method

Once drugs and attributes are extracted, the subsequent step is to link drug names to the corresponding attributes. For this task, we experimented with a rule-based method engineered for the task and a context-aware long-short term memory (LSTM) model, where the positions of the involved entities were encoded using marker embeddings.

Context-Aware LSTM

We used a context-aware LSTM [56] that considers other relations in the sentential context while predicting the target relation. It uses an LSTM-based encoder to jointly learn representations for all relations in the text. Thus, the representation of the target relation and representations of the context relations are combined to make the final prediction. Figure 3 presents the architecture of the LSTM model for RE. It consists of an embedding layer, an LSTM layer, and a softmax layer. The embedding layer maps a portion of the text that contains a target entity pair into a high-level representation vector. First, each token in the text is mapped to its WE vector. Second, every 2 entities (ie, a drug and its associated attribute) in the text are paired as candidate entities for a possible relation. All other tokens are then marked as either belonging to a drug (as the main actor of all relations) or not. Afterward, each token's marker embeddings are concatenated to the WEs to generate a single vector. This vector is then passed to the LSTM, which calculates the hidden states by processing the sequence of token representations. Finally, the LSTM layer's output is routed into the softmax layer to map the nonnormalized output to the final output vector that contains the probability for each relation type.

Figure 3. The architecture of context-aware long-short term memory for the relation extraction model. e: embedding; LSTM: long-short term memory.

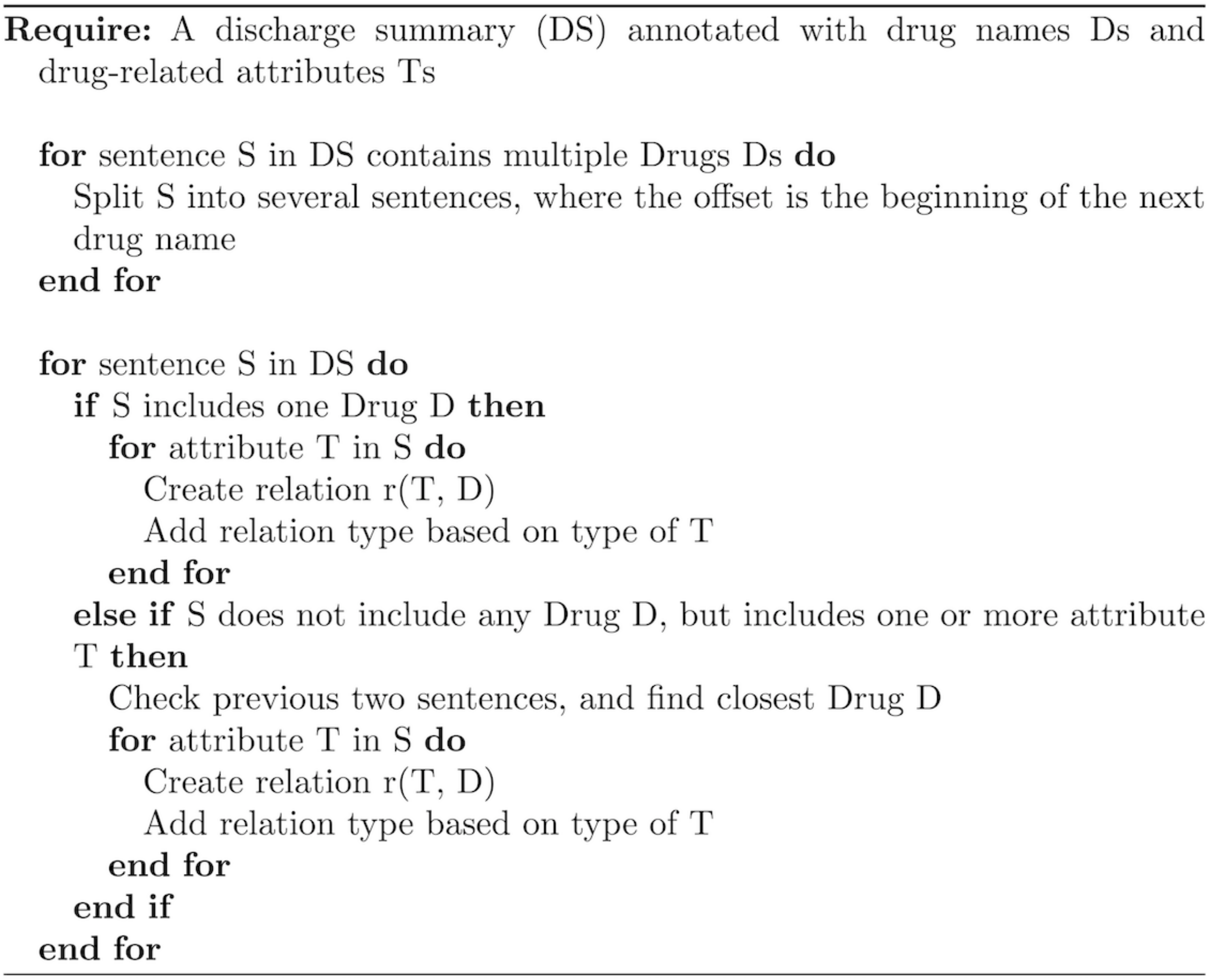


Rule-Based Method

In this approach, we examined patterns of prescription information in discharge summaries in the training set and manually implemented a set of rules using regular expressions. These regular expressions were designed and implemented in the General Architecture of Text Engineering environment [57] (Figure 4). First, the discharge summaries were split into sentences. For sentences that include only one drug name D, all drug attributes found in that sentence will be linked to drug D. However, for sentences that include multiple drug names, the sentences are split into several segments, where the segment's start offset is the beginning of the next drug name.

If a sentence does not include a drug name but contains other entities, then the previous 2 sentences are checked. If they contain a drug name, then the attributes are linked to the closest drug name. For example, "Patient will be on Topiramate *25mg PO BID* until 22/3 PM. Then increase to *50mg po BID for seven days*. Then increase to *75mg ongoing*". All the italicized entities are linked to the drug *topiramate* that appears in the first sentence.

Figure 4. Rule-based method for linking drug names to corresponding attributes in discharge summaries.



```
Require: A discharge summary (DS) annotated with drug names Ds and
  drug-related attributes Ts

  for sentence S in DS contains multiple Drugs Ds do
    Split S into several sentences, where the offset is the beginning of the next
    drug name
  end for

  for sentence S in DS do
    if S includes one Drug D then
      for attribute T in S do
        Create relation r(T, D)
        Add relation type based on type of T
      end for
    else if S does not include any Drug D, but includes one or more attribute
    T then
      Check previous two sentences, and find closest Drug D
      for attribute T in S do
        Create relation r(T, D)
        Add relation type based on type of T
      end for
    end if
  end for
```

RE Model Training and Tuning of Hyperparameters

We used the same procedure and the same approach for hyperparameter settings that we have used previously in the NER models. Specifically, we trained the LSTM model using the same hyperparameters that we have used previously in the NER models. We used marker embeddings with 10-dimensional vectors.

The regular expressions in the RE rule-based method were implemented based on manual observation of the training set, followed by an initial evaluation of the validation set. The regular expressions were then refined based on an error analysis of the output from the validation process, and the final evaluation was performed on the official test set.

Evaluation

We considered the available annotations in the corpus as the gold standard when evaluating the models. To assess the performance of the proposed models, we performed hold-out cross-validation (using training and testing sets) and used the official n2c2 evaluation script provided with the data. It uses standard evaluation methods in information retrieval (ie, precision, recall, and F-score). We report the lenient micro-and macroaveraging for each NER experiment. Lenient matches refer to cases where the overlapped boundaries between the gold standard and the system's predictions are allowed. Macroaveraging calculates the metrics on a per-document basis and then averages the results. Microaveraging, on the other hand, refers to the pooling of the results of all classified instances into a single contingency table.

In addition, we evaluated the performance of the NER models with the best-performing RE model as an end-to-end system. This allows us to measure the effect of missing entities in the NER models on the RE task. As shown in Table 1, attributes could be associated with more than one drug. Thus, when an NER model fails to recognize an entity (either drug or attribute), then all of its semantic relations (ie, associations) will also be missed. Finally, the best-performed end-to-end system was chosen for our DrugEX system.

Results

NER Task

Table 3 shows the lenient precision, recall, and F-score for all models in the NER task. The best result in the NER task was achieved by PWE+CE embeddings (micro F-score of 0.921). Interestingly, NER (PWE), which ranked second in F-score, achieved a slightly higher precision, and NER (PWE+SFE) achieved a higher recall than any other model. NER (PWE+SFE) also yielded a better balance between precision and recall. Concerning individual F-scores, PWE performed better than the baseline (RIWE) for every entity type. The SFEs with the PWEs in NER (PWE+SFE) allow the model to perform better than others on some individual entity types, especially frequency, duration, and reason. An analysis at the per-entity type level shows that most entity types (ie, drugs, strength, form, dosage, frequency, and route) are associated with excellent performance (F-scores above 0.90). Duration and reason, however, are associated with lower performance. This might be amplified by the fact that there were few examples of duration and reason entities in the training data (Table 1).

Table 3. Evaluation results of the named entity recognition models on the test set (lenient evaluation).

Entity	RIWE[a]			PWE[b]			(PWE+CE)[c]			(PWE+SFE)[d]		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
Drug	0.942	0.892	0.917	0.963	0.930	0.946	0.946	0.953	0.949	0.952	0.947	*0.950*[e]
Strength	0.977	0.959	0.968	0.979	0.970	0.975	0.973	0.976	0.974	0.977	0.977	*0.977*
Duration	0.893	0.706	0.789	0.883	0.762	0.818	0.910	0.698	0.790	0.903	0.786	*0.840*
Route	0.964	0.928	0.946	0.964	0.938	0.951	0.956	0.948	*0.952*	0.953	0.943	0.948
Form	0.964	0.935	0.949	0.965	0.940	0.952	0.969	0.944	*0.956*	0.972	0.932	0.951
Dosage	0.928	0.912	0.920	0.932	0.931	*0.931*	0.931	0.928	0.929	0.928	0.931	0.930
Frequency	0.945	0.925	0.935	0.965	0.952	0.959	0.980	0.933	0.956	0.968	0.968	*0.968*
Reason	0.771	0.458	0.575	0.821	0.497	0.620	0.860	0.452	0.593	0.621	0.653	*0.637*
Micro	0.943	0.863	0.901	0.951	0.892	0.921	0.950	0.894	*0.921*	0.927	0.913	0.920
Macro	0.936	0.840	0.883	0.951	0.876	0.910	0.949	0.884	*0.914*	0.923	0.901	0.910

[a] RIWE: bidirectional long-short term memory with conditional random fields with random word embeddings.

[b] PWE: bidirectional long-short term memory with conditional random fields with pretrained word embeddings.

[c] (PWE+CE): bidirectional long-short term memory with conditional random fields with pretrained word embeddings and character embeddings.

[d] (PWE+SFE): bidirectional long-short term memory with conditional random fields with pretrained word embeddings and semantic-feature embeddings.

[e] The best results for each metric are italicized.

To explore the complementarity of the methods, we created an ensemble model using the outputs of all the proposed NER models. The ensemble output for each task was generated using a majority voting scheme. In addition to its type, the entire named entity phrase is taken as 1 prediction instance. The ensemble model showed precision, recall, and F-scores of 0.961, 0.884, and 0.921, respectively. As expected, the ensemble showed performance gains in precision when compared with the best individual models. This indicates that the 3 models did not learn the same patterns from the data set. However, the difference in recall and F-score is not evident, even for specific attributes (Table 4).

Table 4. Evaluation results of pretrained word embeddings+character embedding named entity recognition model, pretrained word embeddings+character embedding named entity recognition model, and the ensemble model on the test set (lenient evaluation).

Entity	(PWE+CE)[a]			(PWE+SFE)[b]			Ensemble		
	Precision	Recall	F-score	Recall	Precision	F-score	Precision	Recall	F-score
Drug	0.946	0.953	0.949	0.952	0.947	*0.950*[c]	0.962	0.939	*0.950*
Strength	0.973	0.976	0.974	0.977	0.977	*0.977*	0.981	0.972	*0.977*
Duration	0.910	0.698	0.790	0.903	0.786	*0.840*	0.919	0.720	0.807
Route	0.956	0.948	0.952	0.953	0.943	0.948	0.963	0.944	*0.953*
Form	0.969	0.944	*0.956*	0.972	0.932	0.951	0.972	0.939	0.955
Dosage	0.931	0.928	0.929	0.928	0.931	0.930	0.943	0.930	*0.936*
Frequency	0.980	0.933	0.956	0.968	0.968	*0.968*	0.979	0.915	0.946
Reason	0.860	0.452	0.593	0.621	0.653	*0.637*	0.858	0.476	0.613
Micro	0.950	0.894	*0.921*	0.927	0.913	0.920	0.961	0.884	*0.921*
Macro	0.949	0.884	*0.914*	0.923	0.901	0.910	0.962	0.869	0.911

[a] (PWE+CE): bidirectional long-short term memory with conditional random fields with pretrained word embeddings and character embeddings.

[b] (PWE+SFE): bidirectional long-short term memory with conditional random fields with pretrained word embeddings and semantic-feature embeddings.

[c] The best results for each metric are italicized.

We further conducted paired *t* tests to determine whether the differences between the models were statistically significant. Differences were considered significant if the *P* value was <.05. The samples used in this test were the microaverage F-scores from each document in the test set (ie, document-level NER performance). Table 5 shows the post hoc analysis of variance for the NER task. The statistical significance test showed that there were no statistically significant differences between any of the models (PWE, PWE+CE, and PWE+SFE), despite the presence of apparently important and computationally expensive clinical information such as the type of entities (ie, problems, signs, and symptoms) in some of the models. However, the 3 models (PWE, PWE+CE, and PWE+SFE) were statistically significantly different from the baseline (ie, RIWE), where random embeddings were used. This means that pretraining embeddings on the target domain (ie, discharge summaries from MIMIC-III) helped in comparison with the random initialization of WEs.

Table 5. Post-hoc analysis of variance (ANOVA) of the named entity recognition models: *P* values of two-tailed paired t tests for each pair of models.[a]

Named entity recognition	PWE[b], *P* value	PWE+CE[c], *P* value	PWE+SFE[d], *P* value
RIWE[e]	<.001	<.001	<.001
PWE	N/A[f]	.94	.99
PWE+CE	N/A	N/A	.95

[a] RIWE is significantly worse than the rest of the models. At the same time, there is no statistically significant difference between PWE, PWE+CE, and PWE+SFE.

[b] PWE: pretrained word embeddings.

[c] CE: character embedding.

[d] SFE: semantic-feature embeddings.

[e] RIWE: randomly initialized word embeddings.

[f] N/A: not applicable.

RE Models

Table 6 shows the performances of the RE models using the gold-standard entities, whereas Table 7 shows the performances of the RE model using the output from the NER models (end-to-end). Using the gold-standard entities and using the output from the best NER model (end-to-end), we achieved micro F-scores of 0.927 for rules and 0.855 for (PWE+CE)+rules, respectively. Thus, the traditional rule-based method performed surprisingly well relative to the context-aware LSTM for this task. Relations between form and frequency to drugs are examples of such success: there was at least a 4% improvement in F-score over the LSTM model. The microaverage F-score for the end-to-end task was notably lower than that for the NER tasks and RE using gold-standard entities. This was expected because prediction in the end-to-end compounded the errors in both the NER and RE steps. A major factor behind the low score is the reasons-drug relation type, which was often not recognized because the NER did not recognize the reason attribute. However, the prediction of this relation itself (ie, reason-drug) is also challenging, as evidenced by the F-score of 0.734 in the RE task (rules) on the gold-standard entities. This might be because the text span between 2 entities in this relation is often relatively long; thus, none of the methods explored in this study could capture this.

Table 6. Evaluation results of the relation extraction models (using gold-standard entities) on the test set (lenient evaluation).

Relation type	LSTM[a]			Rules[b]		
	Precision	Recall	F-score	Precision	Recall	F-score
Strength-drug	0.973	0.961	0.967	0.963	0.988	*0.975*[c]
Dosage-drug	0.963	0.958	0.961	0.956	0.976	*0.966*
Duration-drug	0.909	0.892	0.901	0.942	0.880	*0.910*
Frequency-drug	0.962	0.904	0.932	0.964	0.988	*0.975*
Form-drug	0.982	0.918	0.949	0.970	0.992	*0.981*
Route-drug	0.958	0.934	0.946	0.962	0.972	*0.967*
Reason-drug	0.741	0.830	*0.783*	0.767	0.704	0.734
Micro	0.922	0.913	0.918	0.937	0.917	*0.927*
Macro	0.914	0.910	0.909	0.935	0.902	*0.917*

[a] LSTM: long-short term memory method.

[b] Rules: rule-based method.

[c] The best results for each metric are italicized.

Table 7. Evaluation results of the end-to-end models (ie, output from the best-performing named entity recognition and relation extraction models) on the test set (lenient evaluation).

Relation type	RIWE[a]+rules			PWE[b]+rules			(PWE+CE)[c]+rules			(PWE+SFE)[d]+rules		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
Strength-drug	0.919	0.914	0.917	0.952	0.943	0.947	0.948	0.950	0.949	0.948	0.964	*0.956*[e]
Dosage-drug	0.848	0.853	0.851	0.890	0.888	0.889	0.892	0.884	0.888	0.894	0.897	*0.895*
Duration-drug	0.837	0.615	0.709	0.842	0.662	0.741	0.889	0.617	0.729	0.860	0.678	*0.759*
Frequency-drug	0.878	0.874	0.876	0.931	0.919	0.925	0.949	0.902	0.925	0.934	0.947	*0.940*
Form-drug	0.894	0.888	0.891	0.939	0.915	0.927	0.944	0.919	0.931	0.959	0.920	*0.939*
Route-drug	0.885	0.866	0.875	0.924	0.895	0.909	0.919	0.904	0.911	0.920	0.908	*0.914*
Reason-drug	0.635	0.333	0.437	0.702	0.371	0.485	0.744	0.343	0.470	0.503	0.472	*0.487*
Micro	0.865	0.770	0.815	0.909	0.802	0.852	0.918	0.797	*0.855*	0.871	0.830	0.850
Macro	0.859	0.733	0.784	0.902	0.770	0.824	0.918	0.765	*0.824*	0.849	0.801	0.821

[a] RIWE: bidirectional long-short term memory with conditional random fields with random word embeddings.

[b] PWE: bidirectional long-short term memory with conditional random fields with pretrained word embeddings.

[c] PWE+CE: bidirectional long-short term memory with conditional random fields with pretrained word embeddings and character embeddings.

[d] PWE+SFE: bidirectional long-short term memory with conditional random fields with pretrained word embeddings and semantic-feature embeddings.

[e] The best results for each metric are italicized.

The statistical significance test for the RE task showed that the differences between the LSTM and rule-based models were insignificant (*P*=.41). For the end-to-end task, similar to the NER task, there was no statistically significant difference between any of the models (PWE, PWE+CE, and PWE+SFE); however, the 3 models were statistically significantly different from the RIWE (Table 8). Accordingly, the best-performed end-to-end system, (PWE+CE)+rules, was chosen for our DrugEx system.

Table 8. Post-hoc analysis of variance (ANOVA) of the end-to-end models: *P* values of two-tailed paired *t* tests for each pair of models.

End-to-end models	PWE[a]+rules, *P* value	(PWE+CE[b])+rules, *P* value	(PWE+SFE[c])+rules, *P* value
RIWE[d]+rules	.01	.01	.03
PWE+rules	N/A[e]	.99	.99
(PWE+CE)+rules	N/A	N/A	.99

[a] PWE: pretrained word embeddings.

[b] CE: character embedding.

[c] SFE: semantic-feature embeddings.

[d] RIWE: randomly initialized word embeddings.

[e] N/A: not applicable.

Discussion

Principal Findings

The models explored in this study demonstrated high F-scores of 0.921 for NER, 0.927 for RE, and 0.855 for the end-to-end approach. The overall highest F-scores (achieved by different teams) in the n2c2 challenge in the NER, RE, and end-to-end tasks were 0.942, 0.963, and 0.891, respectively [13]. The top-ranked NER used a BiLSTM-CRF with ELMo language model [45], CFEs, and normalized section titles as features. The top-ranked RE and end-to-end tasks used a joint concept-relation extraction system that uses 2 layers of BiLSTM-CRFs [58].

The results for our NER models showed that PWE+CE had the highest classification efficiency, followed by PWE and PWE+SFE, which had similar scores among themselves and above the baseline. RE models' results showed that the rule-based method achieved significantly higher accuracy than the context-aware LSTM for most relation types. Interestingly, the LSTM model performed notably better in the reason-drug relations, which were missed more than all other relation types.

We observed that external resources (ie, SFEs) contributed to the attribute extraction. Presumably, plentiful labeled data already available and complementary information from these external resources appear to have been helpful for performance. Nevertheless, simpler methods, such as PWE and rule-based methods, can match these sophisticated and expensive methods.

Error Analysis

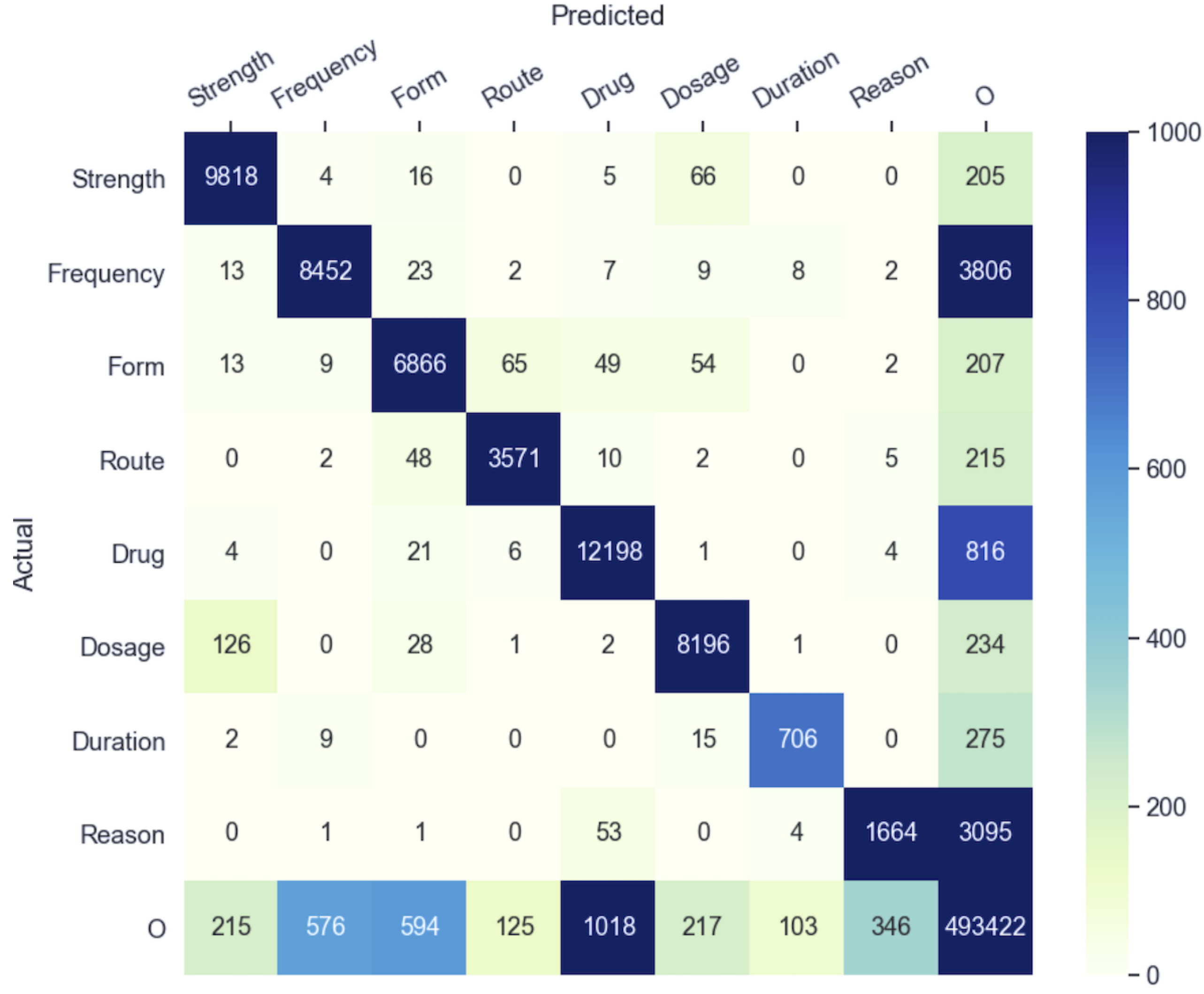
We further analyzed false positives and false negatives from the NER to obtain deeper insights into the common classification errors. Note that the focus in the error analysis was on the NER only, as it appears to be the main factor of the relatively low F-score in RE.

To gain an insight into where errors are made and how models can be improved, we manually reviewed false negatives (entities identified in the gold standard but incorrectly rejected, ie, missed, by the models) and false positives (entities identified by the models when they are not in the gold standard) in the best-performing model. Errors were then grouped into different categories based on their causes, including (1) context error: when an entity is captured as one of the drug-related attributes, although it is not, or when an entity is missing because of the context; (2) type error: when an attribute is extracted but with an incorrect annotation type; and (3) gold-standard error: possible error in the gold standard. We also generated a confusion matrix to subdivide the errors made by the method based on which type of mistake was made.

Context error was a major category of errors. These mostly resulted from previously unseen information (eg, “He was given a loading dose of amiodarone,” where the dosage *loading dose* was missed), atypical expression formats (eg, “One (1) Tablet,” where dosage *one (1)* was missing because of the parentheses), and abbreviations (eg, “Dig level 2.1,” where drug *dig*—which should be *digoxin*—was missed). Context errors may also result from the complexity of language expressions; for example, *200 units* in the phrase “was started on a 7d course of DRUG 200 units daily” could be a dosage when considered as a single phrase, or it could be 2 concepts: 200 (a strength) and unit (a form). Gold annotation preferred the latter, whereas our method identified the former.

Another interesting cause of error is the ambiguity between attributes, where an attribute is recognized, but the type is incorrect. Figure 5 presents the confusion matrix for the BiLSTM-CRF (PWE+CE) and indicates how often each entity is predicted. The confusion of dosage for strength and strength for dosage is the most frequent type of error, accounting for 28% ([66+126]/693) of the errors. The following example illustrates this type of error: “Meropenem 500 mg Intravenous every eight (8) hours.” The dosage *500 mg* is wrongly predicted as strength; usually, the mg unit is associated with strength. The substitution of dosage for strength is a common error, and these entities are often mislabeled as each other—both are often numeric quantities and used in similar contexts. A common solution for this issue is to merge these 2 types into 1 annotation type [59]. However, extracting them separately may be important for some applications.

Figure 5. Confusion matrix (token-level) from the output of bidirectional long-short term memory with conditional random field (with pretrained word embeddings and character embeddings) on the National NLP Clinical Challenges test set. The diagonal entries indicate labels that were correctly predicted, and the off-diagonal entries indicate errors. The total number of errors (sum of off-diagonal cells) was 693.



	Strength	Frequency	Form	Route	Drug	Dosage	Duration	Reason	O
Strength	9818	4	16	0	5	66	0	0	205
Frequency	13	8452	23	2	7	9	8	2	3806
Form	13	9	6866	65	49	54	0	2	207
Route	0	2	48	3571	10	2	0	5	215
Drug	4	0	21	6	12198	1	0	4	816
Dosage	126	0	28	1	2	8196	1	0	234
Duration	2	9	0	0	0	15	706	0	275
Reason	0	1	1	0	53	0	4	1664	3095
O	215	576	594	125	1018	217	103	346	493422

The second most frequent type of this error, which accounts for 16% ([48+65]/693) of the errors, is the confusion of form for route and route for form. These entities are often annotated as the gold standard in various ways. For example, the word *injection* is sometimes annotated as a form and sometimes as a route; in the training set, it is annotated as a form 68 times and as a route 53 times, which makes learning from these examples challenging.

The confusion of drugs with general words is one of the other sources of error. We found that there were several causes of this confusion among drug names. These include (1) generic drug names (eg, *glucose*, *IVF*, *blood*, *D5W*, and *chemo*) corresponding to prescribed medications but not occurring in expected contexts; (2) words such as *pressor*, *fluids*, *agents*, or *medication* that may be considered to be underspecified, but should be extracted, at least in this data set; (3) some classes of drugs (eg, *antiinflammatory drugs* and *hypertension medications*) missing in the training sources; (4) new drug names that did not occur within an expected context or semantic patterns (eg, *Dig level 2.5*), so they were not extracted by the NER methods; and (5) abbreviations (eg, *aspirin325* and *ABX*).

The analysis also showed a few potential omissions and inconsistencies in human annotations. Gold-standard errors fall into 2 different categories: missing in the gold standard and potential problems in gold standards. The more common error in this category is missing in the gold standard, where the method annotates entities that are not annotated in the data set. For example, *four weeks* in the phrase, “adding DRUG cover for the first four weeks of treatment,” is not annotated as a duration in the gold standard, whereas it appears to be a potentially correct attribute. Inconsistency may also appear in annotation spans; for example, dosage or strength, and form were annotated separately sometimes and jointly in others.

Conclusions

In this study, we constructed an end-to-end system (DrugEx) composed of bidirectional LSTM, CRF, and rule-based methods for extracting drug-related information from free-text discharge summaries. We studied various token representations (ie, WE, CE, and SFE) for extracting drug attributes from free-text discharge summaries. We also proposed a rule-based method for relations between drugs and attributes and compared this method with a context-aware deep learning method. The results showed that the proposed system can be used successfully for extracting and linking drug attributes in discharge summaries, although some attributes (ie, reason and duration) are still challenging. The results also showed that domain-tailored embeddings (ie, PWE) perform better than random embeddings (RIWE) in this task. Concatenating PWE with CE or SE achieved a comparable overall performance when compared between themselves. NER (PWE+CE) ranked best in F-score among other proposed models; however, NER (PWE+SE) performed better on some individual entity types, especially frequency, duration, and reason. Semantic embeddings also yielded a better balance between precision and recall. However, a simpler method (eg, WE and CE) can match these sophisticated and expensive methods. Incorporating external knowledge (eg, of a drug's reason, proposed treatment, and a drug's reactions) and incorporating a larger context may improve performance.

Concerning RE, the rule-based method achieved higher accuracy than the context-aware LSTM for most relations. Interestingly, the LSTM model performs notably better on some of the most challenging relations (eg, reason-drug).

In future work, we aim to investigate contextual embeddings, such as ELMo and BERT, which have been proven to provide considerable improvements in other tasks that include complex language structures, ambiguous word use, and unseen words in training. We also consider assessing the performance and transferability of the models across different biomedical data sets and tasks.

Finally, the medication NER and RE tasks are important not only from a research perspective but also because they have applications as steps in practical information extraction pipelines. The current level of performance indicates that these models should be good enough for large-scale statistical and epidemiological studies. However, applications that require patient-specific information may need NER systems with even higher recall and precision, ensemble and multiple-step systems (ie, systems that combine the output of multiple classifiers), or be subject to semiautomated verification.

Acknowledgments

This work was partially supported by the Saudi Arabian Ministry of Education, the Saudi Arabian Cultural Bureau in London, and the Healthcare Text Analytics Network (Heal-tex, grant EP/N027280/1, funded by the UK Engineering and Physical Sciences Research Council). The authors would like to thank Sumithra Velupillai and Natalia Viani (King's College London) for their discussions on the error analysis. The authors would also like to acknowledge the help from Haifa Alrdahi and Nikola Milosevic (University of Manchester) during their participation in the n2c2 shared task.

Authors' Contributions

GA and MB conducted the experiments and analyzed their output. GA drafted the manuscript. NP and GN revised the manuscript. All authors read and approved the final version of the manuscript. GN and NP supervised all steps of the work.

Conflicts of Interest

None declared.

References

1. Abhyankar S, Demner-Fushman D, Callaghan FM, McDonald CJ. Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. J Am Med Inform Assoc 2014;21(5):801-807 [FREE Full text] [doi: 10.1136/amiajnl-2013-001915] [Medline: 24384230]
2. Evans DA, Brownlow ND, Hersh WR, Campbell EM. Automating concept identification in the electronic medical record: an experiment in extracting dosage information. Proc AMIA Annu Fall Symp 1996:388-392 [FREE Full text] [Medline: 8947694]
3. Karystianis G. Extraction and representation of key characteristics from epidemiological literature. The University of Manchester. 2014. URL: https://tinyurl.com/bv927sfthttps://tinyurl.com/645sksnd [accessed 2021-03-31]
4. MacKinlay AD, Verspoor KM. Extracting structured information from free-text medication prescriptions using dependencies. In: Proceedings of the ACM sixth international workshop on Data and text mining in biomedical informatics. 2012 Presented at: CIKM'12: 21st ACM International Conference on Information and Knowledge Management; October, 2012; Maui Hawaii USA p. 35-40. [doi: 10.1145/2390068.2390076]
5. Sohn S, Clark C, Halgrim SR, Murphy SP, Chute CG, Liu H. MedXN: an open source medication extraction and normalization tool for clinical text. J Am Med Inform Assoc 2014;21(5):858-865 [FREE Full text] [doi: 10.1136/amiajnl-2013-002190] [Medline: 24637954]
6. Spasic I, Sarafraz F, Keane JA, Nenadic G. Medication information extraction with linguistic pattern matching and semantic rules. J Am Med Inform Assoc 2010;17(5):532-535 [FREE Full text] [doi: 10.1136/jamia.2010.003657] [Medline: 20819858]
7. Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010;17(5):514-518 [FREE Full text] [doi: 10.1136/jamia.2010.003947] [Medline: 20819854]

8. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc 2010;17(1):19-24 [FREE Full text] [doi: 10.1197/jamia.M3378] [Medline: 20064797]
9. Yang H. Automatic extraction of medication information from medical discharge summaries. J Am Med Inform Assoc 2010;17(5):545-548 [FREE Full text] [doi: 10.1136/jamia.2010.003863] [Medline: 20819861]
10. Krallinger M, Leitner F, Rabal O, Vazquez M, Oyarzabal J, Valencia A. CHEMDNER: the drugs and chemical names extraction challenge. J Cheminform 2015 Jan 19;7(S1). [doi: 10.1186/1758-2946-7-s1-s1]
11. Segura-Bedmar I, Martínez P, Herrero-Zazo M. SemEval-2013 Task 9 : extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In: Proceedings of the Seventh International Workhop on Semantic Evaluation (SemEval 2013) and Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2. 2013 Presented at: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2 and Seventh International Workshop on Semantic Evaluation (SemEval 2013); June, 2013; Atlanta, Georgia, USA p. 341-350.
12. Jagannatha A, Liu F, Liu W, Yu H. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). Drug Saf 2019 Jan;42(1):99-111 [FREE Full text] [doi: 10.1007/s40264-018-0762-z] [Medline: 30649735]
13. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inform Assoc 2020 Jan 01;27(1):3-12 [FREE Full text] [doi: 10.1093/jamia/ocz166] [Medline: 31584655]
14. Karystianis G, Sheppard T, Dixon WG, Nenadic G. Modelling and extraction of variability in free-text medication prescriptions from an anonymised primary care electronic medical record research database. BMC Med Inform Decis Mak 2016 Mar 09;16:18 [FREE Full text] [doi: 10.1186/s12911-016-0255-x] [Medline: 26860263]
15. Leaman R, Khare R, Lu Z. Challenges in clinical natural language processing for automated disorder normalization. J Biomed Inform 2015 Oct;57:28-37 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.010] [Medline: 26187250]
16. Kolárik C, Hofmann-Apitius M, Zimmermann M, Fluck J. Identification of new drug classification terms in textual resources. Bioinformatics 2007 Jul 01;23(13):264-272. [doi: 10.1093/bioinformatics/btm196] [Medline: 17646305]
17. Chhieng D, Day T, Gordon G, Hicks J. Use of natural language programming to extract medication from unstructured electronic medical records. AMIA Annu Symp Proc 2007 Oct 11:908. [Medline: 18694008]
18. Sirohi E, Peissig P. Study of effect of drug lexicons on medication extraction from electronic medical records. Pac Symp Biocomput 2005:308-318 [FREE Full text] [doi: 10.1142/9789812702456_0029] [Medline: 15759636]
19. Lowe DM, Sayle RA. LeadMine: a grammar and dictionary driven approach to entity recognition. J Cheminform 2015 Jan 19;7(S1). [doi: 10.1186/1758-2946-7-s1-s5]
20. Gold S, Elhadad N, Zhu X, Cimino JJ, Hripcsak G. Extracting structured medication event information from discharge summaries. AMIA Annu Symp Proc 2008 Nov 06:237-241 [FREE Full text] [Medline: 18999147]
21. Hamon T, Grabar N. Linguistic approach for identification of medication names and related information in clinical narratives. J Am Med Inform Assoc 2010;17(5):549-554 [FREE Full text] [doi: 10.1136/jamia.2010.004036] [Medline: 20819862]
22. Xu R, Morgan A, Das AK, Garber A. Investigation of unsupervised pattern learning techniques for bootstrap construction of a medical treatment lexicon. In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing. 2009 Presented at: BioNLP '09: Workshop on Current Trends in Biomedical Natural Language Processing; June 4-5, 2009; Boulder, Colorado p. 63-70. [doi: 10.3115/1572364.1572373]
23. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. J Am Med Inform Assoc 2010;17(5):524-527 [FREE Full text] [doi: 10.1136/jamia.2010.003939] [Medline: 20819856]
24. Leaman R, Wei C, Lu Z. tmChem: a high performance approach for chemical named entity recognition and normalization. J Cheminform 2015 Jan 19;7(S1). [doi: 10.1186/1758-2946-7-s1-s3]
25. Lu Y, Ji D, Yao X, Wei X, Liang X. CHEMDNER system with mixed conditional random fields and multi-scale word clustering. J Cheminform 2015 Jan 19;7(S1). [doi: 10.1186/1758-2946-7-s1-s4]
26. Campos D, Matos S, Oliveira JL. A document processing pipeline for annotating chemical entities in scientific documents. J Cheminform 2015 Jan 19;7(S1). [doi: 10.1186/1758-2946-7-s1-s7]
27. Lamurias A, Grego T, Couto FM. Chemical compound and drug name recognition using CRFs and semantic similarity based on ChEBI. Washington, DC USA: BioCreative challenge evaluation workshop, vol. 2; 2013. URL: https://biocreative.bioinformatics.udel.edu/media/store/files/2013/bc4_v2_9.pdf [accessed 2021-03-31]
28. Sikdar UK, Ekbal A, Saha S. Domain-independent model for chemical compound and drug name recognition. Washington, DC USA: BioCreative Challenge Evaluation Workshop. Vol 2; 2013. URL: https://biocreative.bioinformatics.udel.edu/media/store/files/2013/bc4_v2_22.pdf [accessed 2021-03-31]
29. Akhondi SA, Hettne KM, van der Horst E, van Mulligen EM, Kors JA. Recognition of chemical entities: combining dictionary-based and grammar-based approaches. J Cheminform 2015 Jan 19;7(S1). [doi: 10.1186/1758-2946-7-s1-s10]
30. He L, Yang Z, Lin H, Li Y. Drug name recognition in biomedical texts: a machine-learning-based method. Drug Discov Today 2014 May;19(5):610-617. [doi: 10.1016/j.drudis.2013.10.006] [Medline: 24140287]

31. Tikk D, Solt I. Improving textual medication extraction using combined conditional random fields and rule-based systems. J Am Med Inform Assoc 2010;17(5):540-544 [FREE Full text] [doi: 10.1136/jamia.2010.004119] [Medline: 20819860]
32. Korkontzelos I, Piliouras D, Dowsey AW, Ananiadou S. Boosting drug named entity recognition using an aggregate classifier. Artif Intell Med 2015 Oct;65(2):145-153 [FREE Full text] [doi: 10.1016/j.artmed.2015.05.007] [Medline: 26116947]
33. Liu Z, Yang M, Wang X, Chen Q, Tang B, Wang Z, et al. Entity recognition from clinical texts via recurrent neural network. BMC Med Inform Decis Mak 2017 Jul 05;17(Suppl 2):67 [FREE Full text] [doi: 10.1186/s12911-017-0468-7] [Medline: 28699566]
34. Jagannatha AN, Yu H. Structured prediction models for RNN based sequence labeling in clinical text. Proc Conf Empir Methods Nat Lang Process 2016 Nov;2016:856 [FREE Full text] [doi: 10.18653/v1/d16-1082] [Medline: 28004040]
35. Yang X, Bian J, Fang R, Bjarnadottir RI, Hogan WR, Wu Y. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. J Am Med Inform Assoc 2020 Jan 01;27(1):65-72 [FREE Full text] [doi: 10.1093/jamia/ocz144] [Medline: 31504605]
36. Wei Q, Ji Z, Li Z, Du J, Wang J, Xu J, et al. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. J Am Med Inform Assoc 2020 Jan 01;27(1):13-21 [FREE Full text] [doi: 10.1093/jamia/ocz063] [Medline: 31135882]
37. Ju M, Nguyen NT, Miwa M, Ananiadou S. An ensemble of neural models for nested adverse drug events and medication extraction with subwords. J Am Med Inform Assoc 2020 Jan 01;27(1):22-30 [FREE Full text] [doi: 10.1093/jamia/ocz075] [Medline: 31197355]
38. Dai HJ, Su CH, Wu CS. Adverse drug event and medication extraction in electronic health records via a cascading architecture with different sequence labeling models and word embeddings. J Am Med Inform Assoc 2020 Jan 01;27(1):47-55 [FREE Full text] [doi: 10.1093/jamia/ocz120] [Medline: 31334805]
39. Oleynik M, Kugic A, Kasáč Z, Kreuzthaler M. Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. J Am Med Inform Assoc 2019 Nov 01;26(11):1247-1254 [FREE Full text] [doi: 10.1093/jamia/ocz149] [Medline: 31512729]
40. Christopoulou F, Tran TT, Sahu SK, Miwa M, Ananiadou S. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. J Am Med Inform Assoc 2020 Jan 01;27(1):39-46 [FREE Full text] [doi: 10.1093/jamia/ocz101] [Medline: 31390003]
41. Kim Y, Meystre SM. Ensemble method-based extraction of medication and related information from clinical texts. J Am Med Inform Assoc 2020 Jan 01;27(1):31-38 [FREE Full text] [doi: 10.1093/jamia/ocz100] [Medline: 31282932]
42. 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: 26th International Conference on Neural Information Processing Systems - Volume 2; December 2013; Lake Tahoe, Nevada, United States p. 3111-3119 URL: http://dl.acm.org/citation.cfm?id=2999792.2999959
43. 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 (EMNLP). 2014 Presented at: Conference on Empirical Methods in Natural Language Processing (EMNLP); October, 2014; Doha, Qatar. [doi: 10.3115/v1/d14-1162]
44. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguistics 2017 Dec;5:135-146. [doi: 10.1162/tacl_a_00051]
45. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018 Presented at: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers); June, 2018; New Orleans, Louisiana p. 2227-2237. [doi: 10.18653/v1/n18-1202]
46. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. 2018. URL: https://arxiv.org/abs/1810.04805 [accessed 2021-03-31]
47. n2c2 NLP research data sets. Harvard Medical School. 2018. URL: https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/ [accessed 2021-03-31]
48. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
49. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Inform Assoc 2018 Mar 01;25(3):331-336 [FREE Full text] [doi: 10.1093/jamia/ocx132] [Medline: 29186491]
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. Kocmi T, Bojar O. SubGram: extending skip-gram word representation with substrings. In: Text, Speech, and Dialogue. Switzerland: Springer; 2016:182-189.

52. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv. 2013. URL: https://arxiv.org/abs/1301.3781 [accessed 2021-03-31]
53. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. J Am Med Inform Assoc 2017 May 01;24(3):596-606 [FREE Full text] [doi: 10.1093/jamia/ocw156] [Medline: 28040687]
54. Luo Y, Cheng Y, Uzuner O, Szolovits P, Starren J. Segment convolutional neural networks (Seg-CNNs) for classifying relations in clinical notes. J Am Med Inform Assoc 2018 Jan 01;25(1):93-98 [FREE Full text] [doi: 10.1093/jamia/ocx090] [Medline: 29025149]
55. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res. 2012. URL: https://www.jmlr.org/papers/v13/bergstra12a.html [accessed 2021-03-31]
56. Sorokin D, Gurevych I. Context-aware representations for knowledge base relation extraction. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017 Presented at: Conference on Empirical Methods in Natural Language Processing; September, 2017; Copenhagen, Denmark p. 1784-1789. [doi: 10.18653/v1/d17-1188]
57. Cunningham H, Tablan V, Roberts A, Bontcheva K. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. PLoS Comput Biol 2013;9(2) [FREE Full text] [doi: 10.1371/journal.pcbi.1002854] [Medline: 23408875]
58. Xu J, Lee H, Ji Z, Wang J, Wei Q, Xu H. UTH_CCB system for adverse drug reaction extraction from drug labels at TAC-ADR. 2017. URL: https://tinyurl.com/645sksnd [accessed 2021-03-31]
59. Demner-Fushman D, Mork JG, Rogers WJ, Shooshan SE, Rodriguez L, Aronson AR. Finding medication doses in the liteature. AMIA Annu Symp Proc 2018;2018:368-376 [FREE Full text] [Medline: 30815076]

Abbreviations

ADE: adverse drug event
BERT: Bidirectional Encoder Representations from Transformers
BiLSTM: bidirectional long-short term memory
BiLSTM-CRF: bidirectional long-short term memory with conditional random field
BIOES: Begin, Inside, Outside, End, Single
CE: character embedding
CLAMP: Clinical Language Annotation, Modeling, and Processing Toolkit
CRF: conditional random field
cTAKES: Clinical Text Analysis and Knowledge Extraction System
EHR: electronic health record
ELMo: Embeddings from Language Models
LSTM: long-short term memory
MIMIC-III: Medical Information Mart for Intensive Care III
n2c2: National NLP Clinical Challenges
NER: named entity recognition
NLP: natural language processing
PWE: pretrained word embedding
RE: relation extraction
RIWE: randomly initialized word embedding
SFE: semantic-feature embedding
WE: word embedding

Edited by C Lovis; submitted 30.09.20; peer-reviewed by S Fu, M Torii; comments to author 03.11.20; revised version received 15.02.21; accepted 20.02.21; published 05.05.21.

Please cite as:
Alfattni G, Belousov M, Peek N, Nenadic G
Extracting Drug Names and Associated Attributes From Discharge Summaries: Text Mining Study
JMIR Med Inform 2021;9(5):e24678
URL: https://medinform.jmir.org/2021/5/e24678
doi:10.2196/24678
PMID:33949962

Original Paper

An Attention Model With Transfer Embeddings to Classify Pneumonia-Related Bilingual Imaging Reports: Algorithm Development and Validation

Hyung Park[1*], MD; Min Song[2*], PhD; Eun Byul Lee[2], BA; Bo Kyung Seo[2], BA; Chang Min Choi[1,3], MD

[1]Department of Pulmonary and Critical Care Medicine, Asan Medical Center, Seoul, Republic of Korea

[2]Yonsei University, Seoul, Republic of Korea

[3]Department of Oncology, Asan Medical Center, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Chang Min Choi, MD
Department of Pulmonary and Critical Care Medicine
Asan Medical Center
Olympic-ro 43-gil
Seoul, 05505
Republic of Korea
Phone: 82 2 3010 5902
Fax: 82 2 3010 6968
Email: ccm9607@gmail.com

Abstract

Background: In the analysis of electronic health records, proper labeling of outcomes is mandatory. To obtain proper information from radiologic reports, several studies were conducted to classify radiologic reports using deep learning. However, the classification of pneumonia in bilingual radiologic reports has not been conducted previously.

Objective: The aim of this research was to classify radiologic reports into pneumonia or no pneumonia using a deep learning method.

Methods: A data set of radiology reports for chest computed tomography and chest x-rays of surgical patients from January 2008 to January 2018 in the Asan Medical Center in Korea was retrospectively analyzed. The classification performance of our long short-term memory (LSTM)–Attention model was compared with various deep learning and machine learning methods. The area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve, sensitivity, specificity, accuracy, and F1 score for the models were compared.

Results: A total of 5450 radiologic reports were included that contained at least one pneumonia-related word. In the test set (n=1090), our proposed model showed 91.01% (992/1090) accuracy (AUROCs for negative, positive, and obscure were 0.98, 0.97, and 0.90, respectively). The top 3 performances of the models were based on FastText or LSTM. The convolutional neural network–based model showed a lower accuracy 73.03% (796/1090) than the other 2 algorithms. The classification of negative results had an F1 score of 0.96, whereas the classification of positive and uncertain results showed a lower performance (positive F1 score 0.83; uncertain F1 score 0.62). In the extra-validation set, our model showed 80.0% (642/803) accuracy (AUROCs for negative, positive, and obscure were 0.92, 0.96, and 0.84, respectively).

Conclusions: Our method showed excellent performance in classifying pneumonia in bilingual radiologic reports. The method could enrich the research on pneumonia by obtaining exact outcomes from electronic health data.

(JMIR Med Inform 2021;9(5):e24803) doi:10.2196/24803

KEYWORDS

deep learning; natural language process; attention; clinical data; pneumonia; classification; medical imaging; electronic health record; machine learning; model

Introduction

Electronic health records (EHRs) have become increasingly incorporated into clinical practices in hospitals over the past few decades [1]. EHR data are voluminous and can be used as real-world evidence if they are analyzed with proper methods [2]. However, the data are not collected for research purposes [2], and several rule-based methods are used to extract particular outcomes from the data set. There have been numerous studies where analyses were performed using EHR data with labels such as *sepsis* defined by rule-based outcomes [3-6]. However, defining outcomes other than laboratory findings is difficult because the data are unstructured and written as natural language. For this reason, a previous study that used the outcome *pneumonia* defined pneumonia by its International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) code [7,8]. However, the use of ICD codes as a label does not contain temporal information, such as the exact time of diagnosis during hospital admission, and it is hard to perform time series analysis with this limited information.

Although medical imaging reports contain a great deal of information regarding diagnosis and clinical features, it is hard to analyze the information because they are formatted as unstructured free text and are variably written depending on the radiologist.[9] For this reason, medical imaging reports are rarely used as outcomes in big data analysis [10]. However, as long as pneumonia can be identified in radiologic reports, other important information, such as the time of onset and the presence of pneumonia during admission, can also be derived. Moreover, labeled data are essential in deep learning because the analysis requires millions of observations to reach acceptable performance levels [11].

As of 2018, 43 studies using natural language processing for the identification of chronic diseases in EHRs had been published, and only recently have there been more studies conducted on this topic using deep learning [12]. Especially in deep learning, convolutional neural network (CNN)–based models have shown significant accuracy in extracting pulmonary embolism [10] and pulmonary infection from medical reports [1]. The model can be used to classify diagnosis from whole medical records even when they are written in the Chinese language [13], and a recurrent neural network–based model has been used for classifying stroke and identifying its location [14]. However, the use of bilingual clinical reports is common for EHRs in non–English-speaking countries.

The purpose of our study was to classify reports of pneumonia consisting of findings derived during the pre- and postoperative period of a major surgery that were written as bilingual texts (English and Korean). We compared the performance of traditional models with deep learning models, with the latter showing excellent performance in previous studies, and identified the best performing model as an attention-based bidirectional long short-term memory (Bi-LSTM) model neural network.

Methods

Clinical Data

We retrospectively included radiology reports for chest computed tomography (CT) and chest x-rays of surgical patients from January 2008 to January 2018 in the Asan Medical Center in Korea. The patients had undergone upper abdominal and thoracic surgeries, as coded by the ICD-9-CM. Detailed criteria for the surgery are described in Multimedia Appendix 1.

The radiology reports consist of chest CT and chest x-rays (posteroanterior and anteroposterior) that are extracted by radiology procedure codes. The chest x-ray reports have no structured format and only contain descriptions. The chest CT reports consist of the short history of the patients, the findings, and a conclusion; however, the format varies depending on the writing style of the radiologist. The conclusions in around half of the chest CT reports were omitted due to the different writing style of the radiologists. Therefore, we used only the findings of chest CT and the descriptions of chest x-rays to classify the labels, and all the annotation was based solely on the description of each report.

Usually, the pneumonia incidence in surgical patients is around 1%, suggesting that reports of pneumonia are rare. To overcome the imbalance of the positive and negative data sets, we only included radiologic reports that contained pneumonia-related words. The words representing pneumonia were as follows: “pneumoni-,” “consolid-,” “infiltra-,” “bronchiole-,” “hazi-,” “hazzi-,” “opacit-,” and “GGO”.

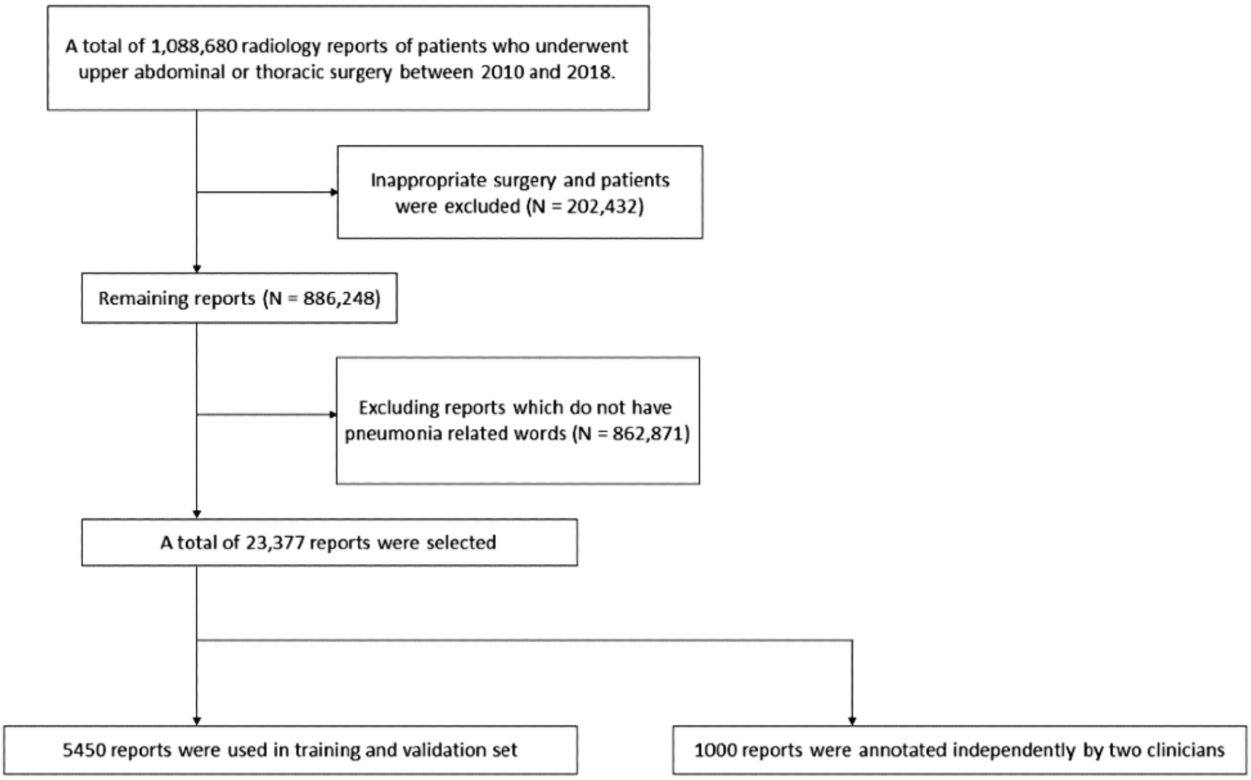
From a total of 1,088,680 radiology reports, 886,248 were included after reports with inappropriate surgical procedures were excluded. The detailed inclusion criteria of the appropriate procedures have been described in a previous study [3]. After extracting the pneumonia-related words, 23,377 reports were included.

Report Annotation

Among the 23,377 reports, a total of 5450 annotated reports were used to train our model. A clinician annotated the 5450 reports and used them for training and validation. After training the model, 2 different clinicians, who worked independently from the first clinician, annotated another 1000 reports for an extra-validation set (Figure 1).

All document-level annotations by clinicians included 3 categories for pneumonia: negative, positive, and unclear (obscure). The positive pneumonia reports included postoperative infection reports and did not contain reports for noninfectious diseases, such as organizing pneumonia or interstitial lung disease, because the label was required to represent pneumonia as a perioperative complication. The excluded reports were labeled as negative reports. It was observed that 895 reports were pneumonia positive, 4005 reports were pneumonia negative, and 550 reports were obscure results. In the extra-validation set, 2 clinicians independently labeled the radiologic reports on the basis of the clinical importance of the findings. To overcome the human error of the 2 clinicians, the consensus label of the 2 clinicians was regarded as the reference standard. An interrater reliability (k score) was calculated by Cohen κ value.

Figure 1. Radiologic reports flowchart.


A total of 1,088,680 radiology reports of patients who underwent upper abdominal or thoracic surgery between 2010 and 2018.
Inappropriate surgery and patients were excluded (N = 202,432)
Remaining reports (N = 886,248)
Excluding reports which do not have pneumonia related words (N = 862,871)
A total of 23,377 reports were selected
5450 reports were used in training and validation set
1000 reports were annotated independently by two clinicians

Ethics Approval

This study was approved by the ethics committee of the Asan Medical Center (approval no. 2018-1122), and the need to obtain informed consent was waived because of the retrospective observational nature of the study. The clinical data that were extracted using the Asan Biomedical Research Environment system were indexed by deidentified encrypted patient ID numbers so that the researchers would not be able identify the patients [15,16].

Proposed Approach

As most of the verbs and adjectives in clinical reports are written in Korean, and most of nouns (usually the names of the diseases) are written in English, we had to consider 2 different languages. Therefore, we proposed a new method for a bilingual clinical data set based on the classification algorithm of combining substring and translation embeddings (Kor2Eng) with an attention-based Bi-LSTM neural network (LSTM-Attention). Multimedia Appendix 1 Figure S4 shows the architecture of our proposed model.

The proposed method includes 3 steps: (1) text preprocessing; (2) word representation, which is composed of substring and Korean-to-English (Kor2Eng) embeddings; and (3) training of the classification model.

Our data set, which is a description of x-ray and CT, is composed of a mix of Korean and English sentences. Therefore, specific preprocessing is required before the statements are fed into the classification model. The detailed methods for text preprocessing and training are described in Multimedia Appendix 1.

Kor2Eng Transfer Embedding

Training word vectors require a considerable amount of data and time. Therefore, we applied embeddings by training them independently on monolingual data and pretraining them with Wikipedia data. However, due to the characteristics of data, the text of the clinical notes was a mixture of English and Korean. If a monolingual embedding were to be used for this data, one side of the information would be lost. To reduce the loss of information, we used a translation method that converts the vector of Korean words into the vector of English words with similar meanings. The unsupervised method of translating the source language into the target language was proposed by Lample et al [17]. In this method, the process of learning a mapping occurs between the 2 sets of embedding in the shared space. We trained the subword embedding model to learn Korean-to-English mapping using the unsupervised method without any parallel data.

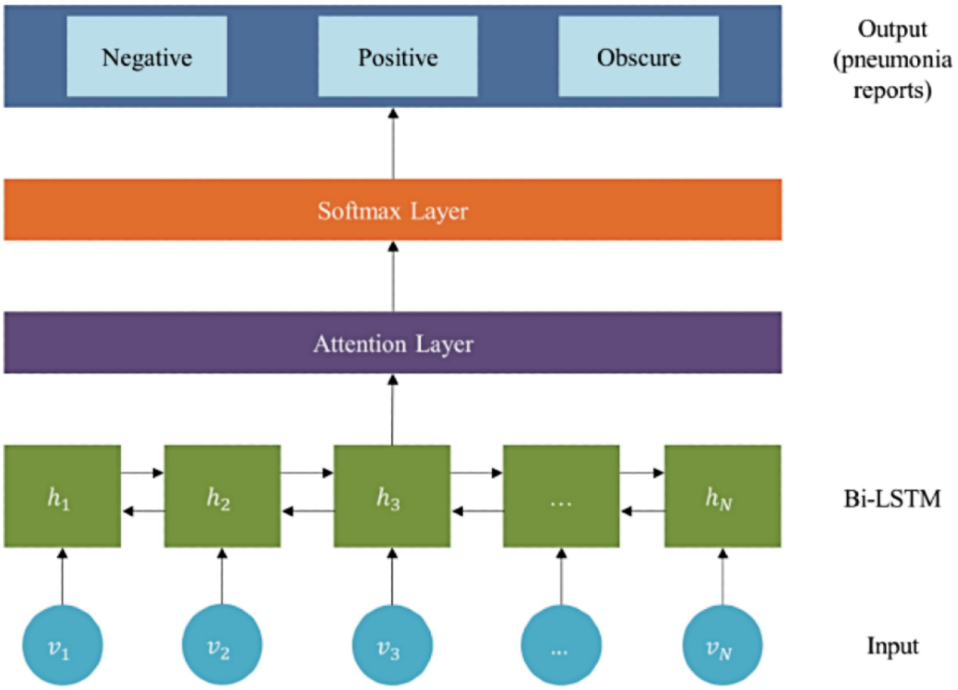
Deep Learning–Based Classification Model

We built an attention-based deep neural network using LSTM. LSTM is a recurrent neural network variant that alleviates the vanishing gradient problem by learning and remembering long-term dependencies [18] and consists of a cell memory state and 3 gates.

The Bi-LSTM consists of a forward–backward LSTM layer [19]. Both layers are connected to the same output layer. Our classification model used Bi-LSTM with the attention mechanism. This allowed the model to simultaneously handle information from different positions.

Figure 2 shows the architecture of the deep learning–based classification model. First, the input is fed into the Bi-LSTM layer. Second, the output of the Bi-LSTM layer is fed into the attention layer (Bi-LSTM–Attention) for attending important words. Finally, the output of the attention weight passes through the softmax layer for classification.

Figure 2. The architectures of a deep learning-based classification model. Each input receives an embedding of English translated from Korean. In the attention layer, each word has an attention weight which is translated into the importance for prediction. Bi-LSTM: bidirectional long short-term memory model.



The performance metrics (ie, precision, recall [sensitivity], and F_1 score) were used to evaluate the models. The accuracy, area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC) were used to compare the models. For analyzing the multilabel data set, labels were treated as interested labels and other labels in evaluating each metric. For example, when we treated the precision for negative labels, only the true negative data were treated as true labels while positive and obscure labels were treated as false labels. F_1 score is the weighted average of precision and recall, and it is used to measure the performance of a model when the data consist of uneven class distributions [20]. The statistical analysis was performed on Python 3.7.6 (Python Software Foundation).

Results

In this section, we evaluated the performance of the various classification models. To demonstrate the performance of our method, we compare the proposed model with traditional machine learning and other deep learning models. The machine learning models included logistic regression [21], support vector machine [22], Naïve Bayes regression [23], K-nearest neighbors algorithm [24], decision tree [25], and random forest [26]. The deep learning models included the word-to-vector representation model (Word2Vec) [27], FastText [17], CNN [28], and LSTM [29]. The details of each model are described in Multimedia Appendix 1.

Out of 5450 data sets, 4005 did not contain pneumonia, 895 contained pneumonia, and 550 were obscure, with 80% being used in the training set and the remaining 20% in test set. The test set was composed of no pneumonia (n=801), pneumonia (n=179), and obscure (n=110) classifications. The extra-validation set was annotated by 2 independent clinicians. Out of a total of 1000 radiologic reports, 803 labels were agreed upon by 2 independent clinicians. Among these labels, 498 did not contain pneumonia, 185 contained pneumonia, and 120 were obscure cases.

Accuracy of Our Model as Compared to Previous Models

We evaluated the performance of the different models to find the best model. As shown in Table 1, the prediction accuracy changed depending on the model. The traditional models (ie, support vector machine, Naïve Bayes, etc) achieved an accuracy between 64.03% and 83.03%. The logistic regression showed a reasonable performance with an accuracy of 83.03% (Multimedia Appendix 1 Table S1).

The deep learning–based methods (ie, FastText, Word2Vec with Bi-LSTM–Attention, and the proposed model) outperformed the traditional models. The prediction accuracy of the deep learning models was 90.00%, 88.99%, and 91.01% for FastText, Word2Vec with Bi-LSTMAttention, and the proposed model, respectively. These deep learning models showed a 10% higher accuracy than did the traditional machine learning methods because sentence classification required the interpretation of complex features. The proposed model achieved the highest performance compared to the other deep learning models (Multimedia Appendix 1 Table S1).

Model Accuracy Based on the Different Representation Methods of Words

We evaluated the performance based on different methods of word representation. The Word2Vec with Bi-LSTM–Attention model is a more commonly used language representation model. The model showed a higher accuracy and F_1 score than did the traditional models; however, the drawback associated with this model is that the foreign language is not represented (Table 1). We implemented another representation method with a substring using the FastText model. This method involves slicing of words to bunches of characters, which can be a better expression for the foreign language. The substring with FastText model achieved a precision of 93% for negative, 84% for positive, and 74% for obscure classifications; and a recall of 93% for negative, 84% for positive, and 47% for obscure classifications. The substring with FastText model showed a better performance than did the Word2Vec model according to F_1 score.

Our proposed model (Kor2Eng) translated Korean to English before the prediction process. The proposed model achieved a precision of 96%, 86%, and 61%, and a recall of 97%, 80%, and 64% for positive, negative, and obscure classifications, respectively. The AUROC of the model was 0.98 for negative, 0.97 for positive, and 0.90 for obscure classifications, while the AUPRC was 0.99 for negative, 0.87 for positive, and 0.62 for obscure classifications (Multimedia Appendix 1 Figure S5). Compared to the classification of the negative labels, which was a relatively easy task (96% of negative), classifying positive or obscure labels was a harder task and showed a rather lower F_1 score (83% for positive and 62% for obscure). For classifying the obscure classification, our model showed the highest performance among different representation methods (substring with FastText, Word2Vec, and Kor2Eng).

Table 1. The detailed performance of the top 3 best-performing models.

Models	Precision, n/N (%)	Recall, n/N (%)	F_1 score (%)	AUROC[a]	AUPRC[b]
Substring+FastText [17]					
Negative	776/819 (94.7)	776/801 (96.9)	96	0.82	0.92
Positive	153/593 (25.8)	153/179 (85.5)	83	0.74	0.34
Obscure	52/73 (71.2)	52/110 (47.3)	57	0.71	0.22
Word2Vec[c]+Bi-LSTM[d]–Attention					
Negative	772/849 (90.9)	772/801 (96.4)	94	0.95	0.98
Positive	153/222 (68.9)	153/179 (85.5)	81	0.96	0.87
Obscure	47/80 (58.8)	47/110 (42.7)	49	0.88	0.51
Proposed model (Kor2Eng[e])					
Negative	776/809 (95.9)	776/801 (96.9)	96	0.98	0.99
Positive	153/182 (84.1)	153/179 (85.5)	83	0.97	0.87
Obscure	70/115 (60.9)	70/110 (63.6)	62	0.90	0.62

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

[b] AUPRC: area under the precision-recall curve.

[c] Word2Vec: the word-to-vector representation model.

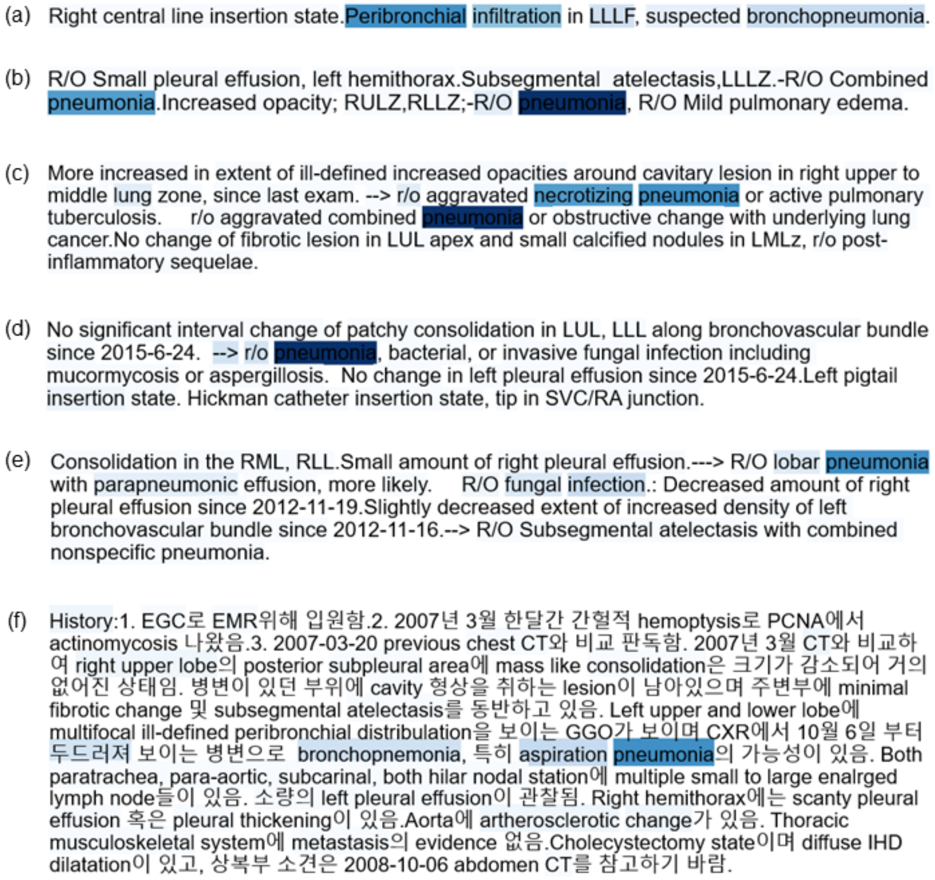
[d] Bi-LSTM: bidirectional long short-term memory model.

[e] Kor2Eng: Korean to English.

Visualization of Relative Importance

We visualized the weighted words when the proposed model classified the input data. In the attention model, the weight of each word could be used for classifying the reports. Based on the intensity of color, the importance of a word was indicated when the proposed model determined the class of the input data. Darker colors indicated a higher importance for classifying pneumonia. Figure 3 shows the instances where the proposed model predicted pneumonia reports correctly. For example, the highlighted words “Peribronchial,” “infiltration,” “suspected,” and “bronchopneumonia” indicate pneumonia (Figure 3a). In the bilingual texts (Figure 3f), the following words are important to classifying pneumonia-reports: “두드러져,” “bronchopneumonia,” “aspiration,” and “pneumonia.”

Figure 3. Visualization of the importance of words by attention weights.The darker the color is, the greater the importance of the words for predicting the pneumonia label. High attention weight is depicted in the darker color. Words with high attention weights are shown.



(a) Right central line insertion state.Peribronchial infiltration in LLLF, suspected bronchopneumonia.

(b) R/O Small pleural effusion, left hemithorax.Subsegmental atelectasis,LLLZ.-R/O Combined pneumonia.Increased opacity; RULZ,RLLZ;-R/O pneumonia, R/O Mild pulmonary edema.

(c) More increased in extent of ill-defined increased opacities around cavitary lesion in right upper to middle lung zone, since last exam. --> r/o aggravated necrotizing pneumonia or active pulmonary tuberculosis. r/o aggravated combined pneumonia or obstructive change with underlying lung cancer.No change of fibrotic lesion in LUL apex and small calcified nodules in LMLz, r/o post-inflammatory sequelae.

(d) No significant interval change of patchy consolidation in LUL, LLL along bronchovascular bundle since 2015-6-24. --> r/o pneumonia, bacterial, or invasive fungal infection including mucormycosis or aspergillosis. No change in left pleural effusion since 2015-6-24.Left pigtail insertion state. Hickman catheter insertion state, tip in SVC/RA junction.

(e) Consolidation in the RML, RLL.Small amount of right pleural effusion.---> R/O lobar pneumonia with parapneumonic effusion, more likely. R/O fungal infection.: Decreased amount of right pleural effusion since 2012-11-19.Slightly decreased extent of increased density of left bronchovascular bundle since 2012-11-16.--> R/O Subsegmental atelectasis with combined nonspecific pneumonia.

(f) History:1. EGC로 EMR위해 입원함.2. 2007년 3월 한달간 간헐적 hemoptysis로 PCNA에서 actinomycosis 나왔음.3. 2007-03-20 previous chest CT와 비교 판독함. 2007년 3월 CT와 비교하여 right upper lobe의 posterior subpleural area에 mass like consolidation은 크기가 감소되어 거의 없어진 상태임. 병변이 있던 부위에 cavity 형상을 취하는 lesion이 남아있으며 주변부에 minimal fibrotic change 및 subsegmental atelectasis를 동반하고 있음. Left upper and lower lobe에 multifocal ill-defined peribronchial distribulation을 보이는 GGO가 보이며 CXR에서 10월 6일 부터 두드러져 보이는 병변으로 bronchopnemonia, 특히 aspiration pneumonia의 가능성이 있음. Both paratrachea, para-aortic, subcarinal, both hilar nodal station에 multiple small to large enalrged lymph node들이 있음. 소량의 left pleural effusion이 관찰됨. Right hemithorax에는 scanty pleural effusion 혹은 pleural thickening이 있음.Aorta에 artherosclerotic change가 있음. Thoracic musculoskeletal system에 metastasis의 evidence 없음.Cholecystectomy state이며 diffuse IHD dilatation이 있고, 상복부 소견은 2008-10-06 abdomen CT를 참고하기 바람.

Extra Validations

As an extra validation of our proposed model, 2 clinicians labeled an additional data set. The data set was randomly selected from the entire data set, excluding the previously trained data. For precise labeling, 2 medical doctors each labeled the records. Of the 1000 records, 803 were agreed upon by 2 independent physicians. The Cohen κ value of the clinicians' label was 0.63 (95% CI 0.59-0.67). Table 2 shows the performance results of the proposed model with the extra-validation data set. The AUROC and AUPRC for positive labels were slightly lower in the extra-validation set than in the test set (Figure 4). The F_1 score of positive labels was similar to that of the training data; however, predicting negative and obscure labels showed a relatively poor performance as compared to the training data set according to F_1 score. The overall accuracy of our model was 80.0%.

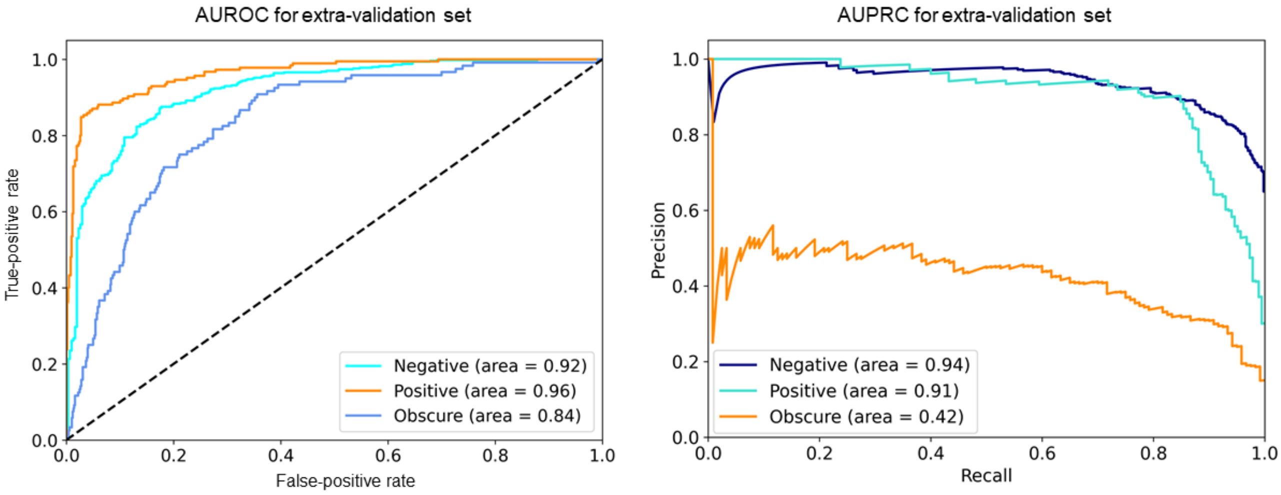
Table 2. Extra validation of the proposed Korean-to-English (Kor2Eng) model.

Class	Precision, n/N (%)	Recall, n/N (%)	F_1 score	AUROC[a]	AUPRC[b]
Negative	422/470 (89.8%)	422/498 (84.7%)	87%	0.92	0.94
Positive	142/155 (91.6%)	142/185 (76.8%)	84%	0.96	0.91
Obscure	77/178 (43.3%)	77/120 (64.2%)	52%	0.84	0.42

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

[b] AUPRC: area under the precision-recall curve

Figure 4. AUROC and AUPRC of our proposed model in the extra-validation set. AUROC: area under the receiver operating characteristic curve; AUPRC: area under the precision-recall curve.



Discussion

The purpose of the Kor2Eng model is to classify pneumonia-related medical records written in Korean and English. Our proposed model showed 91.01% accuracy in the test set and 80.0% accuracy in the extra-validation set for classifying pneumonia reports. Appropriate classification of radiologic reports is mandatory for further analysis regarding pneumonia through EMRs. As compared to other models, such as CNN or traditional machine learning models, our model showed better performance. The 3 best-performing models (Word2Vec with Bi-LSTM–Attention, FastText, and the proposed model) demonstrated better performance than did the traditional and CNN models, and our proposed model provided the highest AUROC and AUPRC among the top 3 models. Because too many false-positives may lead to clinician exhaustion, a model with excellent performance is desirable. We consider that a model with an AUROC of at least 0.95 can be used in clinical practice or for labeling the data set. The false-positive results of pneumonia reports can be additionally filtered with other clinical findings such as respiratory symptoms or antibiotics use, as pneumonia is defined by respiratory symptoms with radiologic findings [30].

The label balance of the data set was a consequence of excluding irrelevant labels to our target. As the reports that do not have pneumonia-related words can be considered pneumonia-negative radiologic reports, the reports requiring classification must contain at least one of the pneumonia-related words such as “consolidation” or “haziness”. Excluding the irrelevant label is clinically appropriate and balances the data set with each label, with the balanced data set mitigating the overestimation of the model. Furthermore, filtering radiologic reports containing relevant words might make the data set rather homogenous, which makes classification a hard task. Our model showed an excellent performance in classifying pneumonia, and thus, it can be used for auto-labeling in classifying pneumonia reports.

A notable observation is the discrepancy between the test and extra-validation set. The model showed a rather similar performance in classifying negative and positive cases and a relatively poor performance in obscure cases. One reason for this discrepancy might be that 2 different clinicians annotated the entire extra-validation set. As some of the obscure cases are classified by the nuance of the context, the 2 clinicians might have differed in labeling the obscure cases. Therefore, the labeling of the obscure classification in the extra-validation set might have been different from that of the training set. The pneumonia cases in the report should only be decided by clinical situations, and thus, the importance of obscure cases should be evaluated in subsequent studies.

Several studies have been conducted for classifying radiologic reports as positive or negative for a given disease [1,10,31,32] or for classifying various diagnoses from medical records written in Chinese [13]. Most of the studies used a CNN-based model and showed a better performance than did our model [1,10,31,32]. In our study, we compared several deep learning models from logistic regression to LSTM with attention. The CNN model, which showed an excellent performance in previous studies [1,10,31,32], was inferior to the attention-based LSTM model in our data set. The reason for its relatively poor performance might be explained by our data selection. We selected radiologic reports that had at least one of the pneumonia-related words. This selection made the radiologic reports relatively homogeneous compared to those used in previous studies, which might contain a wider variety of radiologic reports. As we compared the performance with the CNN model, our proposed model was found to be comparably accurate with those of previous studies and showed better performance.

Radiologic reports in this study consisted of 2 languages: English and Korean. Compared to the English data set, the Korean word data set has a lack of studies in embedding and analyzing in deep learning. To overcome this limitation, we used unsupervised translation of Korean words to English words, which had pretrained embedding [17]. Compared to the Word2Vec with Bi-LSTM–Attention model, the attention/LSTM model with transfer embedding showed a better performance in classification, especially for obscure labels. This method might be especially important in bilingual reports.

Our study has several limitations. First, we only included reports from a single tertiary center of surgical in-patients. Our model might be inaccurate in a reporting style different from the one that we have incorporated. Thus, if the model used a data set from another reporting style, the model would need to be validated again. However, in this case, more labeled data might be available, and thus the applied method would show better performance in another data set, especially for bilingual text reports. Second, we could not compare the exact same models with the previous models that showed good performance. However, we compared our model with various deep learning models that were used in previous studies, which is sufficient to compare the performance of different model structures.

In summary, our proposed model showed superior performance as compared to other algorithms in the classification of pneumonia from radiologic reports. In bilingual radiologic reports, the proposed method of transferring and Bi-LSTM–Attention model showed significant improvement in performance than did the previous high-performing models. We hope that this method could be used to enrich the research about pneumonia by obtaining exact outcomes from electronic health data.

Acknowledgments

This work was supported by the Bio-Synergy Research Project (no. NRF-2013M3A9C4078138) of the Ministry of Science, Information and Communications Technology, and Future Planning through the National Research Foundation.

Authors' Contributions

HJP and CMC contributed to the conception and design of the study, as well as the data acquisition. HJP, BYS, EBL, and MS contributed to the analysis and interpretation of the data. HJP, BYS, EBL, and MS drafted the manuscript. HJP, CMC, and MS contributed to the critical revision of the paper, and all authors gave final approval for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary method and figures.
[DOCX File , 603 KB - medinform_v9i5e24803_app1.docx]

References

1. Kehl KL, Elmarakeby H, Nishino M, Van Allen EM, Lepisto EM, Hassett MJ, et al. Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. JAMA Oncol 2019 Oct 01;5(10):1421. [doi: 10.1001/jamaoncol.2019.1800]
2. Sherman RE, Anderson SA, Dal PGJ, Gray GW, Gross T, Hunter NL, et al. Real-world evidence - What is it and what can it tell us? N Engl J Med 2016 Dec 08;375(23):2293-2297. [doi: 10.1056/NEJMsb1609216] [Medline: 27959688]
3. Park HJ, Jung DY, Ji W, Choi CM. Detection of bacteremia in surgical in-patients using recurrent neural network based on time series records: development and validation study. J Med Internet Res 2020 Aug 04;22(8):e19512 [FREE Full text] [doi: 10.2196/19512] [Medline: 32669261]
4. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. JMIR Med Inform 2016 Sep 30;4(3):e28 [FREE Full text] [doi: 10.2196/medinform.5909] [Medline: 27694098]
5. Saqib M, Sha Y, Wang MD. Early prediction of sepsis in EMR records using traditional ML techniques and deep learning LSTM networks. Annu Int Conf IEEE Eng Med Biol Soc 2018 Jul 04;2018(8):4038-4041 [FREE Full text] [doi: 10.1109/EMBC.2018.8513254] [Medline: 30441243]
6. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med 2018 Apr 14;46(4):547-553 [FREE Full text] [doi: 10.1097/CCM.0000000000002936] [Medline: 29286945]
7. Huh K, Hong J, Jung J. Association of meteorological factors and atmospheric particulate matter with the incidence of pneumonia: an ecological study. Clin Microbiol Infect 2020 Dec;26(12):1676-1683. [doi: 10.1016/j.cmi.2020.03.006] [Medline: 32184173]
8. Liu WC, Lin CS, Yeh CC, Wu HY, Lee YJ, Chung CL, et al. Effect of influenza vaccination against postoperative pneumonia and mortality for geriatric patients receiving major surgery: a nationwide matched study. J Infect Dis 2018 Feb 14;217(5):816-826. [doi: 10.1093/infdis/jix616] [Medline: 29216345]
9. Mirończuk M. Information extraction system for transforming unstructured text data in fire reports into structured forms: a Polish case study. Fire Technol 2019 Jul 26;56(2):545-581. [doi: 10.1007/s10694-019-00891-z]

10. 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]
11. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med 2016 Sep 29;375(13):1216-1219 [FREE Full text] [doi: 10.1056/NEJMp1606181] [Medline: 27682033]
12. 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]
13. Zhou S, Li X. Feature engineering vs. deep learning for paper section identification: toward applications in Chinese medical literature. Information Processing & Management 2020 May;57(3):102206. [doi: 10.1016/j.ipm.2020.102206]
14. Ong CJ, Orfanoudaki A, Zhang R, Caprasse FPM, Hutch M, Ma L, et al. Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports. PLoS One 2020;15(6):e0234908 [FREE Full text] [doi: 10.1371/journal.pone.0234908] [Medline: 32559211]
15. Shin SY, Lyu Y, Shin Y, Choi HJ, Park J, Kim WS, et al. Lessons learned from development of de-identification system for biomedical research in a Korean tertiary hospital. Healthc Inform Res 2013 Jun;19(2):102-109 [FREE Full text] [doi: 10.4258/hir.2013.19.2.102] [Medline: 23882415]
16. Shin SY, Park YR, Shin Y, Choi HJ, Park J, Lyu Y, et al. A de-identification method for bilingual clinical texts of various note types. J Korean Med Sci 2015 Jan;30(1):7-15 [FREE Full text] [doi: 10.3346/jkms.2015.30.1.7] [Medline: 25552878]
17. Guillaume L, Alexis C, Ludovic D, Marc'Aurelio R. Unsupervised machine translation using monolingual corpora only. 2018 Presented at: Sixth International Conference on Learning Representations; Apr 30-May 3 2018; Vancouver URL: https://arxiv.org/abs/1711.00043
18. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
19. Dzmitry B, Kyunghyun C, Yoshua B. Neural machine translation by jointly learning to align and translate. 2015 Presented at: 3rd International Conference on Learning Representations; 2015 May 7-9; San Diego.
20. Lee SM, Seo JM, Yun J, Cho YH, Vogel-Claussen J, Schiebler ML, et al. Deep learning applications in chest radiography and computed tomography: current state of the art. J Thorac Imaging 2019 Mar;34(2):75-85. [doi: 10.1097/RTI.0000000000000387] [Medline: 30802231]
21. Cox D. The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological) 2018 Dec 05;20(2):215-232. [doi: 10.1111/j.2517-6161.1958.tb00292.x]
22. Corinna C, Vladimir V. Support-vector networks. Machine learning 1995:273-297.
23. Sebastiani F. Machine learning in automated text categorization. ACM Comput. Surv 2002 Mar;34(1):1-47. [doi: 10.1145/505282.505283]
24. Soucy P, Mineau GW. A simple KNN algorithm for text categorization. : IEEE; 2001 Presented at: IEEE International Conference on Data Mining; Nov 29-Dec 2 2001; San Jose.
25. Quinlan J. Induction of decision trees. Mach Learn 1986 Mar;1(1):81-106. [doi: 10.1007/bf00116251]
26. Leo B. Random forests. Machine learning 2001 Jan:5-32.
27. Tomas M, Kai C, Greg C, Jeffrey D. Efficient estimation of word representations in vector space. 2013 Presented at: 2013 International Conference on Learning Representations,; May 2-May 4 2013; Scottsdale, AX.
28. Keiron O, Ryan N. An introduction to convolutional neural networks. 2015 Dec 2. URL: https://arxiv.org/abs/1511.08458 [accessed 2021-04-30]
29. Greff K, Srivastava RK, Koutnik J, Steunebrink BR, Schmidhuber K. LSTM: a search space odyssey. IEEE Trans Neural Netw Learn Syst 2017 Oct;28(10):2222-2232. [doi: 10.1109/TNNLS.2016.2582924] [Medline: 27411231]
30. Ranzani OT, Prina E, Menéndez R, Ceccato A, Cilloniz C, Méndez R, et al. New sepsis definition (sepsis-3) and community-acquired pneumonia mortality. A validation and clinical decision-making study. Am J Respir Crit Care Med 2017 Nov 15;196(10):1287-1297. [doi: 10.1164/rccm.201611-2262OC] [Medline: 28613918]
31. Wang Y, Sohn S, Liu S, Shen F, Wang L, Atkinson EJ, et al. A clinical text classification paradigm using weak supervision and deep representation. BMC Med Inform Decis Mak 2019 Jan 07;19(1):1 [FREE Full text] [doi: 10.1186/s12911-018-0723-6] [Medline: 30616584]
32. Shi X, Hu Y, Zhang Y, Li W, Hao Y, Alelaiwi A, et al. Multiple disease risk assessment with uniform model based on medical clinical notes. IEEE Access 2016;4:7074-7083. [doi: 10.1109/ACCESS.2016.2614541]

Abbreviations

AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
Bi-LSTM: bidirectional long short-term memory model
CNN: convolutional neural network
CT: computed tomography
EHR: electronic health record

ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification
Kor2Eng: Korean to English
LSTM: long short-term memory model
Word2Vec: word-to-vector representation model

Edited by G Eysenbach; submitted 06.10.20; peer-reviewed by S Sheikhalishahi, M Aczon, R Abeysinghe, X Wang; comments to author 28.10.20; revised version received 21.12.20; accepted 04.04.21; published 17.05.21.

Please cite as:
Park H, Song M, Lee EB, Seo BK, Choi CM
An Attention Model With Transfer Embeddings to Classify Pneumonia-Related Bilingual Imaging Reports: Algorithm Development and Validation
JMIR Med Inform 2021;9(5):e24803
URL: https://medinform.jmir.org/2021/5/e24803
doi:10.2196/24803
PMID:33820755

Original Paper

Combining External Medical Knowledge for Improving Obstetric Intelligent Diagnosis: Model Development and Validation

Kunli Zhang[1*], PhD; Linkun Cai[1*], MSc; Yu Song[1], MSc; Tao Liu[1], MSc; Yueshu Zhao[2], MSc

[1]School of Information Engineering, Zhengzhou University, Zhengzhou, China
[2]The Third Affiliated Hospital of Zhengzhou University, Zhengzhou, China
[*]these authors contributed equally

Corresponding Author:
Yu Song, MSc
School of Information Engineering
Zhengzhou University
No 100, Science Avenue
Zhengzhou, 450000
China
Phone: 86 137 0084 2398
Email: ieysong@zzu.edu.cn

Abstract

Background: Data-driven medical health information processing has become a new development trend in obstetrics. Electronic medical records (EMRs) are the basis of evidence-based medicine and an important information source for intelligent diagnosis. To obtain diagnostic results, doctors combine clinical experience and medical knowledge in their diagnosis process. External medical knowledge provides strong support for diagnosis. Therefore, it is worth studying how to make full use of EMRs and medical knowledge in intelligent diagnosis.

Objective: This study aims to improve the performance of intelligent diagnosis in EMRs by combining medical knowledge.

Methods: As an EMR usually contains multiple types of diagnostic results, the intelligent diagnosis can be treated as a multilabel classification task. We propose a novel neural network knowledge-aware hierarchical diagnosis model (KHDM) in which Chinese obstetric EMRs and external medical knowledge can be synchronously and effectively used for intelligent diagnostics. In KHDM, EMRs and external knowledge documents are integrated by the attention mechanism contained in the hierarchical deep learning framework. In this way, we enrich the language model with curated knowledge documents, combining the advantages of both to make a knowledge-aware diagnosis.

Results: We evaluate our model on a real-world Chinese obstetric EMR dataset and showed that KHDM achieves an accuracy of 0.8929, which exceeds that of the most advanced classification benchmark methods. We also verified the model's interpretability advantage.

Conclusions: In this paper, an improved model combining medical knowledge and an attention mechanism is proposed, based on the problem of diversity of diagnostic results in Chinese EMRs. KHDM can effectively integrate domain knowledge to greatly improve the accuracy of diagnosis.

(JMIR Med Inform 2021;9(5):e25304) doi:10.2196/25304

KEYWORDS

intelligent diagnosis; obstetric electronic medical record; medical knowledge; attention mechanism

Introduction

Intelligent diagnosis is a way to provide clinical decision support for doctors by means of artificial intelligence technology. In the clinic, intelligent diagnosis plays an important role and can be applied to a variety of practical situations. Intelligent diagnosis can help doctors diagnose a patient's condition, significantly improving the efficiency and accuracy of the diagnosis, and the results can also become an important basis for future diagnosis. The continuous development of modern diagnosis and treatment technology has made medical information increasingly complex. Doctors obtain a large amount of clinical diagnostic information every day and need to make comprehensive decisions based on a large amount of data representing clinical information [1]. In addition, the occurrence of complications during pregnancy poses a challenge to doctors.

Electronic medical records (EMRs) are the most detailed and direct form of clinical medical activities [2]. With the rapid growth of EMRs, many methods of intelligent diagnosis using EMRs have become available, enabling significant progress in this field. Early intelligent diagnosis works mainly relied on artificially designed feature templates [3,4] or used single traditional machine learning methods, treating intelligent diagnosis as a classification problem. Goldstein et al [5] used the Informatics for Integrating Biology & the Bedside 2008 dataset to train a classifier for each disease category to classify obesity and 15 other complications. Medhekar et al [6] developed a decision support system based on data mining that used a naïve Bayes classifier to model heart disease. Roopa et al [7] used principal component analysis to extract the characteristics of a diabetes dataset and then used a linear regression model to predict whether a patient had diabetes. These methods promoted the application of machine learning and natural language processing in intelligent diagnosis but are still in the early stages (eg, using relatively simple classification methods and a shallow analysis of the EMRs).

Recently, an increasing number of researchers have focused on neural networks to model intelligent diagnosis and related tasks. Yang et al [8] proposed a clinical assistant diagnosis method based on a multilayer convolutional neural network [9]. This method uses self-learning to automatically extract the high-level semantic information from EMRs. Chen et al [10] used an end-to-end hierarchical neural network to investigate breast cancer problems using EMRs. Hao et al [11] used a deep belief network [12] to integrate patients' structured data characteristics to predict the risk of cerebral infarction. Hao et al [13] proposed a diagnostic modeling and reasoning system using the dynamic uncertain causality graph and improved the diagnostic accuracy of jaundice. Jeddi et al [14] applied the C5.0 algorithm to draw a multibranch decision tree used to aid in the diagnosis of complicated skin diseases.

When the scale of the training data is limited in a traditional neural network, the advantage of using external knowledge is more obvious. These methods ignore the fact that neural networks and external knowledge can benefit from each other.

The rapid development of computer technology and biotechnology has enabled the rapid growth of biomedical text resources. These resources contain valuable knowledge that can be used to promote the development of medical informatics. A doctor's diagnostic process is a combination of their own clinical experience and general medical knowledge. Therefore, medical knowledge is indispensable in the diagnosis process. Fang et al [15] proposed a method to diagnose chronic obstructive pulmonary disease based on a knowledge graph and integrated models. Liang et al [1] designed a system framework for the data mining of EMRs based on pediatric diseases. This framework combines medical knowledge with a data-driven model and uses logistic regression for the disease hierarchical diagnosis. These efforts provide new methods for medical data analysis, but intelligent diagnosis based on EMRs is still hindered by the following problems:

- An EMR usually involves multiple diagnostic results, such as normal diagnosis, pathological diagnosis, and complications.
- In the aspect of external knowledge, the above methods simply splice the knowledge with the model, which fails to capture the key information well and requires a large number of calculations.
- To achieve the most advanced performance, doctors not only care about the diagnostic results but also need to know what medical knowledge contributed to the diagnosis.

Therefore, in this paper, we design a novel intelligent diagnosis model based on deep learning. Specifically, to capture the important details of the original documents, we use bidirectional gated recurrent units (Bi-GRUs) [16] with a hierarchical attention mechanism to model the correlations among words and sentences in EMRs and knowledge documents. Given an analysis of the correlation between the EMRs and medical knowledge documents, we select the most supportive external knowledge to support intelligent diagnosis. Considering the diversity of diagnostic results, we need to conduct intelligent diagnosis in the multilabel classification paradigm. The major contributions of this paper are summarized as follows:

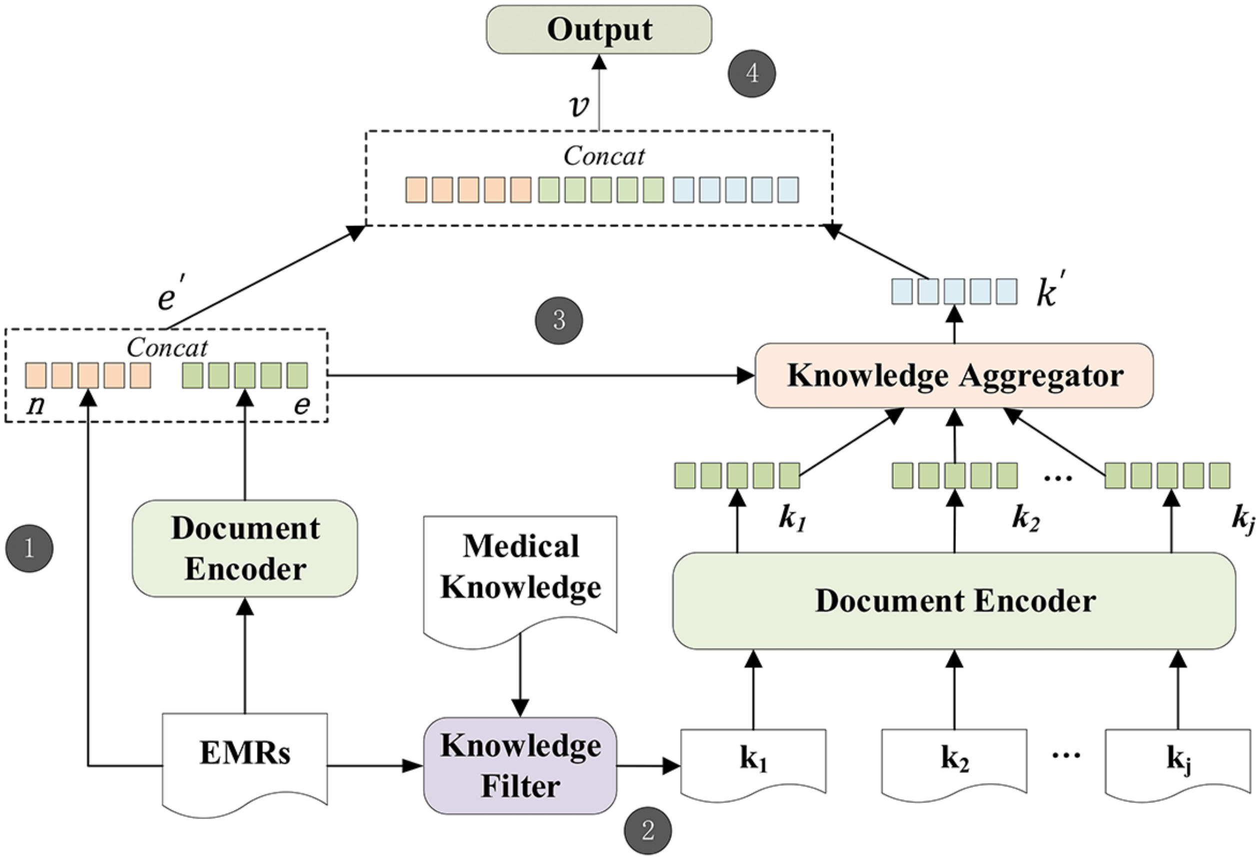
- Knowledge-aware hierarchical diagnosis model (KHDM) makes full use of the hierarchical deep language model to encode the EMRs and external knowledge documents.
- Language model is enriched with high-quality knowledge, combining the advantages of both to perform a knowledge-aware diagnosis.
- Experimental results on real-word Chinese obstetric EMRs achieve superior performance over baselines. In addition, we discuss the importance and interpretability of external medical knowledge.

Methods

Overview

KHDM contains the following steps, as depicted in Figure 1.

Figure 1. Overview of knowledge-aware hierarchical diagnosis model.



1. Enter the EMR into the document encoder to obtain the document embedding *e* and concatenate it with the numerical features *n* to get the final EMR embedding *e′*.
2. Input the EMRs and external knowledge documents into the knowledge filter for preliminary screening of the external knowledge, and send the filtered knowledge documents to the document encoder to obtain the knowledge embedding *k*.
3. Input the EMR embedding and knowledge embedding jointly into the knowledge aggregator. Through the simultaneous analysis of the EMRs and knowledge documents, our model learns a knowledge-side attention component in order to carefully select the most supportive knowledge document *k′* from the external knowledge to support intelligent diagnosis.
4. *e′* and *k′* are concatenated and passed to a sigmoid classifier for the diagnosis. In this section, we introduce the document encoder, knowledge attention module (including the knowledge filter and knowledge aggregator), and output.

Document Encoder

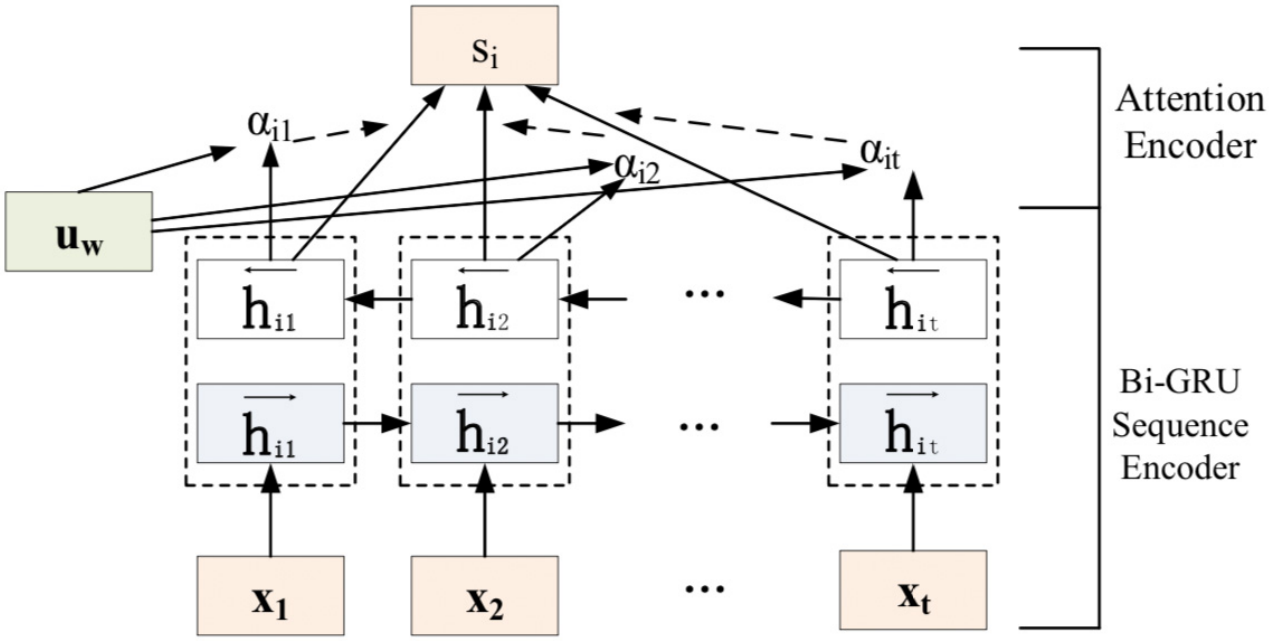
The purpose of the document encoder is to encode the original EMRs and knowledge documents into continuous low-dimensional embeddings to capture semantic relationships. EMRs and medical knowledge documents usually have potential hierarchical structures. A document consists of several sentences, and a sentence consists of several words. Intuitively, the document embedding problem can be converted into two sequence embedding problems [17]. Modeling the semantics of the EMR and external knowledge by word-level and sentence-level representations can fully capture the hierarchical laws and dependencies.

The words and sentences in a document provide different information and have different degrees of importance. Inspired by Yang et al [18], we successively apply the attention mechanism [19] at the word level and sentence level so that it can differentiate more important information when constructing the document representation. The attention mechanism not only improves the performance of the deep learning model but also intuitively shows the contributions of words and sentences to the classification decision.

We use the Bi-GRU sequence encoder with an attention mechanism to encode the EMRs and knowledge documents. Numerical features, such as physiological indicators and laboratory results, are also important in EMRs. To enable more complete use of the EMRs, we separately extract the numerical features and concatenate them with EMRs. Next, we introduce the Bi-GRU sequence encoder, attention encoder, and numerical features in detail.

Although the word-level and sentence-level encoders can have different structures, we use the same structure here for simplicity, as shown in Figure 2.

Figure 2. Document encoder framework.



Bi-GRU Sequence Encoder

The importance of words and sentences is highly context dependent. In other words, the same words or sentences may have different degrees of importance in different contexts. We model the semantics of EMRs and external knowledge documents by including word-level and sentence-level representations that can fully capture hierarchical dependencies. Taking the word level as an example, we use Bi-GRU to make a word compilation of the meaning of an entire sentence, where the GRU uses a gate control mechanism to memorize the information of the previous cells.

The GRU has two gates: the reset gate r_t and the update gate z_t. The reset gate is used to determine the degree to which the previous information is forgotten, and the update gate is used to decide which information to forget and which new information to enter. r_t and z_t jointly control the calculation from hidden state $h_{t–1}$ to hidden state h_t. $h_t^{\sim}$ is a candidate hidden layer. At time t, the GRU is calculated as follows:

where W* is the weight matrix. x_t is the sequence vector at time t, and σ is the activation sigmoid function that converts the values of each cell state into the range of 0 to 1 to act as a gate signal. The reset gate r_t receives the values of $h_{t–1}$ and x_t. If r_t is zero, then the previous state is not saved. In other words, at this time, $h_t^{\sim}$ only contains the information of the current word. Afterward, the update gate z_t controls how much information needs to be forgotten from the hidden state $h_{t–1}$ at the previous moment and how much hidden layer information $h_t^{\sim}$ needs to be added at the moment. The final hidden layer information h_t can then be output.

Bi-GRU uses forward and backward GRUs to encode the sequence in two directions so that the associations between different words (sentences) are taken into account when encoding. Specifically, consider an EMR e = $[s_1,s_2,^{\cdots},s_L]$, where L is the number of sentences and $s_i(1 \leq I \leq L)$ represents the i^{th} sentence in the document. For each sentence in the document $s_i = [w_{i1},w_{i2},^{\cdots},w_{iT}]$, $w_{im}(1 \leq m \leq T)$ represents the m^{th} word in s_i. w_{im} is the embedding representation of w_{im}, and the encoding method is to concatenate the feature representations of Bi-GRU; that is, the forward hidden state $h_{it}^{\rightarrow}$ and backward hidden state $h_{it}^{\leftarrow}$ at time t are weighted sums:

Attention Encoder

Not all words have the same effect on the meaning of a sentence, as is the case for sentences within documents. The attention mechanism has become an effective mechanism for mining local differences and highlighting vital elements of data. Therefore, we add an attention mechanism at the word and sentence levels to indicate their importance to the previous level. Compared with the general word-level attention mechanism, the sentence-level attention mechanism plays a more important role in medical documents because certain domain phrases often appear. At the word level, the attention mechanism is introduced to extract those words that are important to the meaning of the sentence, and the representations of these informative words are aggregated to form a sentence vector. The final sentence vector representation s_i is defined as follows:

where the weight a_{it} indicates the importance of a word to the meaning of the sentence. The context vector u_w is an attention matrix obtained by a random initialization method. It is a cumulative sum of the different probability weights assigned by the attention mechanism and the performance of each hidden layer state. We measure the importance of the word as similarity of w_{it} with a word-level context vector u_w and get a normalized importance weight α_{it} through a softmax function. We use the same method to obtain the context-level representation of u_s and finally to obtain the document vector *e*:

Numerical Features

Numerical features are very important indicators in Chinese obstetric EMRs. For example, physiological indicators such as the age of the pregnant woman, the number of menopause months, and the uterine height are important factors affecting the clinical judgement. However, there are some cases where the numerical units of EMRs are not uniform. Taking the number of menopause months as an example, it is generally described as "menopause X months," but some EMRs also use the description method "menopause Y weeks," We unified the units of this indicator as months, relying on the equation that "4 weeks" is approximately "1 month" in the feature extraction. We also need to consider the validity of the data. According to medical professional knowledge, numerical features have a certain value range. For example, when extracting the physiological parameters of a pregnant woman's uterine height, if a value is found to be "29 m," it can be speculated that this data point is incorrect, which will affect the experimental results. This paper determines the accuracy of the data by setting thresholds for each physiological index, and the error data are directly deleted. Detailed thresholds descriptions are provided in Multimedia Appendix 1. After extracting the numerical features *n*, they are concatenated with the document vector *e* as the final representation of the EMR:

Knowledge Attention Module

Integrating all the external knowledge into the model is very time-consuming, and not all knowledge has enough discernment to support the final classification. Our knowledge attention module aims to alleviate these problems, ensuring that our model can select reliable and useful knowledge for each candidate. This module consists of a knowledge filter and knowledge aggregator. The knowledge filter can preliminarily filter out irrelevant knowledge documents, and the knowledge aggregator uses the attention mechanism to select the most supported knowledge. Considering that external knowledge has too much noise, such an attention mechanism explores the correlation between the EMRs and knowledge documents. KHDM mainly uses this module to make a knowledge-aware diagnosis.

Knowledge Filter

We consider the task of the knowledge filter to be text similarity calculation. By calculating the similarity between the input EMRs and the medical knowledge documents, the knowledge not related to the input EMRs will be filtered out. Due to the special nature of medical texts, symptoms and diagnostic methods vary by disease. Therefore, we use the term frequency–inverse document frequency (TF-IDF) to extract the text features of the EMRs and external knowledge. TF(x) represents word frequency, which counts the frequency of each word in an EMR. IDF(x) represents the inverse text frequency and returns the frequency of word x in the corpus, reflecting the importance of words in the text:

where N(x) represents the number of occurrences of word x in the document, N is the total number of words in the document, and D is the total number of documents. D(x) indicates how many documents the word x appears in. Due to professionalism in the medical field, the IDF is smoothed so that domain words that do not appear in all documents can also obtain a suitable IDF value:

The set of documents and knowledge is then viewed as a set of vectors in a vector space. The cosine function is used to measure the similarity between the document and any knowledge. If the similarity score is less than 0.5, we consider these knowledge documents irrelevant and vice versa. After that, we use the document encoder mentioned above to encode the relevant knowledge document. Finally, we obtain the relevant knowledge vector representation: $k = [k_1,k_2,\cdots,k_j]$.

Knowledge Aggregator

This submodule aims to find further medical knowledge that supports intelligent diagnosis and generates an aggregated knowledge embedding *k*′. Therefore, we use the attention mechanism to select the key knowledge documents that are the most critical to the task objective. When generating an aggregated knowledge embedding, more attention is paid to the most important knowledge:

The attention weight α_t generated by k_t and *e*′ can be regarded as the correlation between the external knowledge and the input EMRs. The top *k*-related knowledge is selected according to the attention weight after sorting. The number of related knowledge documents less than *k* will be padded with zero vectors. We define *k* as the average label number per document.

Output

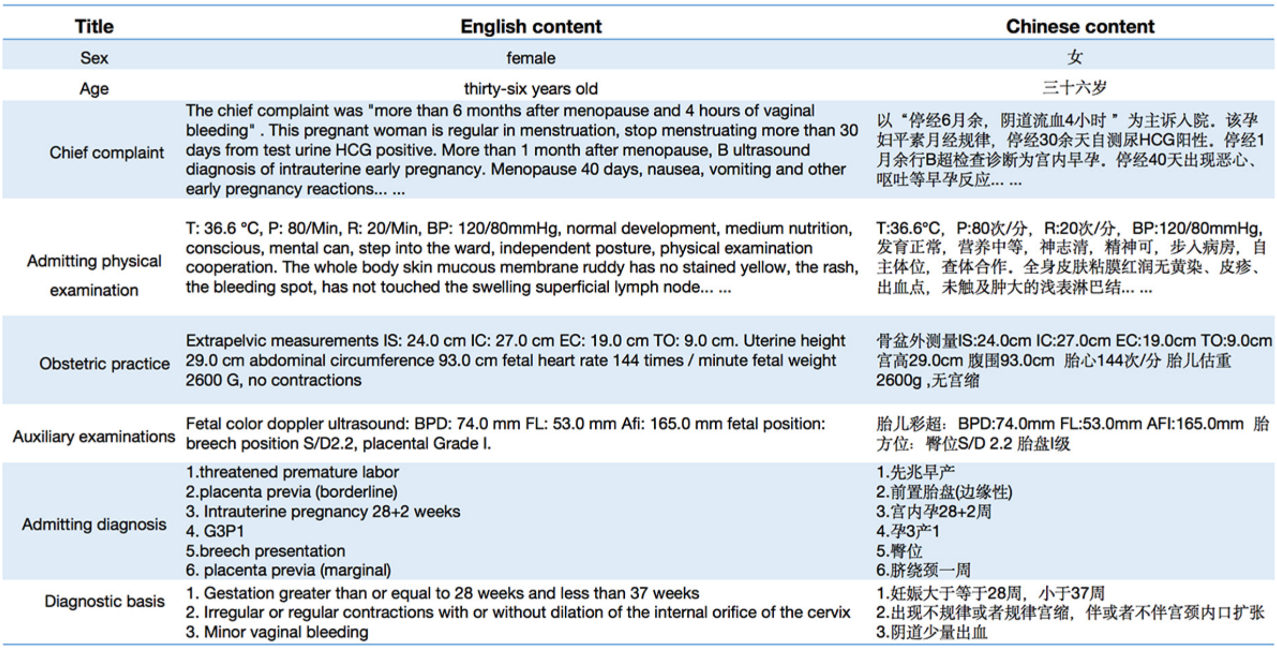
To make the final diagnosis prediction, we first concatenate the EMR embedding *e*′ and the knowledge embedding *k*′ and feed it into two fully connected layers to generate a new vector, which is then passed to a sigmoid classifier to produce the predicted results. We consider that all diseases with an output probability greater than τ are positive predictions. The input to the first fully connected layer can also be only *e*′ or *k*′, which means we use only EMRs or external knowledge to make the diagnosis. The loss function for the training is the cross entropy:

Results

Dataset Details

We collected 24,192 Chinese obstetric EMRs randomly selected by multiple hospitals as the research material, and each EMR corresponds to one patient. Due to the different writing habits of doctors, there are many different forms of expression for the same diagnostic results. Therefore, the medical thesaurus *International Classification of Diseases, Tenth Revision* [20] is used as the basis for the standardization of disease naming. To protect the privacy of patients, personal identifying information such names and ID numbers of patients was removed [21]. The dataset focuses on inpatient department data and consists primarily of structured and unstructured text data. Structured data include the basic information on the patient such as age, ethnicity, and laboratory examination data. Unstructured data mainly refer to the patient's main complaint, admission, and physical examination. Detailed data descriptions are shown in Figure 3. The dataset contains 59 types of disease diagnostic results and is divided into 21,772 training sets and 2420 test sets according to the results distribution.

Figure 3. Chinese obstetric electronic medical record sample.



Title	English content	Chinese content
Sex	female	女
Age	thirty-six years old	三十六岁
Chief complaint	The chief complaint was "more than 6 months after menopause and 4 hours of vaginal bleeding" . This pregnant woman is regular in menstruation, stop menstruating more than 30 days from test urine HCG positive. More than 1 month after menopause, B ultrasound diagnosis of intrauterine early pregnancy. Menopause 40 days, nausea, vomiting and other early pregnancy reactions... ...	以“停经6月余，阴道流血4小时”为主诉入院。该孕妇平素月经规律，停经30余天自测尿HCG阳性。停经1月余行B超检查诊断为宫内早孕。停经40天出现恶心、呕吐等早孕反应... ...
Admitting physical examination	T: 36.6 °C, P: 80/Min, R: 20/Min, BP: 120/80mmHg, normal development, medium nutrition, conscious, mental can, step into the ward, independent posture, physical examination cooperation. The whole body skin mucous membrane ruddy has no stained yellow, the rash, the bleeding spot, has not touched the swelling superficial lymph node... ...	T:36.6°C，P:80次/分，R:20次/分，BP:120/80mmHg，发育正常，营养中等，神志清，精神可，步入病房，自主体位，查体合作。全身皮肤粘膜红润无黄染、皮疹、出血点，未触及肿大的浅表淋巴结... ...
Obstetric practice	Extrapelvic measurements IS: 24.0 cm IC: 27.0 cm EC: 19.0 cm TO: 9.0 cm. Uterine height 29.0 cm abdominal circumference 93.0 cm fetal heart rate 144 times / minute fetal weight 2600 G, no contractions	骨盆外测量IS:24.0cm IC:27.0cm EC:19.0cm TO:9.0cm 宫高29.0cm 腹围93.0cm 胎心144次/分 胎儿估重2600g ,无宫缩
Auxiliary examinations	Fetal color doppler ultrasound: BPD: 74.0 mm FL: 53.0 mm Afi: 165.0 mm fetal position: breech position S/D2.2, placental Grade I.	胎儿彩超：BPD:74.0mm FL:53.0mm AFI:165.0mm 胎方位：臀位S/D 2.2 胎盘I级
Admitting diagnosis	1.threatened premature labor 2.placenta previa (borderline) 3. Intrauterine pregnancy 28+2 weeks 4. G3P1 5.breech presentation 6. placenta previa (marginal)	1.先兆早产 2.前置胎盘(边缘性) 3.宫内孕28+2周 4.孕3产1 5.臀位 6.脐绕颈一周
Diagnostic basis	1. Gestation greater than or equal to 28 weeks and less than 37 weeks 2. Irregular or regular contractions with or without dilation of the internal orifice of the cervix 3. Minor vaginal bleeding	1.妊娠大于等于28周，小于37周 2.出现不规律或者规律宫缩，伴或者不伴宫颈内口扩张 3.阴道少量出血

For external knowledge, we collected descriptions of medical concepts from the authoritative textbook *Obstetrics and Gynecology* [22] and a medical encyclopedia. The medical concepts mainly include the disease definition, symptoms, and treatment methods. In the end, we collect a total of 72 medical definition documents that make up our external knowledge. All external knowledge was chosen under the guidance of medical experts.

Hyperparameter Setting

Since all EMRs and external knowledge documents are written in Chinese, we first use PKUSEG [23] to segment the document and set the maximum document length to 1600 characters. We use the GloVe [24] model to train word embedding on the corpus of EMRs after word segmentation. The hidden state size of the GRU is set to 100. For text convolutional neural network (TextCNN), this paper sets the filter width to (2, 3, 4, 5), and each filter size is 25 to maintain consistency. After the connection, the representation size of our model becomes 200. Finally, a 200 * c fully connected layer is added (c is the number of labels).

Since we use the sigmoid function for classification, the prediction threshold τ is set to 0.5. Average label number per document k is 2.688, so we set $k = 3$. We use Adam [25] as the optimizer. During the training period, EMRs are selected by random sampling method. We set the learning rate to 0.001 and the batch size to 32.

Performance on an Obstetric EMR Dataset

In multilabel learning, each sample may have multiple category labels. Many evaluation metrics for multilabel learning have been proposed [26]. We use the average precision, 1-error, hamming loss, ranking loss, and coverage as evaluation metrics. The following text classification models were used as baselines for comparison:

- Classifier chains [27] integrate multiple single classification methods into one model to solve the problem of multilabel classification.
- Multilabel k–nearest neighbor [28] considers the k instances with the smallest distance from the new instance in the feature space as a set.
- Long short-term memory (LSTM) [29] uses the last hidden state as the representation of the whole document.
- Bidirectional long short-term memory (Bi-LSTM) is a bidirectional LSTM that can obtain long-term context information in the direction of the input.
- TextCNN [9] uses multiple kernels of different sizes to extract the key information in sentences to better capture the local relevance.

All text classification models are trained in the multilabel framework. The experimental results on the Chinese obstetric EMR dataset are summarized in Table 1.

Table 1. Comparative results on Chinese obstetric electronic medical record dataset.

Method	Average precision	1-error	Hamming loss	Ranking loss	Coverage
CC[a]	0.5083	0.4880	0.0308	0.1366	19.7917
ML-KNN[b]	0.6109	0.2488	0.0258	0.0709	10.2347
LSTM[c]	0.8651	0.0836	0.0166	0.0190	4.4612
Bi-LSTM[d]	0.8721	0.0775	0.0164	0.0186	4.4625
TextCNN[e]	0.8652	0.0961	0.0188	0.0203	4.6035
KHDM[f]	0.8929	0.0713	0.0156	0.0165	4.0833

[a] CC: classifier chains.

[b] ML-KNN: multilabel k–nearest neighbor.

[c] LSTM: long short-term memory.

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

[e] TextCNN: text convolutional neural networks.

[f] KHDM: knowledge-aware hierarchical diagnosis model.

According to the experimental results, compared with the traditional machine learning methods, the neural network method has achieved better results. The main reason is that the neural network can capture richer features and deeper semantic information. Considering the structured context information, a bidirectional network can significantly improve the performance. For example, Bi-LSTM gives an average precision of 0.8721, while that of the LSTM is 0.8651. In addition, our model is largely superior to other traditional neural network methods. The TextCNN is usually connected to the pooling layer after the convolution layer. Its operation logic is to retain the strongest features from the feature vectors obtained from a convolution kernel so it cannot retain the relative position information of the original input, resulting in information loss. LSTM has a sequence dependency problem and does not perform well when the document is too long. Our model uses a hierarchical structure to divide the document into sentences without the problems of distance dependence and information loss. In general, our model is much better than the other models in all of the evaluation metrics applied, with improvements of 3% to 30%. Making full use of the attention mechanism to integrate external medical knowledge is undoubtedly an important way to improve the effectiveness of intelligent diagnosis.

Performance on Public Dataset

This paper takes the obstetric intelligent diagnosis problem into a multilabel classification framework. Therefore, we test the classification effect on two public datasets: DeliciousMIL [30] and Hep categories. The former consists of a number of tagged pages on the social bookmarking site delicious.com, with categories including programming, style, and reference, and the latter is a public multilabel dataset available on Magpie, with subject categories relevant to high-energy physics (HEP) abstracts, including astrophysics, experiment-HEP, gravitation and cosmology, phenomenology-HEP, and theory-HEP. Table 2 provides a brief description of each dataset. The selected external knowledge *k* values of the two datasets are 3 and 1, respectively.

The external knowledge data for the DeliciousMIL and Hep categories datasets are derived from Wikipedia entry definitions. Table 3 and Table 4 present the results. Similar to the results on the obstetric EMR dataset, it can be clearly observed that our model performs best in multilabel text classification, proving that KHDM is universal for text classification tasks.

Table 2. Description of public datasets.

Dataset	Field	Instances	Labels	AL[a]
DeliciousMIL	Social networking sites	12,234	20	2.9574
Hep categories	High-energy physics	1000	5	1.1920

[a] AL: average label number per document.

Table 3. Comparative results on public dataset DeliciousMIL.

Method	Average precision	1-error	Hamming loss	Ranking loss	Coverage
CC[a]	0.3208	0.8134	0.2054	0.4183	12.9241
ML-KNN[b]	0.3703	0.7621	0.4748	0.3488	11.0213
LSTM[c]	0.5813	0.3947	0.1641	0.1518	6.9928
Bi-LSTM[d]	0.5968	0.3786	0.1610	0.1615	6.9648
TextCNN[e]	0.6299	0.3639	0.1760	0.1344	6.0637
KHDM[f]	0.6386	0.3312	0.1255	0.1284	5.9101

[a] CC: classifier chains.

[b] ML-KNN: multilabel k–nearest neighbor.

[c] LSTM: long short-term memory.

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

[e] TextCNN: text convolutional neural networks.

[f] KHDM: knowledge-aware hierarchical diagnosis model.

Table 4. Comparative results on public dataset Hep categories.

Method	Average precision	1-error	Hamming loss	Ranking loss	Coverage
CC[a]	0.5606	0.6290	0.2982	0.4381	1.9410
ML-KNN[b]	0.5733	0.5800	0.3460	0.4433	2.2300
LSTM[c]	0.6807	0.5422	0.2740	0.2437	0.9642
Bi-LSTM[d]	0.7055	0.4816	0.2200	0.2251	0.9455
TextCNN[e]	0.7903	0.3429	0.2420	0.1550	0.6207
KHDM[f]	0.8929	0.0713	0.0156	0.0165	4.0833

[a] CC: classifier chains.

[b] ML-KNN: multilabel k–nearest neighbor.

[c] LSTM: long short-term memory.

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

[e] TextCNN: text convolutional neural networks.

[f] KHDM: knowledge-aware hierarchical diagnosis model.

Discussion

Ablation Test

KHDM is a combination of a knowledge attention mechanism and external medical knowledge representation. We conducted an ablation test to assess the contributions of these two components in our model. Table 5 presents the performance of our model and its ablations on the obstetric EMR dataset. *w/o Knowledge* means using only the EMRs for the intelligent diagnosis, and *w/o Att* means we remove the attention mechanism and all the medical knowledge documents directly concatenated with the EMRs and do not use the knowledge attention module.

Table 5. Results of the ablation test.

Method	Average precision	One error	Hamming loss	Ranking loss	Coverage
w/o Knowledge	0.8789	0.1047	0.0184	0.0212	4.2364
w/o Att[a]	0.8519	0.1022	0.0164	0.0181	4.3210
TextCNN[b]	0.8652	0.0961	0.0188	0.0203	4.6035
TextCNN + knowledge	0.8700	0.0912	0.0167	0.0199	4.3516
KHDM[c]	0.8929	0.0713	0.0156	0.0165	4.0833

[a] Att: attention.

[b] TextCNN: text convolutional neural networks.

[c] KHDM: knowledge-aware hierarchical diagnosis model.

From the experimental results, the following can be seen:

- When the external knowledge is not introduced or the attention mechanism is not used, the model performance deteriorates.
- The models incorporating knowledge are superior to ordinary text classification models with a drop to 0.8789 of model *w/o Knowledge* after the supplementary knowledge is removed. The effectiveness of using external knowledge information is confirmed, and medical knowledge contributes to intelligent diagnosis.
- When fusing the medical knowledge, performances of .*w/o Att* and *TextCNN + knowledge* significantly increase by simply concatenating the knowledge document.

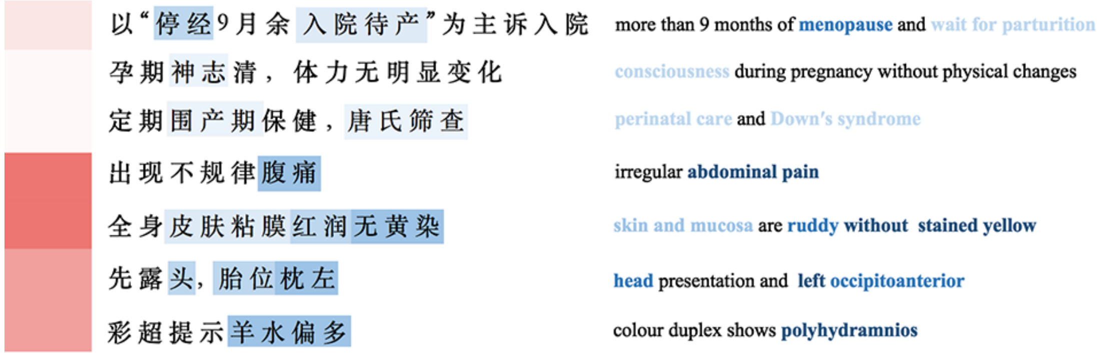
However, these models do not use the knowledge attention mechanism but directly concatenate with the external knowledge, which will introduce a large amount of noise. We can see KHDM improves more than 2 percentage points on most evaluation metrics. These ablation test results reflect the importance and rationality of using the attention mechanism to capture the interactions between multiple inputs.

Interpretability of the Attention Mechanism

Interpretability is very important for model evaluation, especially in the medical field, as it allows doctors to understand the rationale behind the diagnostic results. To verify that our model can capture the most important sentences and words in a document, we first visualized the hierarchical attention mechanism in the document encoder on the Chinese obstetric EMR dataset.

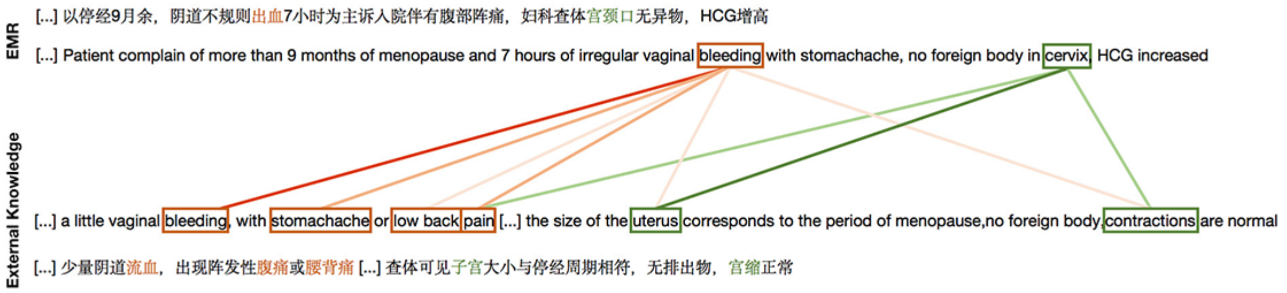
As shown in Figure 4, every line is a sentence, and we normalize the sentence weights and word weights to ensure that only the important words in the most important sentences are emphasized. Red denotes the weight of a sentence and blue denotes the weight of a word, where the darker the color is, the greater the weight. We know that doctors often diagnose patients by analyzing their clinical symptoms and test results. Our model accurately locates the words *abdominal pain* and *no yellow stain* and their corresponding sentences.

Figure 4. Visualization of attention in document encoder (attention encoder).


以“停经9月余入院待产”为主诉入院
more than 9 months of menopause and wait for parturition
孕期神志清，体力无明显变化
consciousness during pregnancy without physical changes
定期围产期保健，唐氏筛查
perinatal care and Down's syndrome
出现不规律腹痛
irregular abdominal pain
全身皮肤粘膜红润无黄染
skin and mucosa are ruddy without stained yellow
先露头，胎位枕左
head presentation and left occipitoanterior
彩超提示羊水偏多
colour duplex shows polyhydramnios

Next, we choose a representative example to illustrate the role of the attention mechanism in the knowledge aggregator. We remove all attention values less than 10^{-3} from the visualization. As can be seen in Figure 5, our model pays more attention to the clinical symptom *blood (red part)* and site *cervix (green part)* within the medical knowledge. The darker the color of the line, the higher the attention. Similarly, medical concepts are essential in clinical diagnosis, so medical knowledge with a higher attention score through localization of symptoms and sites will be selected.

Figure 5. Visualization of attention in knowledge aggregator (knowledge attention).


EMR
[...] 以停经9月余，阴道不规则出血7小时为主诉入院伴有腹部阵痛，妇科查体宫颈口无异物，HCG增高
[...] Patient complain of more than 9 months of menopause and 7 hours of irregular vaginal bleeding with stomachache, no foreign body in cervix, HCG increased
External Knowledge
[...] a little vaginal bleeding, with stomachache or low back pain [...] the size of the uterus corresponds to the period of menopause,no foreign body,contractions are normal
[...] 少量阴道流血，出现阵发性腹痛或腰背痛 [...] 查体可见子宫大小与停经周期相符，无排出物，宫缩正常

Limitations

We used only external medical knowledge related to obstetric diseases, but obstetric diagnosis also involves immunology, cytology, genetics, pathology, and other multilevel knowledge. For cardiovascular and cerebrovascular diseases requiring blood pressure and routine blood tests, the numerical features are very important for the diagnosis, and our proposed method provide support. These numerical features are very important for the diagnosis, and our proposed method can provide support. But for diseases such as cancer, text data alone is not enough and must be combined with other types of medical information such as medical images and signals. To improve the interpretability of intelligent diagnosis model, communication with the clinic and selection of an appropriate interpretation method in terms of complementing the doctor's workflow and habits is still necessary. Another limitation that needs to be addressed in achieving intelligent diagnosis based on EMRs is imbalanced datasets. This paper selects common diseases as the research object. In future work, we will focus on diseases with lower frequency.

Conclusions

In this paper, we propose KHDM that synchronously and effectively uses Chinese obstetric EMRs and external knowledge. Particularly, the use of the knowledge attention module to selectively leverage medical knowledge not only improves performance but also provides a basis for intelligent diagnosis. The experimental results on a real obstetric EMR dataset show that KHDM can effectively use external knowledge to enhance the language model, thereby improving the performance.

Acknowledgments

This work is supported by grant 18ZDA315 from the Major Program of the National Social Science Foundation of China, grant 2019TQ0286v from the China Postdoctoral Science Foundation, grant 192102210260 from the Science and Technique Program of Henan Province, grant SB201901021from the Medical Science and Technique Program Cosponsored by Henan Province and Ministry, and grants 19A520003 and 20A520038 from the Key Scientific Research Program of Higher Education of Henan Province. The funding body did not play any role in the design of the study; collection, analysis, and interpretation of data; or in writing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Threshold descriptions in numerical features.
[DOCX File , 14 KB - medinform_v9i5e25304_app1.docx]

References

1. Liang H, Tsui BY, Ni H, Valentim CCS, Baxter SL, Liu G, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med 2019 Mar;25(3):433-438. [doi: 10.1038/s41591-018-0335-9] [Medline: 30742121]
2. Hornberger J. Electronic health records: a guide for clinicians and administrators. JAMA 2009 Jan 07;301(1):110. [doi: 10.1001/jama.2008.910]
3. Turchin A, Shubina M, Breydo E, Pendergrass ML, Einbinder JS. Comparison of information content of structured and narrative text data sources on the example of medication intensification. J Am Med Inform Assoc 2009;16(3):362-370 [FREE Full text] [doi: 10.1197/jamia.M2777] [Medline: 19261947]
4. Piao M, Lee HG, Pok C. A data mining approach for dyslipidemia disease prediction using carotid arterial feature vectors. 2010 Presented at: International Conference on Computer Engineering and Technology; 2010; Chengdu p. 16-18. [doi: 10.1109/iccet.2010.5485249]
5. Goldstein I, Uzuner O. Specializing for predicting obesity and its co-morbidities. J Biomed Inform 2009 Oct;42(5):873-886 [FREE Full text] [doi: 10.1016/j.jbi.2008.11.001] [Medline: 19041423]

6. Pattekari SA, Parveen A. Prediction system for heart disease using Naïve Bayes. Int J Adv Comput Math Sci 2012;3(3):290-294. [doi: 10.1109/CIMCA.2016.8053261]
7. Roopa H, Asha T. A linear model based on principal component analysis for disease prediction. IEEE Access 2019;7:105314-105318. [doi: 10.1109/access.2019.2931956]
8. Yang Z, Huang Y, Jiang Y, Sun Y, Zhang Y, Luo P. Clinical assistant diagnosis for electronic medical record based on convolutional neural network. Sci Rep 2018 Apr 20;8(1):6329 [FREE Full text] [doi: 10.1038/s41598-018-24389-w] [Medline: 29679019]
9. Kim Y. Convolutional neural networks for sentence classification. 2014 Presented at: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October 25-29; Doha, Qatar p. 1746-1751.
10. Chen D, Qian G, Pan Q. Breast cancer classification with electronic medical records using hierarchical attention bidirectional networks. 2018 Presented at: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 3-6; Madrid, Spain p. 983-988. [doi: 10.1109/bibm.2018.8621479]
11. Hao Y, Usama M, Yang J, Hossain MS, Ghoneim A. Recurrent convolutional neural network based multimodal disease risk prediction. Future Generation Computer Systems 2019 Mar;92:76-83. [doi: 10.1016/j.future.2018.09.031]
12. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science 2006 Jul 28;313(5786):504-507 [FREE Full text] [doi: 10.1126/science.1127647] [Medline: 16873662]
13. Hao S, Geng S, Fan L, Chen J, Zhang Q, Li L. Intelligent diagnosis of jaundice with dynamic uncertain causality graph model. J Zhejiang Univ Sci B 2017 May;18(5):393-401 [FREE Full text] [doi: 10.1631/jzus.B1600273] [Medline: 28471111]
14. Jeddi FR, Arabfard M, Kermany ZA. Intelligent diagnostic assistant for complicated skin diseases through C5's algorithm. Acta Inform Med 2017 Sep;25(3):182-186 [FREE Full text] [doi: 10.5455/aim.2017.25.182-186] [Medline: 29114111]
15. Fang Y, Wang H, Wang L, Di R, Song Y. Diagnosis of COPD based on a knowledge graph and integrated model. IEEE Access 2019;7:46004-46013. [doi: 10.1109/access.2019.2909069]
16. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. 2015 Presented at: 3rd International Conference on Learning Representations; May 7-9; San Diego, California URL: http://arxiv.org/abs/1409.0473
17. Tang D, Qin B, Liu T. Document modeling with gated recurrent neural network for sentiment classification. 2015 Presented at: 2015 conference on empirical methods in natural language processing (EMNLP); September 17-21; Lisbon, Portugal p. 1422-1432 URL: https://www.aclweb.org/anthology/D15-1167.pdf [doi: 10.18653/v1/d15-1167]
18. Yang Z, Yang D, Dyer C. Hierarchical attention networks for document classification. 2016 Presented at: 15th conference of the North American chapter of the association for computational linguistics: human language technologies; June 12-17; San Diego, California p. 1480-1489 URL: https://www.aclweb.org/anthology/N16-1174.pdf [doi: 10.18653/v1/n16-1174]
19. Vaswani A, Shazeer N, Parmar N. Attention is all you need. 2017 Presented at: 31th Conference on Neural Information Processing Systems (NIPS); December 4-9; Long Beach, California p. 5998-6008 URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf [doi: 10.5040/9781350101272.00000005]
20. Sundararajan V, Henderson T, Perry C, Muggivan A, Quan H, Ghali WA. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. J Clin Epidemiol 2004 Dec;57(12):1288-1294. [doi: 10.1016/j.jclinepi.2004.03.012] [Medline: 15617955]
21. Zhao Y, Zhang K, Ma H, Li K. Leveraging text skeleton for de-identification of electronic medical records. BMC Med Inform Decis Mak 2018 Mar 22;18(Suppl 1):18 [FREE Full text] [doi: 10.1186/s12911-018-0598-6] [Medline: 29589571]
22. Xie X, Kong BT, Duan T. Obstetrics and Gynecology. 9th edition. Beijing: People's Medical Publishing House; 2018.
23. Luo R, Xu J, Zhang Y. PKUSEG: a toolkit for multi-domain Chinese word segmentation. ArXiv.. Preprint posted on Jun 27, 2019. URL: http://arxiv.org/abs/1906.11455 [accessed 2019-06-27]
24. Pennington J, Socher R, Manning CD. Glove: global vectors for word representation. 2014 Presented at: 2014 conference on empirical methods in natural language processing (EMNLP); October 25-29; Doha, Qatar p. 1532-1543. [doi: 10.3115/v1/D14-1162]
25. Kingma DP, Ba J. Adam: a method for stochastic optimization. 2015 Presented at: 3rd International Conference on Learning Representations; May 7-9; San Diego, California URL: http://arxiv.org/abs/1412.6980
26. Zhang M, Zhou Z. A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng 2014 Aug;26(8):1819-1837. [doi: 10.1109/TKDE.2013.39]
27. Read J, Pfahringer B, Holmes G. Classifier chains for multi-label classification. 2008 Presented at: Joint European Conference on Machine Learning and Knowledge Discovery in Databases; September 14-18; Ghent, Belgium p. 254-269. [doi: 10.1007/978-3-642-04174-7_17]
28. Zhang M, Zhou Z. ML-KNN: a lazy learning approach to multi-label learning. Pattern Recognition 2007 Jul;40(7):2038-2048. [doi: 10.1016/j.patcog.2006.12.019]
29. Liu P, Qiu X, Huang X. Recurrent neural network for text classification with multi-task learning. 2016 Presented at: Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI); July 9-15; New York p. 2873-2879 URL: https://www.ijcai.org/Proceedings/16/Papers/408.pdf [doi: 10.1016/0004-3702(82)90046-7]

30. Soleimani H, Miller DJ. Semi-supervised multi-label topic models for document classification and sentence labeling. 2016 Presented at: 25th ACM international on conference on information and knowledge management (CIKM); October 24-26; Indianapolis p. 105-114. [doi: 10.1145/2983323.2983752]

Abbreviations

Bi-GRU: bidirectional gated recurrent unit
Bi-LSTM: bidirectional long short-term memory
EMR: electronic medical record
HEP: high-energy physics
KHDM: knowledge-aware hierarchical diagnosis model
LSTM: long short-term memory
TextCNN: text convolutional neural network
TF-IDF: term frequency–inverse document frequency

Edited by C Lovis; submitted 27.10.20; peer-reviewed by L Li, H Liang; comments to author 14.01.21; revised version received 11.03.21; accepted 18.04.21; published 10.05.21.

Please cite as:
Zhang K, Cai L, Song Y, Liu T, Zhao Y
Combining External Medical Knowledge for Improving Obstetric Intelligent Diagnosis: Model Development and Validation
JMIR Med Inform 2021;9(5):e25304
URL: https://medinform.jmir.org/2021/5/e25304
doi:10.2196/25304
PMID:33970113

Original Paper

A Multimodal Imaging–Based Deep Learning Model for Detecting Treatment-Requiring Retinal Vascular Diseases: Model Development and Validation Study

Eugene Yu-Chuan Kang[1,2*], MD; Ling Yeung[2,3*], MD; Yi-Lun Lee[4], BSc; Cheng-Hsiu Wu[2,3], MD; Shu-Yen Peng[2,3], MD; Yueh-Peng Chen[4], PhD; Quan-Ze Gao[4], PhD; Chihung Lin[4], PhD; Chang-Fu Kuo[2,4], PhD, MD; Chi-Chun Lai[2,3], MD

[1]Department of Ophthalmology, Chang Gung Memorial Hospital, Linkou Medical Center, Taoyuan, Taiwan

[2]College of Medicine, Chang Gung University, Taoyuan, Taiwan

[3]Department of Ophthalmology, Keelung Chang Gung Memorial Hospital, Keelung, Taiwan

[4]Center for Artificial Intelligence in Medicine, Chang Gung Memorial Hospital, Linkou Medical Center, Taoyuan, Taiwan

[*]these authors contributed equally

Corresponding Author:
Chi-Chun Lai, MD
Department of Ophthalmology
Keelung Chang Gung Memorial Hospital
No. 222, Maijin Rd
Keelung
Taiwan
Phone: 886 24313131 ext 6314
Email: Chichun.lai@gmail.com

Abstract

Background: Retinal vascular diseases, including diabetic macular edema (DME), neovascular age-related macular degeneration (nAMD), myopic choroidal neovascularization (mCNV), and branch and central retinal vein occlusion (BRVO/CRVO), are considered vision-threatening eye diseases. However, accurate diagnosis depends on multimodal imaging and the expertise of retinal ophthalmologists.

Objective: The aim of this study was to develop a deep learning model to detect treatment-requiring retinal vascular diseases using multimodal imaging.

Methods: This retrospective study enrolled participants with multimodal ophthalmic imaging data from 3 hospitals in Taiwan from 2013 to 2019. Eye-related images were used, including those obtained through retinal fundus photography, optical coherence tomography (OCT), and fluorescein angiography with or without indocyanine green angiography (FA/ICGA). A deep learning model was constructed for detecting DME, nAMD, mCNV, BRVO, and CRVO and identifying treatment-requiring diseases. Model performance was evaluated and is presented as the area under the curve (AUC) for each receiver operating characteristic curve.

Results: A total of 2992 eyes of 2185 patients were studied, with 239, 1209, 1008, 211, 189, and 136 eyes in the control, DME, nAMD, mCNV, BRVO, and CRVO groups, respectively. Among them, 1898 eyes required treatment. The eyes were divided into training, validation, and testing groups in a 5:1:1 ratio. In total, 5117 retinal fundus photos, 9316 OCT images, and 20,922 FA/ICGA images were used. The AUCs for detecting mCNV, DME, nAMD, BRVO, and CRVO were 0.996, 0.995, 0.990, 0.959, and 0.988, respectively. The AUC for detecting treatment-requiring diseases was 0.969. From the heat maps, we observed that the model could identify retinal vascular diseases.

Conclusions: Our study developed a deep learning model to detect retinal diseases using multimodal ophthalmic imaging. Furthermore, the model demonstrated good performance in detecting treatment-requiring retinal diseases.

(JMIR Med Inform 2021;9(5):e28868) doi:10.2196/28868

KEYWORDS

deep learning; retinal vascular diseases; multimodal imaging; treatment requirement; machine learning; eye; retinal; imaging; treatment; model; detection; vascular

Introduction

Background

Retinal vascular diseases, including diabetic macular edema (DME), neovascular age-related macular degeneration (nAMD), myopic choroidal neovascularization (mCNV), and retinal vein occlusion (RVO), highly affect visual function and lead to loss of working ability and impaired life quality [1-4]. Anti–vascular endothelial growth factor (VEGF) can improve visual outcomes for patients with retinal diseases [5]. Early disease detection and timely management can prevent disease progression and advanced visual impairment.

With its advancement in recent years, artificial intelligence has recently been used for several applications in the medical field, including for disease monitoring, diagnosis, and treatment [6]. In ophthalmology, deep learning—an artificial intelligence technique—can potentially detect eye diseases, such as diabetic retinopathy, glaucoma, nAMD, and retinopathy of prematurity, as well as refractive errors [7]. Different ocular pathologies can be identified using different imaging modalities. Multiple imaging modalities are available for retinal vascular disease diagnosis. Although the use of retinal fundus photography for diagnosis is feasible, robust diagnosis may require further imaging, such as through the use of optical coherence tomography (OCT), chorioretinal angiography (ie, fluorescein angiography [FA] and indocyanine green angiography [ICGA]), and optical coherence tomography angiography (OCTA). Deep learning has been applied for various imaging techniques. In addition to color fundus images, which are commonly used for detecting eye diseases [7], other imaging modalities are useful in deep learning–based applications. For example, OCT has been used for diagnosis and referral in patients with retinal diseases [8,9], and OCTA has been used for identifying nonperfusion areas in the retina [10].

Objective

Multimodal imaging in ophthalmology could improve the accuracy of disease diagnosis. The increased application of multiple imaging modalities for disease detection has led to advancements in deep learning–assisted disease diagnosis. An et al [11] used OCT combined with retinal fundus photography for glaucoma diagnosis. Meanwhile, Vaghefi et al [12] demonstrated an increased accuracy when using multimodal imaging to train an algorithm for OCT, OCTA, and retinal fundus photography for detecting dry AMD. However, little research has investigated the use of deep learning techniques in multimodal imaging for determining retinal vascular diseases. In our study, we developed a deep learning–based model for detecting retinal vascular diseases and diseases requiring anti-VEGF treatment through the use of multimodal retinal imaging, including color fundus photography, OCT, and FA with or without ICGA (FA/ICGA).

Methods

Study Participants

In this retrospective study, we included patients who underwent clinical examinations involving retinal fundus photography, OCT, and FA/ICGA from 2013 to 2019 at Chang Gung Memorial Hospital, Linkou Medical Center, Taipei and Keelung branches. The retinal fundus photos were obtained using 1 of the 2 color fundus cameras (Topcon Medical Systems; digital non-mydriatic retinal camera: Canon). OCT was performed using OCT machines (Heidelberg Engineering Inc; Avanti, Optovue Inc), and FA/ICGA images were obtained using fundus angiography machines (Heidelberg Engineering, Inc). The study protocol was approved by the Institutional Review Board of Chang Gung Memorial Hospital (no. 201900477B0), and the study adhered to the tenets of the Declaration of Helsinki.

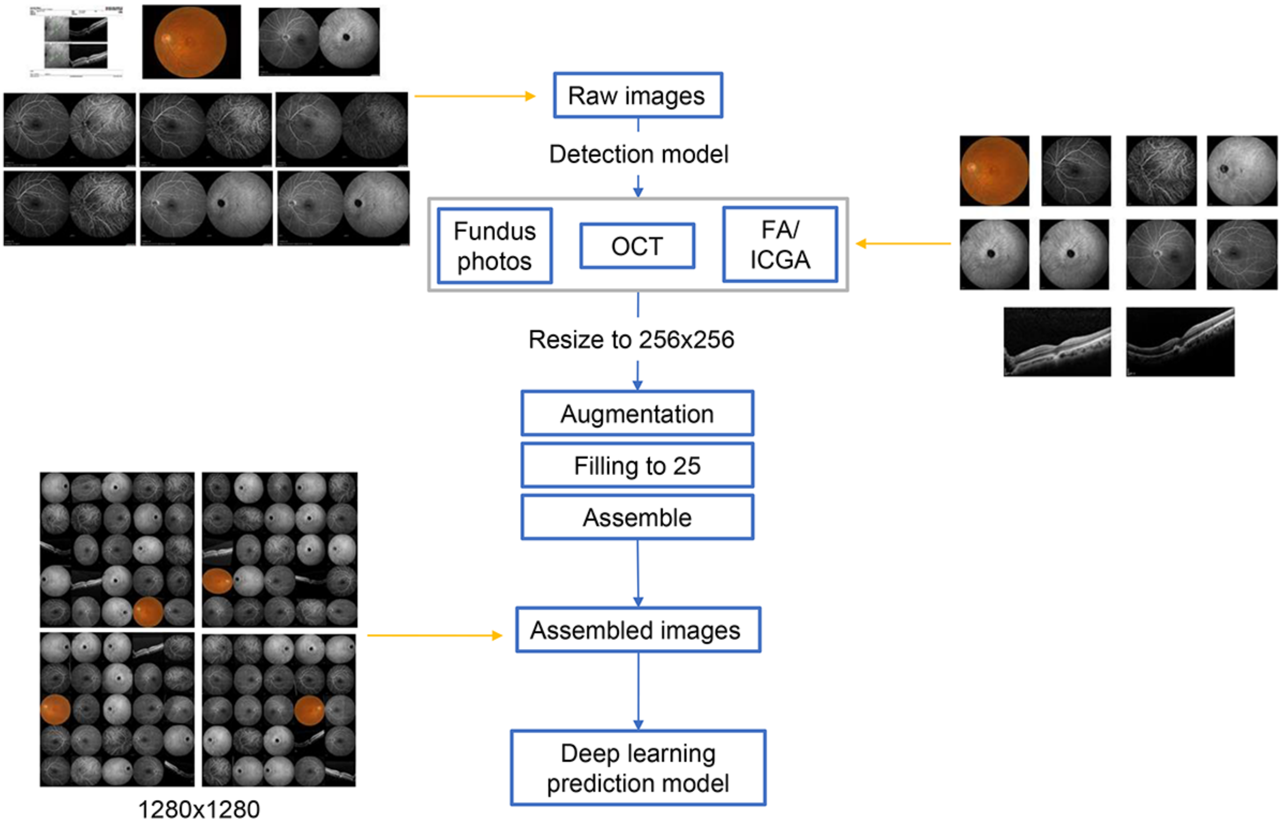
Data Classification

In our study, we identified retinal vascular diseases, including DME, nAMD, mCNV, branch retinal vein occlusion (BRVO), and central retinal vein occlusion (CRVO). Patients without a history of anti-VEGF treatment were included. After review of the multimodal images of each eye, disease diagnoses and need for anti-VEGF treatment were determined by 3 trained retinal ophthalmologists (LY, CHW, and SYP, who had 20, 10, and 6 years of clinical experience, respectively). Eye images were first reviewed by 2 of the retinal ophthalmologists (CHW and SYP). The ophthalmologists (CHW and SYP) excluded images with poor quality or nondifferentiable diagnosis. When the disease labels assigned by the ophthalmologists differed, a consensus was reached through discussion among all 3 retinal ophthalmologists. The senior retinal ophthalmologist (LY) again confirmed the image labels that were consistent in the first labeling. The patients were classified into DME, nAMD, mCNV, BRVO, and CRVO groups according to their disease diagnosis. The retinal ophthalmologists further defined diseases as anti-VEGF treatment requiring or non–treatment requiring. Based on the published literature, the treatment requirement was defined separately in each retinal vascular disease according to the features in different images [1,2,13-15]. Moreover, in the control group, we included patients who had undergone retinal fundus photography, OCT, and FA/ICGA examination for clinical purposes, but the examinations revealed no remarkable lesions or only lesions not related to retinal vascular diseases. For multimodal imaging, retinal fundus photos were macular centered; OCT images were fovea centered; and FA/ICGA images, which were randomly selected from different phases, were macular centered.

Data Management

The data management and image processing were performed on the same eye. We collected images of retinal fundus photography, OCT, and FA/ICGA from each eye. The flowchart of the image collection process is displayed in Figure 1. First, images were evaluated by the detection model to select and crop for different image types. The detection model, Cascade R-CNN [16], was trained with 599 images in different imaging modalities. The isolated images were first resized to 256 × 256 pixels. Subsequently, isolated images were augmented by slight adjustment of the brightness and contrast level, foggy masking, compression, rotation, horizontal flipping, and the addition of side lines. Then, 25 images were randomly selected from different imaging modalities and assembled. At least one image was required from each imaging modality. The assembled image package consisted of 25 segmented images from the same eye based on a combination of images with various augmentations and components of fundus retinal photography, OCT, and FA/ICGA. The size of the assembled images was 1280 × 1280 pixels, consisting of 25 images with a size of 256 × 256. Then, the image package was sent to the model for prediction.

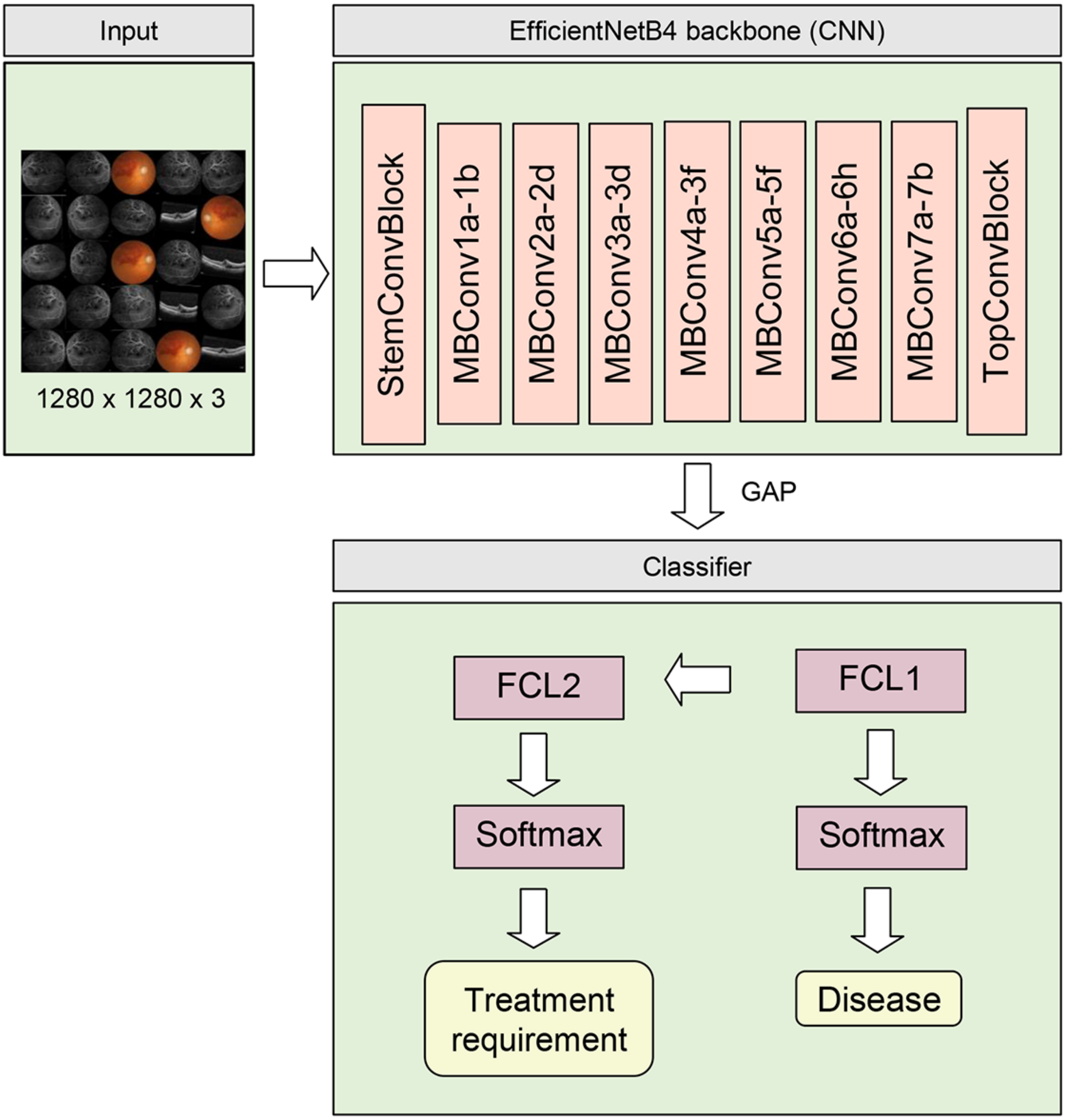
Figure 1. Flowchart of multimodal image management and processing. OCT: optical coherence tomography; FA/ICGA: fluorescein angiography with or without indocyanine green angiography.



Model Architecture

In our study, EfficientNetB4 was used as the convolutional neural network (CNN) for the classification model (Figure 2). Because our goal was to aid disease diagnosis and the detection of disease severity, the models had 2 outputs: (1) disease classification and (2) treatment requirement determination. However, features indicating severity may differ based on the disease. Our model first delivered disease prediction for differentiating different retinal vascular diseases. We then designed a layer consisting of a fully connected, reshaped, and weighted sum to facilitate the model classification of treatment requirement partially according to the results from the disease prediction part. In addition, to visualize the features for model prediction, heat maps were generated using gradient-weighted class activation mapping [17], which used the gradient based on the output scores to show the activation map for the specific image. The features of the heat maps were highlighted in a lighter color.

Figure 2. Architecture of the deep learning prediction model. CNN: convolutional neural network; FCL: fully connected layer; GAP: global average pooling.



Model Training

Image packages were split into training, validation, and testing data sets in a 5:1:1 ratio, respectively. The model was trained based on noisy student [18] pretrained weight and optimized using an AdamW optimizer [19]. The model was trained 3 times with different combinations of training and validation data sets. We also tested different parameters including learning rates of 1e-4, 1e-5, and 5e-5, and batch sizes of 8, 12, and 16. Subsequently, the model with the best performance in the training and validation data sets was selected and evaluated in the testing data set (Multimedia Appendices 1 and 2). The learning rate and batch size were set as 5e-5 and 16, respectively. Data preprocessing and the training and evaluation of the model were completed on a NVIDIA DGX-1 server with the Ubuntu 18.04 operating system. Image preprocessing, including conversion, augmentation, and assembly, was conducted using ImageMagick 7.0.10 [20]. Images were evaluated and cropped using Mmdetection 1.0.0 [21] and Pytorch 1.4.0 [22], and the bounding box was labeled using CocoAnnotator [23]. Tensorflow 2.2 [24] was used as the framework to train and evaluate the deep learning model.

Statistical Analysis

Receiver operating characteristic (ROC) curves were used for differentiating different retinal vascular diseases and treatment-requiring diseases, and the area under the curve (AUC) was measured for each ROC curve. Moreover, the sensitivity, specificity, and accuracy of the model were calculated. Regarding model performance in predicting different retinal diseases, the AUC, sensitivity, specificity, and accuracy were based on a one-versus-rest comparison. Additionally, a confusion matrix was created and demonstrated sensitivity in disease prediction. Statistical analysis was performed using the Sklearn 0.23.2 package in Python (Python Software Foundation).

Results

Study Participants and Data Distribution

In total, 2992 eyes of 2185 patients were included in our study. In the first labeling of 2992 eyes, 212 (7.08%) were differently labeled by CHW and SYP, and a consensus was reached after discussion among all 3 retinal ophthalmologists. Among the 2780 eyes with consistent labels in the first step, 144 (5.18%) eyes had different labels after review by LY, and a consensus was reached after discussion among all 3 retinal ophthalmologists. The distribution of the included eyes is shown in Table 1.

Table 1. Number of eyes included in the control and disease groups.

Groups	Total	Treatment-requiring	Non–treatment-requiring
Control	239	N/A[a]	N/A[a]
DME[b]	1209	788	421
nAMD[c]	1008	809	199
mCNV[d]	211	56	155
BRVO[e]	189	144	45
CRVO[f]	136	101	35
Total	2992	1898	855

[a] N/A: not applicable.

[b] DME: diabetic macular edema.

[c] nAMD: neovascular age-related macular degeneration.

[d] mCNV: myopic choroidal neovascularization.

[e] BRVO: branch retinal vein occlusion.

[f] CRVO: central retinal vein occlusion.

The control, DME, nAMD, mCNV, BRVO, and CRVO groups consisted of 239, 1209, 1008, 211, 189, and 136 eyes, respectively. Among all the disease groups, 788, 809, 56, 144, and 101 eyes required treatment in the DME, nAMD, mCNV, BRVO, and CRVO groups, respectively. Subsequently, 2138, 427, and 427 eyes were assigned to the training, validation, and testing data sets, respectively. We used 5117 retinal fundus photos, 9316 OCT images, and 20 922 FA/ICGA images, and the distribution of the images in different data sets is shown in Table 2.

Table 2. Distribution of image number used in different modalities for different data sets.

Modality	Total (n=2992)	Training (n=2138)	Validation (n=427)	Testing (n=427)
Retinal fundus photos	5117	3662	709	746
OCT[a]	9316	6704	1272	1340
FA/ICGA[b]	20922	14932	2959	3031

[a] OCT: optical coherence tomography.

[b] FA/ICGA: fluorescein angiography with or without indocyanine green angiography.

Model Performance

Model performance was evaluated using the testing data set. ROC curves are illustrated in Figure 3, and the AUC for each curve was determined. For disease identification, the overall AUC was 0.987, and the AUC was the highest in the mCNV (0.996) and control (0.996) groups, followed by the DME (0.995), nAMD (0.990), CRVO (0.988), and BRVO (0.959) groups. For predicting diseases requiring anti-VEGF treatment, the AUC was 0.969. Details regarding the model sensitivity and specificity are provided in Table 3. For retinal vascular disease prediction, the sensitivity was the highest for the control (0.971) group, followed by the nAMD (0.956), DME (0.940), and mCNV (0.933) groups, whereas the sensitivity of RVO identification was the lowest (0.690 for BRVO and 0.769 for CRVO). Regarding the prediction of diseases requiring anti-VEGF treatment, the sensitivity was 0.904 and specificity was 0.945. The accuracy for disease prediction was the highest in the control and mCNV (0.984) groups, followed by the BRVO and CRVO (0.977), DME (0.967), and nAMD (0.963) groups.

The accuracy for the detection of treatment-requiring diseases was 0.930. The confusion matrix is shown in Figure 4.

Figure 3. Receiver operating characteristic curves of the model performance for (A) predicting different retinal vascular diseases and (B) identifying treatment-requiring diseases. AUC: area under the curve; BRVO: branch retinal vein occlusion; CRVO: central retinal vein occlusion; DME: diabetic macular edema; mCNV: myopic choroidal neovascularization; nAMD: neovascular age-related macular degeneration.

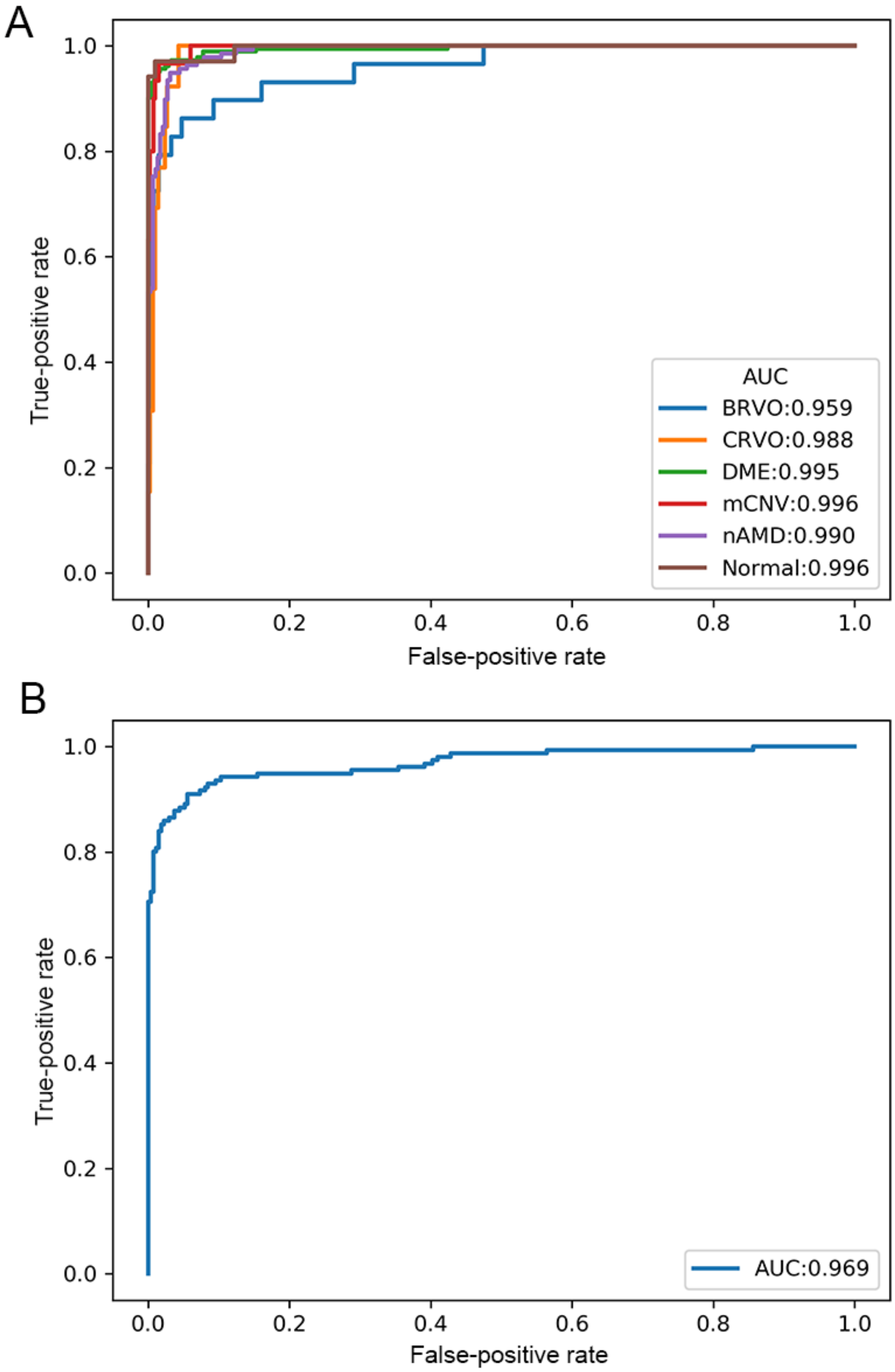


Table 3. Sensitivity, specificity, and accuracy of the model in the prediction of retinal vascular diseases and treatment-requiring diseases.

Value	Sensitivity	Specificity	Accuracy
Control	0.971	0.985	0.984
DME[a]	0.940	0.988	0.967
nAMD[b]	0.956	0.966	0.963
mCNV[c]	0.933	0.987	0.984
BRVO[d]	0.690	0.997	0.977
CRVO[e]	0.769	0.983	0.977
Treatment requirement	0.904	0.945	0.930

[a] DME: diabetic macular edema.

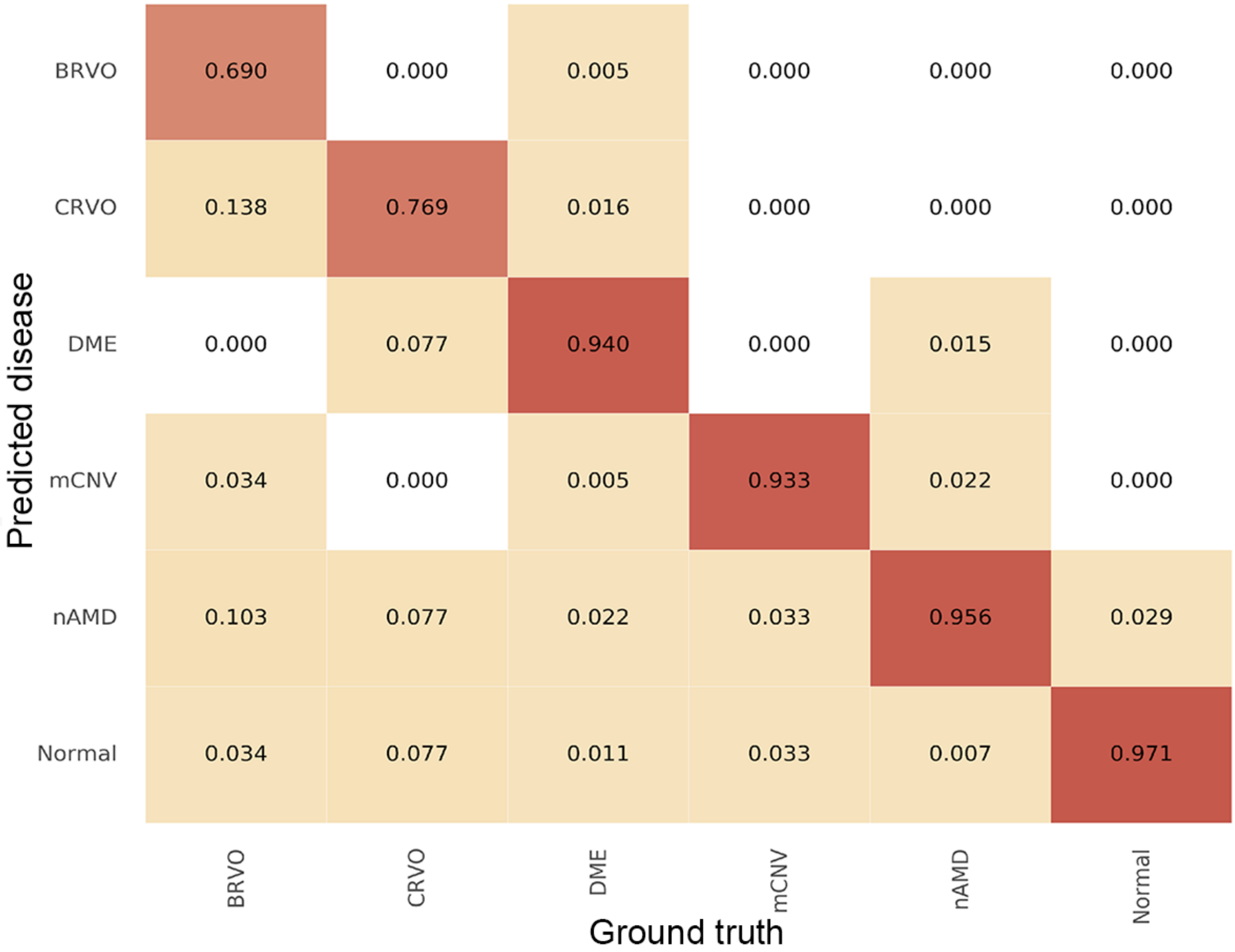
[b] nAMD: neovascular age-related macular degeneration.

[c] mCNV: myopic choroidal neovascularization.

[d] BRVO: branch retinal vein occlusion.

[e] CRVO: central retinal vein occlusion.

Figure 4. Confusion matrix demonstrating the performance of the prediction model in different retinal vascular diseases. BRVO: branch retinal vein occlusion; CRVO: central retinal vein occlusion; DME: diabetic macular edema; mCNV: myopic choroidal neovascularization; nAMD: neovascular age-related macular degeneration.

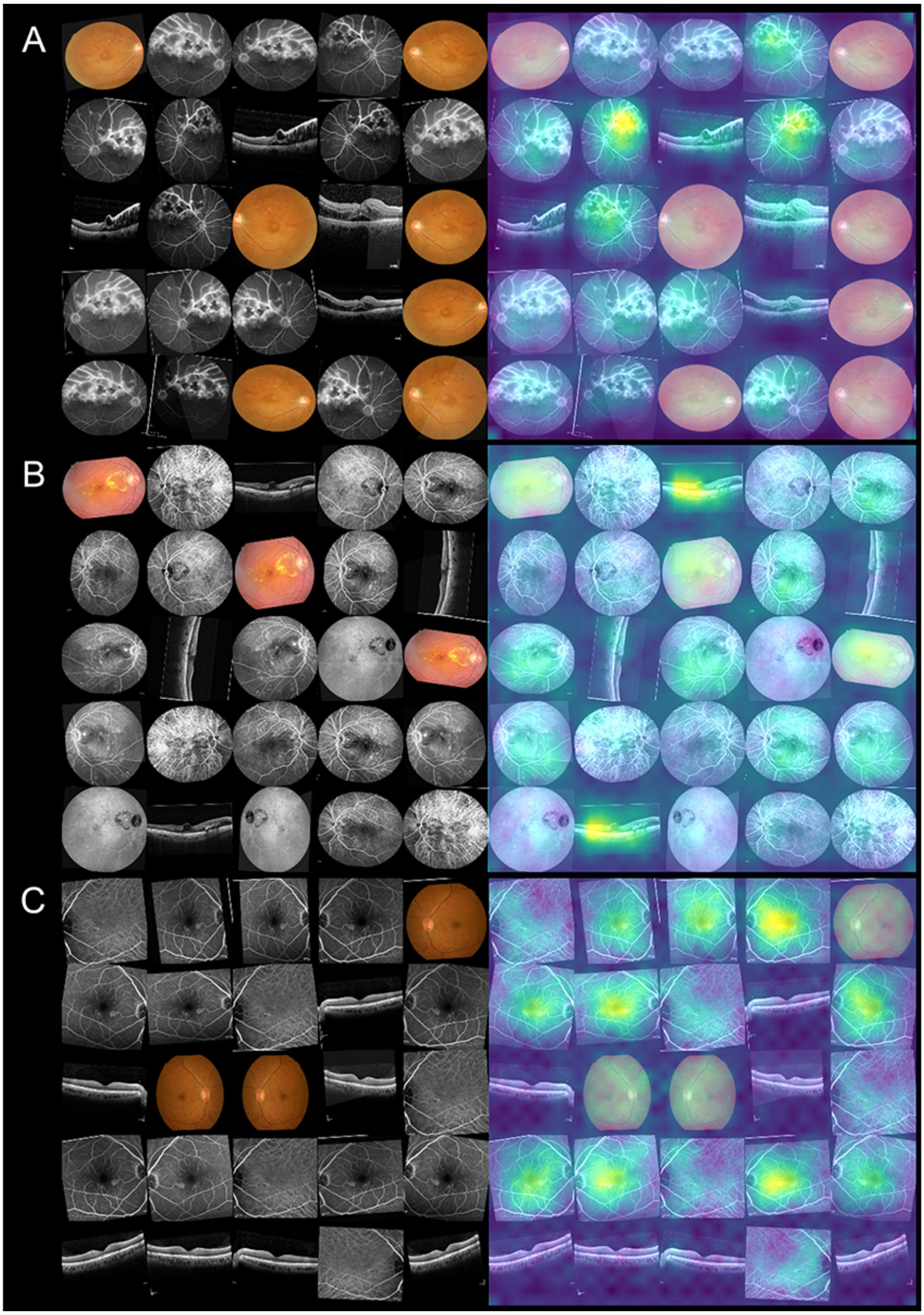


Predicted disease	BRVO	CRVO	DME	mCNV	nAMD	Normal
BRVO	0.690	0.000	0.005	0.000	0.000	0.000
CRVO	0.138	0.769	0.016	0.000	0.000	0.000
DME	0.000	0.077	0.940	0.000	0.015	0.000
mCNV	0.034	0.000	0.005	0.933	0.022	0.000
nAMD	0.103	0.077	0.022	0.033	0.956	0.029
Normal	0.034	0.077	0.011	0.033	0.007	0.971

Heat Maps for Model Prediction

Heat maps for visual explanations of our model predictions were generated using gradient-weighted class activation mapping, and the samples are shown in Figure 5. In the heat maps, the model could simultaneously identify the lesion in different imaging modalities. Regarding different retinal vascular diseases, the model had different weights in different image modalities. For example, in eyes with RVO, the model highlighted the exudates and hemorrhage dot in retinal fundus photos, ischemic area, and leaking point in FA/ICGA. In patients requiring treatment for DME, the model highlighted retinal vessels within the macula in retinal images, the central swelling area in OCT images, and the leaking or staining lesions in FA/ICGA images.

Figure 5. Sample heat maps generated by the prediction model in a true-positive patient with (A) treatment-requiring branch retinal vein occlusion, (B) treatment-requiring diabetic macular edema, and (C) non–treatment-requiring age-related macular degeneration.



Discussion

Main Findings

In this study, we used multimodal imaging to develop a deep learning–based model for the prediction of retinal vascular diseases, including DME, nAMD, mCNV, BRVO, and CRVO, and to determine whether anti-VEGF treatment was required. This model had average AUCs of 0.987 and 0.969 for predicting retinal vascular diseases and for predicting treatment-requiring diseases, respectively. The heat map shows that the model can identify disease features through multimodal retinal imaging.

Ophthalmology Imaging in Deep Learning

Previous studies have proven the efficacy of using different image modalities in deep learning–based models for predicting retinal diseases. In addition to retinal fundus images for identifying diabetic retinopathy, AMD, and glaucoma [7], a deep learning model using OCT for retinal layer segmentation and retinal disease identification was developed by the DeepMind group [8]. Moreover, deep learning could help to detect ischemic zones in retinal vascular diseases through the use of ultra-wide-field FA [25]. The aforementioned studies demonstrated that deep learning can be effectively applied for a single retinal imaging modality. However, few investigations have been conducted to study the application of deep learning models for predicting diseases using more than one retinal imaging modality. OCT and retinal fundus images have been used concomitantly for dry AMD [26] and glaucoma [11] diagnosis. However, previous studies have either used a single imaging modality or focused on predicting a single retinal disease. To date, few studies have evaluated the performance of deep learning models with multimodal retinal imaging for predicting multiple retinal vascular diseases.

Multimodal Imaging–Based Deep Learning Model for Retinal Vascular Diseases

To our knowledge, this is the first study to use multimodal deep learning–based architecture for detecting multiple retinal vascular diseases. In our study, we used multiple image modalities, including retinal fundus photography, OCT, and FA/ICGA, for predicting neovascular retinal diseases, including DME, nAMD, mCNV, and RVO [27]. Furthermore, this model can identify diseases requiring anti-VEGF treatment. In clinical settings, multimodal retinal images are crucial for ophthalmologists to treat retinal diseases. Occasionally, a feature in a retinal image modality may be shared by many retinal diseases. For example, increased central retinal thickness in OCT can be present in DME, nAMD, mCNV, and RVO, but retinal fundus images may vary among these diseases. The features of nAMD and mCNV may appear similar in retinal fundus images, and an ICGA is needed for differentiating them [2]. Therefore, multimodal imaging is required for the diagnosis and treatment determination of different retinal diseases [28]. Our model with multimodal imaging was similar to real-world ophthalmology practice with regard to the diagnosis for multiple retinal diseases and determination of disease severity. In real-world practice, the model may help with the screening of the diseases and treatment-requiring status, saving ophthalmologist's time and effort on reviewing the images. Although the AUC of different retinal vascular diseases demonstrated excellent differentiation, defined as AUC > 0.8 [29], the RVO groups showed relatively low sensitivity. This might be related to the low number of eyes used for model training. In the future investigation, the generative adversarial network may be implemented to synthesize ophthalmic images and solve the problem of an inadequate number of images [30].

Detection of Treatment-Requiring Retinal Vascular Diseases

Because expenses involved in using anti-VEGF drugs in the treatment of retinal vascular diseases are high, patients being administered these drugs may need to meet strict criteria to claim reimbursement from insurance companies in many regions [31]. In Taiwan, the use of intravitreal anti-VEGF treatment for DME, nAMD, mCNV, and RVO requires prereview by members of the Taiwan National Health Insurance program for reimbursement [32,33]. An efficient and accurate method for evaluating a patient's retinal vascular disease status and disease severity may be essential. Our model could not only aid ophthalmologists in disease diagnosis and in determining the need for anti-VEGF treatment for retinal vascular diseases but also help with the prereview of anti-VEGF treatment.

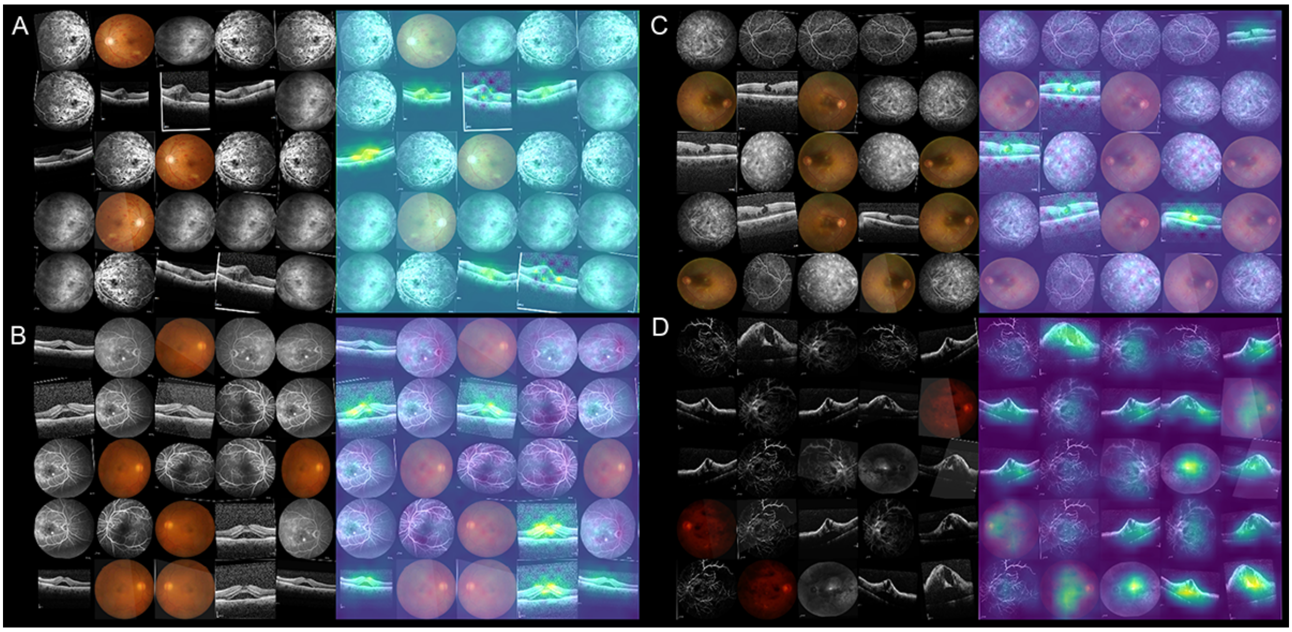
Image Variability for the Model

The model developed in the present study is highly flexible in terms of image input. It does not depend on a fixed image distribution for different modalities. The only requirement is at least one image for each imaging modality. We investigated the model accuracy for packages with different numbers of images, and 25 images in a 5 × 5 matrix had the highest performance. Moreover, we tested different CNN models and different compositions of imaging modalities to determine which could achieve the highest accuracy (Multimedia Appendix 3). Using the CNN of EfficientNetB4 with images of retinal fundus photography, OCT, and FA/ICGA had the best performance. The images from the same eye can be randomly arranged or augmented during the preprocessing stage before being used in the prediction model. The visualized heat maps show that the model has the ability of simultaneous differentiation of retinal diseases with the use of different imaging modalities. With DME, for example, both the central retina in OCT and the leaking points in FA had high weightage. For BRVO, the model highlights areas with hemorrhage in retinal fundus images, increased retinal thickness in OCT images, and nonperfusion in FA images. These findings are compatible with the clinical features of retinal diseases [34,35] and indicate that our model produces reasonable and reliable predictions of retinal vascular diseases.

False Prediction of the Model

Regarding false predictions of retinal diseases, sample heat maps are presented in Figure 6. We observed that the model provided wrong predictions mostly for eyes with advanced-stage diseases or coexisting retinal diseases. Retinal vascular diseases may share undistinguishable features in advanced stages. For example, in an advanced stage of a disease, retinal hemorrhage, retinal nonperfusion, and macular edema could appear to have the same prominence in CRVO as in DME and advanced diabetic retinopathy [36,37]. The coexistence of diabetic retinopathy with DME may produce clinical features similar to those of RVO with macular edema [38]. Additionally, other retinal disorders, such as central serous chorioretinopathy, may display features similar to those of retinal vascular diseases and lead to misdiagnosis by the model. As for diseases requiring anti-VEGF treatment, false prediction was noted in cases with borderline disease activity or other retinal disorders, such as central serous chorioretinopathy and epiretinal membrane, for which anti-VEGF treatment is not indicated.

Figure 6. Sample heat maps for false prediction of the model: (A) false prediction of treatment-requiring diabetic macular edema (DME) in a patient with coexisting DME and central retinal vein occlusion (CRVO); (B) false prediction of treatment-requiring age-related macular degeneration (AMD) in a patient with central serous chorioretinopathy; (C) false prediction of treatment-requiring DME in a patient with epiretinal membrane, lamellar macular hole, and diabetic retinopathy; (D) false prediction of treatment-requiring DME in a patient with advanced CRVO.



Study Limitations

This study had some limitations. First, the model requires the use of multiple image modalities, including OCT and FA/ICGA, which some eye-care facilities may not be equipped with. Although the study focused on deep learning–based prediction with multimodal imaging, clinical application may require more investigation. Second, images used in the study underwent quality checks. The efficacy during application to a real-world clinical setting may be affected by the patient's condition and the image quality [39]. Additionally, some ocular diseases affecting image signal transmission could affect image quality and retinal disease diagnosis [40,41]. Third, images from different machine manufacturers not included in our study might have affected the model accuracy. A transfer learning approach could be adopted in cases where images are obtained from different machine manufacturers. Fourth, we did not consider other retinal vascular diseases, such as retinal neovascularization caused by uveitis or infection. The model is inapplicable to diseases not included in our study. Fifth, we only identified disease statuses that may require anti-VEGF treatment. Disease statuses requiring other treatments, such as laser therapy, were not analyzed in the current study. Furthermore, images of the most advanced disease stages with features such as severe vitreous hemorrhage or diffused chorioretinal atrophy would have been excluded due to nondifferentiable diagnosis. Sixth, a relatively small number of eyes in the RVO groups led to decreased accuracy in disease prediction and more data may be needed for better model performance. Last, the study group only included patients without previous anti-VEGF treatment. The accuracy in patients with a history of anti-VEGF treatment needs further investigation.

Conclusions

We developed a deep learning–based model using multimodal imaging for predicting retinal vascular diseases and determining whether anti-VEGF treatment is required. This model can facilitate the differentiation of DME, nAMD, mCNV, BRVO, and CRVO and help in determining the indication for anti-VEGF treatment.

Acknowledgments

The authors thank Andy Po-Chun Kang, MSc, for English editing and technical support. The authors would also like to thank the Maintenance Project of the Center for Artificial Intelligence in Medicine at Chang Gung Memorial Hospital, Taiwan (CLRPG3H0012, CIRPG3H0012) for statistical assistance and study support. Chang Gung Memorial Hospital, Taiwan, provided additional support to this study (CMRPG3L0251, CMRPG2J0211, and CMRPG2J0212). This manuscript was edited by Wallace Academic Editing. The founder had no role in the study design or interpretation of the results.

Authors' Contributions

EYCK, CCL, LY, and CFK contributed to conception and design of the study. Data were collected by SYP and CHW. Data analysis was conducted by YLL, CHW, and SYP. CFK, CCL, LY, YPC, QZG, and CHL contributed to data interpretation. EYCK and YLL wrote the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Splitting of the data sets and training process of the model.
[PNG File , 100 KB - medinform_v9i5e28868_app1.png]

Multimedia Appendix 2
Performance of the model trained with different (A) learning rates and (B) batch sizes. AUC: area under the curve.
[PDF File (Adobe PDF File), 80 KB - medinform_v9i5e28868_app2.pdf]

Multimedia Appendix 3
Performance of the model with different convolutional neural networks and composition of imaging modality. AUC: area under the curve; CF: color fundus photography; CNN: convolutional neural network; FA/ICGA: fluorescein angiography with or without indocyanine green angiography; OCT: optical coherence tomography.
[PNG File , 115 KB - medinform_v9i5e28868_app3.png]

References

1. Tan GS, Cheung N, Simó R, Cheung GCM, Wong TY. Diabetic macular oedema. Lancet Diabetes Endocrinol 2017 Feb;5(2):143-155. [doi: 10.1016/S2213-8587(16)30052-3] [Medline: 27496796]
2. Cheung CMG, Arnold JJ, Holz FG, Park KH, Lai TYY, Larsen M, et al. Myopic choroidal neovascularization: review, guidance, and consensus statement on management. Ophthalmology 2017 Nov;124(11):1690-1711. [doi: 10.1016/j.ophtha.2017.04.028] [Medline: 28655539]
3. Hayreh SS. Ocular vascular occlusive disorders: natural history of visual outcome. Prog Retin Eye Res 2014 Jul;41:1-25 [FREE Full text] [doi: 10.1016/j.preteyeres.2014.04.001] [Medline: 24769221]
4. Lim LS, Mitchell P, Seddon JM, Holz FG, Wong TY. Age-related macular degeneration. Lancet 2012 May 05;379(9827):1728-1738. [doi: 10.1016/S0140-6736(12)60282-7] [Medline: 22559899]
5. Kim LA, D'Amore PA. A brief history of anti-VEGF for the treatment of ocular angiogenesis. Am J Pathol 2012 Aug;181(2):376-379 [FREE Full text] [doi: 10.1016/j.ajpath.2012.06.006] [Medline: 22749677]
6. Amisha, Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. J Family Med Prim Care 2019 Jul;8(7):2328-2331 [FREE Full text] [doi: 10.4103/jfmpc.jfmpc_440_19] [Medline: 31463251]
7. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol 2019 Feb;103(2):167-175 [FREE Full text] [doi: 10.1136/bjophthalmol-2018-313173] [Medline: 30361278]
8. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 2018 Dec;24(9):1342-1350. [doi: 10.1038/s41591-018-0107-6] [Medline: 30104768]
9. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018 Feb 22;172(5):1122-1131.e9. [doi: 10.1016/j.cell.2018.02.010] [Medline: 29474911]
10. Guo Y, Hormel TT, Xiong H, Wang B, Camino A, Wang J, et al. Development and validation of a deep learning algorithm for distinguishing the nonperfusion area from signal reduction artifacts on OCT angiography. Biomed Opt Express 2019 Jul 01;10(7):3257-3268 [FREE Full text] [doi: 10.1364/BOE.10.003257] [Medline: 31360599]
11. An G, Omodaka K, Hashimoto K, Tsuda S, Shiga Y, Takada N, et al. Glaucoma diagnosis with machine learning based on optical coherence tomography and color fundus images. J Healthc Eng 2019;2019:4061313 [FREE Full text] [doi: 10.1155/2019/4061313] [Medline: 30911364]
12. Vaghefi E, Hill S, Kersten HM, Squirrell D. Multimodal retinal image analysis via deep learning for the diagnosis of intermediate dry age-related macular degeneration: a feasibility study. J Ophthalmol 2020;2020:7493419 [FREE Full text] [doi: 10.1155/2020/7493419] [Medline: 32411434]
13. Brand CS. Management of retinal vascular diseases: a patient-centric approach. Eye (Lond) 2012 Apr;26 Suppl 2:S1-16 [FREE Full text] [doi: 10.1038/eye.2012.32] [Medline: 22495396]

14. Flaxel CJ, Adelman RA, Bailey ST, Fawzi A, Lim JI, Vemulakonda GA, et al. Age-related macular degeneration preferred practice pattern. Ophthalmology 2020 Jan;127(1):P1-P65. [doi: 10.1016/j.ophtha.2019.09.024] [Medline: 31757502]
15. Rehak M, Wiedemann P. Retinal vein thrombosis: pathogenesis and management. J Thromb Haemost 2010 Sep;8(9):1886-1894 [FREE Full text] [doi: 10.1111/j.1538-7836.2010.03909.x] [Medline: 20492457]
16. Cai Z, Vasconcelos N. Cornell University.: Delving into high quality object detection; 2017. URL: https://arxiv.org/abs/1712.00726 [accessed 2021-01-02]
17. Selvaraju R, Das A, Vedantam R, Cogswell M, Parikh D, Batra D. Why did you say that? Visual explanations from deep networks via gradient-based localization. arXiv. Cornell University. Grad-CAM; 2016. URL: https://arxiv.org/abs/1610.02391v3 [accessed 2020-11-07]
18. Xie Q, Hovy E, Luong M, Le Q. Self-training with noisy student improves ImageNet classification. Cornell University. 2019. URL: https://arxiv.org/abs/1911.04252 [accessed 2020-11-07]
19. Loshchilov I, Hutter F. Decoupled weight decay regularization. Cornell University. 2017. URL: https://arxiv.org/abs/1711.05101 [accessed 2020-11-07]
20. ImageMagick. ImageMagick. URL: https://imagemagick.org/ [accessed 2020-11-25]
21. Chen K, Wang J, Pang J, Cao Y, Xiong Y, Li X. Open MMLab detection toolbox and benchmark. Cornell University. 2019. URL: https://arxiv.org/abs/1906.07155 [accessed 2020-12-01]
22. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z. Automatic differentiation in PyTorch. OpenReview. URL: https://openreview.net/forum?id=BJJsrmfCZ [accessed 2020-12-01]
23. Brooks J. COCOAnnotator. GitHub. 2019. URL: https://github.com/jsbroks/coco-annotator [accessed 2020-12-01]
24. TensorFlow. 2015. URL: https://www.tensorflow.org/ [accessed 2020-12-01]
25. Nunez do Rio JM, Sen P, Rasheed R, Bagchi A, Nicholson L, Dubis AM, et al. Deep learning-based segmentation and quantification of retinal capillary non-perfusion on ultra-wide-field retinal fluorescein angiography. J Clin Med 2020 Aug 06;9(8):2537 [FREE Full text] [doi: 10.3390/jcm9082537] [Medline: 32781564]
26. Yoo TK, Choi JY, Seo JG, Ramasubramanian B, Selvaperumal S, Kim DW. The possibility of the combination of OCT and fundus images for improving the diagnostic accuracy of deep learning for age-related macular degeneration: a preliminary experiment. Med Biol Eng Comput 2019 Mar;57(3):677-687. [doi: 10.1007/s11517-018-1915-z] [Medline: 30349958]
27. Taban M, Sharma S, Williams DR, Waheed N, Kaiser PK. Comparing retinal thickness measurements using automated fast macular thickness map versus six-radial line scans with manual measurements. Ophthalmology 2009 May;116(5):964-970. [doi: 10.1016/j.ophtha.2008.12.033] [Medline: 19410954]
28. Novais EA, Baumal CR, Sarraf D, Freund KB, Duker JS. Multimodal imaging in retinal disease: a consensus definition. Ophthalmic Surg Lasers Imaging Retina 2016 Mar;47(3):201-205. [doi: 10.3928/23258160-20160229-01] [Medline: 26985792]
29. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol 2010 Sep;5(9):1315-1316 [FREE Full text] [doi: 10.1097/JTO.0b013e3181ec173d] [Medline: 20736804]
30. Odaibo S. Generative adversarial networks synthesize realistic OCT images of the retina. Cornell University. URL: https://arxiv.org/abs/1902.06676 [accessed 2021-04-16]
31. Baker-Schena L. Expensive drugs. American Academy of Ophthalmology. URL: https://www.aao.org/eyenet/article/expensive-drugs [accessed 2020-12-07]
32. Chou Y, Chen M, Lin T, Chen S, Hwang D. Priority options of anti-vascular endothelial growth factor agents in wet age-related macular degeneration under the National Health Insurance Program. J Chin Med Assoc 2019 Aug;82(8):659-664. [doi: 10.1097/JCMA.0000000000000138] [Medline: 31259835]
33. Tsai M, Hsieh Y, Peng Y. Real-life experience of ranibizumab for diabetic macular edema in Taiwan. Int Ophthalmol 2019 Jul;39(7):1511-1522. [doi: 10.1007/s10792-018-0970-7] [Medline: 29926364]
34. Kwan CC, Fawzi AA. Imaging and biomarkers in diabetic macular edema and diabetic retinopathy. Curr Diab Rep 2019 Aug 31;19(10):95. [doi: 10.1007/s11892-019-1226-2] [Medline: 31473838]
35. Jaulim A, Ahmed B, Khanam T, Chatziralli IP. Branch retinal vein occlusion: epidemiology, pathogenesis, risk factors, clinical features, diagnosis, and complications. An update of the literature. Retina 2013 May;33(5):901-910. [doi: 10.1097/IAE.0b013e3182870c15] [Medline: 23609064]
36. Hayreh SS, Zimmerman MB. Fundus changes in central retinal vein occlusion. Retina 2015 Jan;35(1):29-42 [FREE Full text] [doi: 10.1097/IAE.0000000000000256] [Medline: 25084156]
37. Viswanath K, McGavin DDM. Diabetic retinopathy: clinical findings and management. Community Eye Health 2003;16(46):21-24 [FREE Full text] [Medline: 17491851]
38. Schmidt-Erfurth U, Garcia-Arumi J, Gerendas BS, Midena E, Sivaprasad S, Tadayoni R, et al. Guidelines for the management of retinal vein occlusion by the European Society of Retina Specialists (EURETINA). Ophthalmologica 2019;242(3):123-162 [FREE Full text] [doi: 10.1159/000502041] [Medline: 31412332]
39. Beede E, Baylor E, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. CHI'20 2020 Dec:1-12. [doi: 10.1145/3313831.3376718]

40. Hsieh Y, Chuang L, Jiang Y, Chang T, Yang C, Yang C, et al. Application of deep learning image assessment software VeriSee™ for diabetic retinopathy screening. J Formos Med Assoc 2021 Jan:165-171 [FREE Full text] [doi: 10.1016/j.jfma.2020.03.024] [Medline: 32307321]
41. Kang EY, Hsieh Y, Li C, Huang Y, Kuo C, Kang J, et al. Deep learning-based detection of early renal function impairment using retinal fundus images: model development and validation. JMIR Med Inform 2020 Nov 26;8(11):e23472 [FREE Full text] [doi: 10.2196/23472] [Medline: 33139242]

Abbreviations

AUC: area under the curve
BRVO: branch retinal vein occlusion
CNN: convolutional neural network
CRVO: central retinal vein occlusion
DME: diabetic macular edema
FA: fluorescein angiography
mCNV: myopic choroidal neovascularization
nAMD: neovascular age-related macular degeneration
OCTA: optical coherence tomography angiography
ROC: receiver operating characteristic
VEGF: vascular endothelial growth factor

Edited by G Eysenbach; submitted 17.03.21; peer-reviewed by YT Hsieh, A Staffini; comments to author 08.04.21; revised version received 18.04.21; accepted 03.05.21; published 31.05.21.

Please cite as:
Kang EYC, Yeung L, Lee YL, Wu CH, Peng SY, Chen YP, Gao QZ, Lin C, Kuo CF, Lai CC
A Multimodal Imaging–Based Deep Learning Model for Detecting Treatment-Requiring Retinal Vascular Diseases: Model Development and Validation Study
JMIR Med Inform 2021;9(5):e28868
URL: https://medinform.jmir.org/2021/5/e28868
doi:10.2196/28868
PMID:34057419

Original Paper

Pathway-Driven Coordinated Telehealth System for Management of Patients With Single or Multiple Chronic Diseases in China: System Development and Retrospective Study

Zheyu Wang[1], BSc; Jiye An[1], PhD; Hui Lin[1], BSc; Jiaqiang Zhou[2], MD; Fang Liu[3], MSc; Juan Chen[3], MD; Huilong Duan[1], PhD; Ning Deng[1], PhD

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

[2]Sir Run-Run Shaw Hospital, Zhejiang University School of Medicine, Zhejiang University, Hangzhou, China

[3]General Hospital of Ningxia Medical University, Yinchuan, China

Corresponding Author:
Ning Deng, PhD
Ministry of Education Key Laboratory of Biomedical Engineering
College of Biomedical Engineering and Instrument Science
Zhejiang University
38 Zheda Rd, Zhouyiqing Bldg 512
Yuquan Campus
Hangzhou, 310027
China
Phone: 86 57122952693
Email: zju.dengning@gmail.com

Abstract

Background: Integrated care enhanced with information technology has emerged as a means to transform health services to meet the long-term care needs of patients with chronic diseases. However, the feasibility of applying integrated care to the emerging “three-manager” mode in China remains to be explored. Moreover, few studies have attempted to integrate multiple types of chronic diseases into a single system.

Objective: The aim of this study was to develop a coordinated telehealth system that addresses the existing challenges of the “three-manager” mode in China while supporting the management of single or multiple chronic diseases.

Methods: The system was designed based on a tailored integrated care model. The model was constructed at the individual scale, mainly focusing on specifying the involved roles and responsibilities through a universal care pathway. A custom ontology was developed to represent the knowledge contained in the model. The system consists of a service engine for data storage and decision support, as well as different forms of clients for care providers and patients. Currently, the system supports management of three single chronic diseases (hypertension, type 2 diabetes mellitus, and chronic obstructive pulmonary disease) and one type of multiple chronic conditions (hypertension with type 2 diabetes mellitus). A retrospective study was performed based on the long-term observational data extracted from the database to evaluate system usability, treatment effect, and quality of care.

Results: The retrospective analysis involved 6964 patients with chronic diseases and 249 care providers who have registered in our system since its deployment in 2015. A total of 519,598 self-monitoring records have been submitted by the patients. The engine could generate different types of records regularly based on the specific care pathway. Results of the comparison tests and causal inference showed that a part of patient outcomes improved after receiving management through the system, especially the systolic blood pressure of patients with hypertension (P<.001 in all comparison tests and an approximately 5 mmHg decrease after intervention via causal inference). A regional case study showed that the work efficiency of care providers differed among individuals.

Conclusions: Our system has potential to provide effective management support for single or multiple chronic conditions simultaneously. The tailored closed-loop care pathway was feasible and effective under the “three-manager” mode in China. One direction for future work is to introduce advanced artificial intelligence techniques to construct a more personalized care pathway.

(JMIR Med Inform 2021;9(5):e27228) doi:10.2196/27228

KEYWORDS

chronic disease; telehealth system; integrated care; pathway; ontology

Introduction

Background

Chronic diseases are the most prevalent and costly health conditions worldwide [1]. Patient self-management combined with timely intervention from care providers are essential to control the progression of chronic diseases [2]. However, within traditional care settings, the disconnected and time-consuming management procedure is unable to meet the long-term care needs of patients [3,4]. Furthermore, several patients with chronic diseases live with more than one chronic condition (ie, multiple chronic conditions [MCC]), which create diverse, and sometimes contradictory, needs for health services [5,6].

Integrated care has been proposed as a means to meet the above challenges by transforming traditional health services [7]. In an integrated care setting, health services are provided by a coordinated multidisciplinary team of care providers. The core objective is to implement patient-centered health systems through comprehensive delivery of quality services across the life course [8]. Further, the advent of information technologies has promoted the delivery of integrated care services, which can be understood from different scales. From an individual scale, information technologies are crucial to facilitate the development of shared care plans [7], which clearly articulates the roles of care providers and patients in the care process to deliver more personalized and targeted care [9]. From a group scale, information technologies play a key role in achieving the goals of bidirectional communication within care provider teams and provision of continuous self-management support to patients [10].

As practical applications of information technologies in the health care domain, telehealth systems have demonstrated potential to improve the outcomes of chronic disease management [11-14]. Patient self-monitoring at home and remote guidance from care providers can be realized with the assistance of telehealth systems [15]. However, most of these systems focus on a single chronic disease, and few studies have attempted to integrate multiple types of chronic diseases into one system [16-19]. A telehealth system designed for managing multiple chronic illnesses simultaneously can not only support the management of patients with MCC but can also reduce the cost of developing multiple telehealth systems for managing different chronic diseases.

In China, the current health service system is in a three-tier form: community health service institutions at the bottom, secondary hospitals in the middle, and tertiary hospitals at the top [20]. General practitioners (GPs) are at the core of primary health care (ie, community level), providing basic treatment and long-term care for patients, especially in the management of chronic diseases [21]. In response to the government policy on promoting integrated care [22], specialists and case managers (CMs) are gradually involved in the management to form a coordinated multidisciplinary team called the "three-manager" mode [23]. The specialists are mainly from secondary or tertiary hospitals, providing more specialized and professional treatment [21]. CMs are mainly composed of nurses who work together with GPs to assist them in their daily work, similar to other countries [24,25].

The "three-manager" mode has been implemented in several provinces of China; however, there remain some practical challenges, which result in a significant gap between standards of care and medical practice [26-28]. First, current management guidelines [29-31] do not clearly specify the responsibility of each role in the "three-manager" mode. Second, the unbalanced allocation of medical resources in China leads to a difference in the abilities of primary care providers [32-34]. GPs and CMs in remote areas may rarely perform comprehensive and effective management following the guidelines.

Objectives

In this paper, we present the design, development, and retrospective evaluation of a telehealth system that simultaneously supports the management of single or multiple chronic diseases. The proposed system aims to address the existing challenges of the "three-manager" mode in China through a tailored integrated care model, mainly focusing on specifying the responsibility of involved roles and providing a universal care pathway for common chronic diseases. Currently, the system supports three main chronic conditions in China [35]: hypertension (HTN), type 2 diabetes mellitus (T2DM), and chronic obstructive pulmonary disease (COPD).

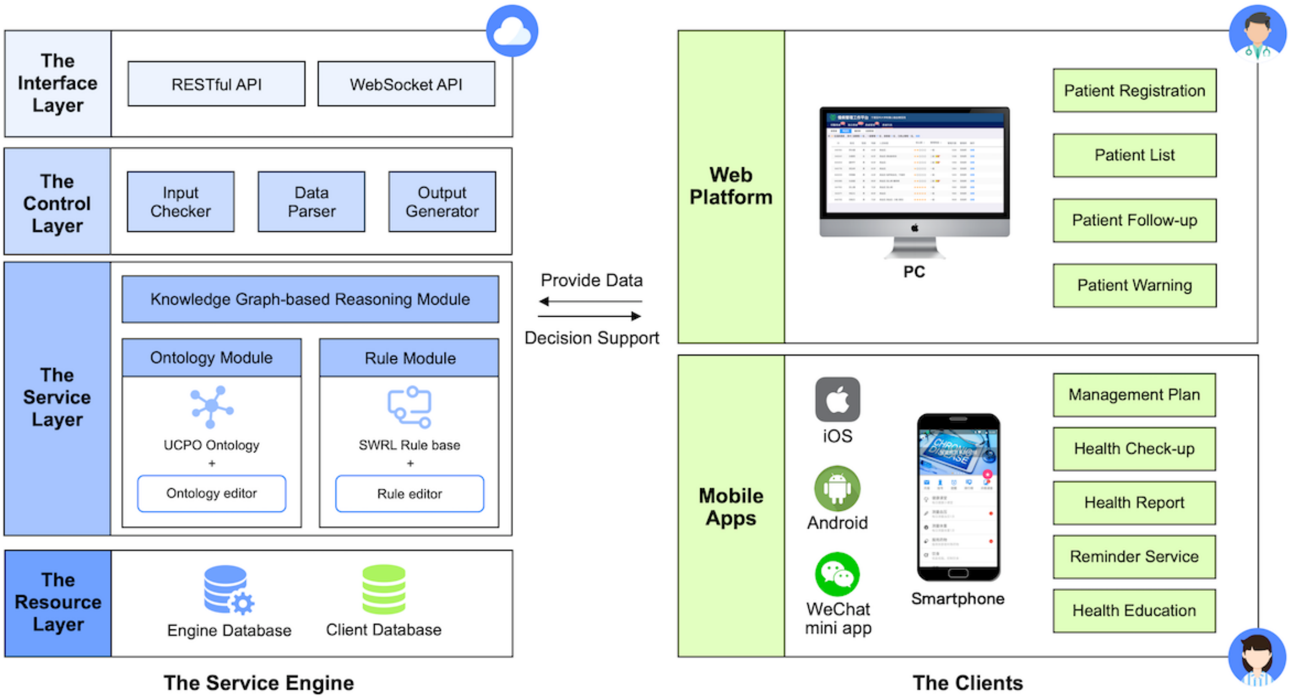
Methods

System Overview

Figure 1 illustrates the system architecture that consists of two components: (1) a service engine for data storage and decision support, and (2) clients for care providers and patients. Concretely, the service engine is a web service deployed on the cloud server, interacting with clients via several types of application programming interfaces (APIs). The core of the engine is a custom ontology called Universal Care Pathway Ontology (UCPO), which represents the knowledge contained in our pathway-driven integrated care model. Given specific patient data, the engine will generate a small-scale knowledge graph based on UCPO to provide personalized decision support.

The clients are represented in different forms for both care providers and patients. For care providers, the client is in the form of a website that can be accessed on their computers in the hospital or health center. Care providers can utilize the client to monitor patients and perform the intervention. For patients, the client is in the form of a mobile app that can be downloaded to their personal smartphones. Patients can use the client to check their self-management plans, perform self-monitoring, and receive health education. We provided three versions of our app: native apps, including Android and iOS versions, for patients who prefer a better user experience, and a WeChat mini program for patients who are more familiar with WeChat.

Figure 1. Architecture of the system. API: application programming interface; SWRL: Semantic Web Rule Language; UCPO: Universal Care Pathway Ontology.



Service Engine for Decision Support

Pathway Construction

The implementation of the service engine was divided into three steps. First, we constructed a pathway-driven integrated care model for the management of common chronic diseases catering to the “three-manager” mode in China. Figure 2 demonstrates the diagram of the model. The tailored model was constructed at an individual scale, mainly concentrating on two aspects: the roles involved in the management process and their responsibilities. According to the “three-manager” mode, four roles participate in the management: specialists, GPs, CMs, and patients. The responsibility of each role was specified through a well-designed universal care pathway. To identify common parts in the management procedures of different diseases, we carried out a qualitative analysis on the management guidelines of the three most prevalent chronic conditions in China: HTN [29], T2DM [30], and COPD [31]. A total of 9 common tasks were defined in the pathway for long-term out-of-hospital management. Furthermore, we held several rounds of discussion with experienced physicians to specify the detailed contents of each task for specific diseases that are not mentioned in the guidelines. Table 1 summarizes the general definition of each task and their specific contents (adopted in our system) for the above three diseases. The detailed description of each disease-specific care pathway can be found in Multimedia Appendix 1.

A practical guideline for effective implementation based on the tailored integrated model can be described as a two-stage process: stage 1 involves the generation of a management plan and stage 2 involves the realization of long-term effective management. In stage 1, a patient should first be diagnosed with a specific disease (or multiple diseases) by specialists in secondary or tertiary hospitals. An initial treatment plan will be formulated for the patient, mainly focusing on drug therapy. If the patient’s clinical situation is stable with no indication for hospitalization, they will be sent to the affiliated primary care clinic or health center. GPs and CMs work collaboratively to perform the out-of-hospital management. The patient needs to register in the corresponding institution (ie, patient archiving in Figure 2) and undergo a risk assessment for the diagnosed disease before starting routine management. GPs should evaluate the associated risk factors of the patient to refine the treatment plan. Subsequently, the patient will enter the initial management period, during which CMs should help the patient to become familiar with self-management tools (eg, the smartphone app). Based on the self-monitoring records during this period as well as the risk assessment results, the patient will be classified into a specific level with a specific intensity of intervention (follow-up). The management level will be dynamically adjusted throughout the management process according to the up-to-date health status of the patient. Given the personalized management plan consisting of a treatment plan and follow-up plan, the patient will enter the formal management period (stage 2).

In stage 2, the patient needs to follow self-management regimens using the provided tools. GPs need to perform follow-ups regularly to obtain a detailed understanding of the patient’s situation (combined with their self-monitoring records) and adjust the treatment plan if necessary. The follow-up schedule should be adjusted according to the management level. CMs need to supervise the patient’s self-monitoring records through a telehealth terminal (eg, the web platform for care providers in our system). Once an unexpected condition occurs, such as low compliance or abnormal self-monitoring data, CMs should respond in a timely manner to the warning, including contacting the patient and reporting the condition to GPs. If the condition is out of control, GPs should suggest a referral for the patient to receive further treatment from specialists. Moreover, CMs also need to provide health education to improve patient awareness and self-management abilities.

Figure 2. Pathway-driven integrated care model for the “three-manager” mode in China. CM: case manager; GP: general practitioner.

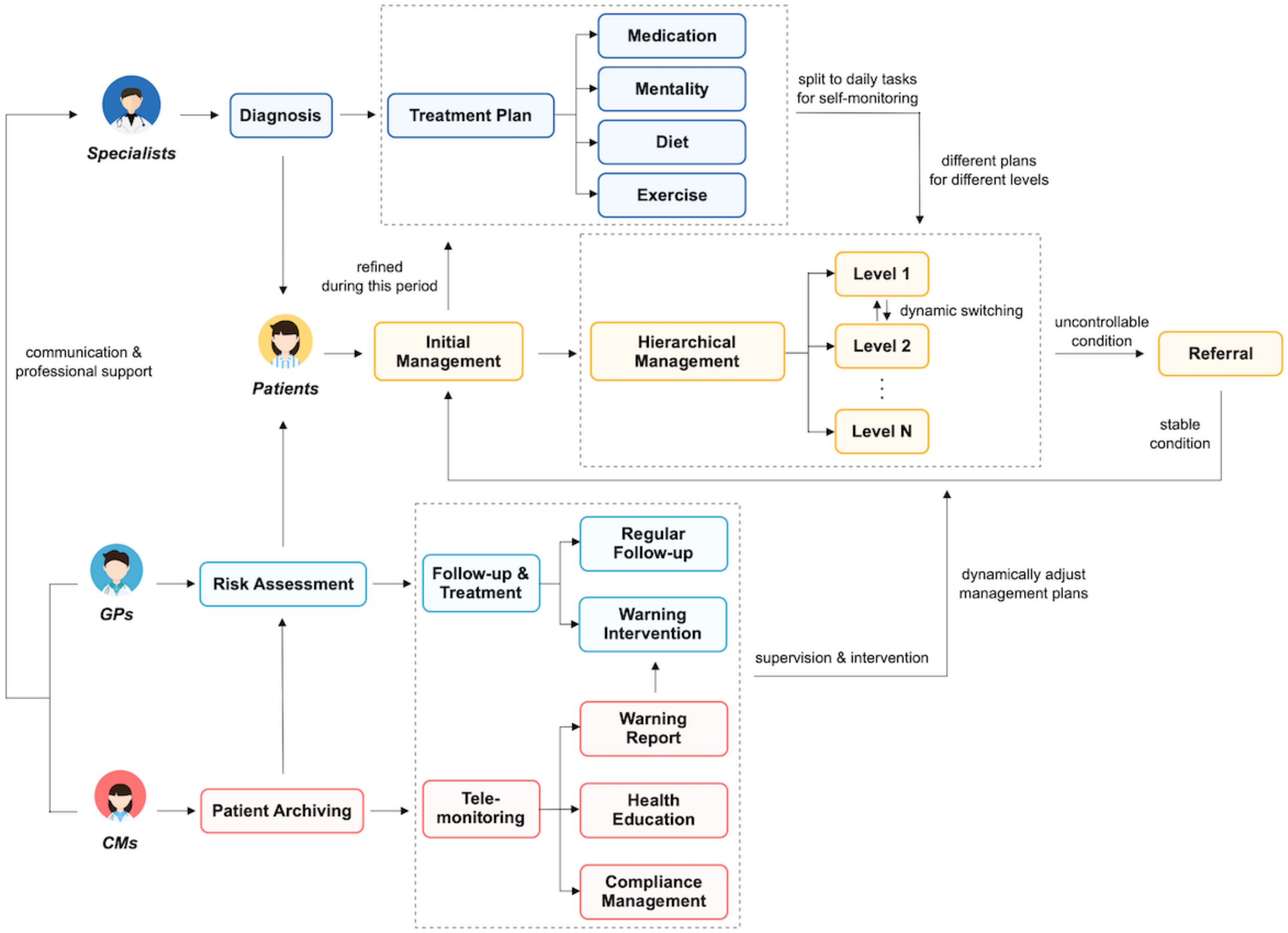


Table 1. Common tasks extracted for out-of-hospital chronic disease management.

Common tasks	General definition	Hypertension	Type 2 diabetes	COPD[a]
Diagnosis	Diagnosis based on specific indicators of specific diseases	Based on clinical BP[b] measurements	Based on clinical BG[c] measurements	Based on pulmonary function test
Risk assessment	Evaluate specific risk factors of specific diseases to refine the treatment plan	Evaluate cardiovascular risk factors, target organ damage, and comorbidity	Evaluate BP, BG, and blood lipid levels	Evaluate PEF[d] level and scales about COPD-specific health status (such as the CAT[e])
Hierarchical management	Divide patients into different levels to effectively utilize existing resources	Divide into 2 levels according to whether the patients reach target BP	Divide into 3 levels according to whether the patients reach target BG	Divide into 4 levels based on the risk assessment results
Regular follow-up	Communicate with patients regularly to perform intervention	Once every 3 months for level I and once every 2-4 weeks for level II	Once every 3 months for level I, once every month for level II, and once every 2 weeks for level III	Once every 2 weeks for each level of patients
Abnormal condition intervention	Emergency treatment for abnormal self-monitoring data	Evaluate single BP values and weekly BP values	Evaluate single BG values and blood ketone levels	Evaluate single PEF values, scale results, and acute exacerbations
Medication guidance	Drug therapy for the specific disease	Select appropriate antihypertensive drugs for patients	Select appropriate hypoglycemic drugs or insulin injection	Select appropriate drugs such as SABA[f], LAMA[g], and LABA[h]
Lifestyle guidance	Nondrug therapy for the specific disease, generally include diet, exercise, and mentality	Reduce sodium intake, control body weight, avoid smoking and drinking, increase exercise, and reduce mental stress	Control body weight, balanced diet, reduce sodium intake, avoid smoking and drinking, moderate exercise, and reduce mental stress	Avoid smoking, increase regular exercise, and perform professional rehabilitation exercises
Health education	Provide knowledge of chronic diseases to increase patients' awareness and self-management ability	Provide basic knowledge about hypertension to patients	Provide basic knowledge about diabetes to patients	Provide basic knowledge about COPD to patients
Compliance management	Enhance the motivation of patients with low self-management compliance	Perform extra follow-ups for patients with low self-management compliance	Perform extra follow-ups for patients with low self-management compliance	Perform extra follow-ups for patients with low self-management compliance

[a] COPD: chronic obstructive pulmonary disease.

[b] BP: blood pressure.

[c] BG: blood glucose.

[d] PEF: peak expiratory flow.

[e] CAT: COPD Assessment Test.

[f] SABA: short-acting beta-agonists.

[g] LAMA: long-acting muscarinic antagonists.

[h] LABA: long-acting beta-agonists.

Ontology-Based Model Representation

To incorporate the proposed model into our system, we utilized an ontological approach to implement pathway-driven decision support. A custom ontology called UCPO was constructed to represent the knowledge in our model, including structural information (ie, relationships among model elements) and medical knowledge (ie, task contents for a specific care pathway). Structural information is represented through a class hierarchy and based on the properties of ontology, whereas medical knowledge is represented through an external rule set compatible with ontology.

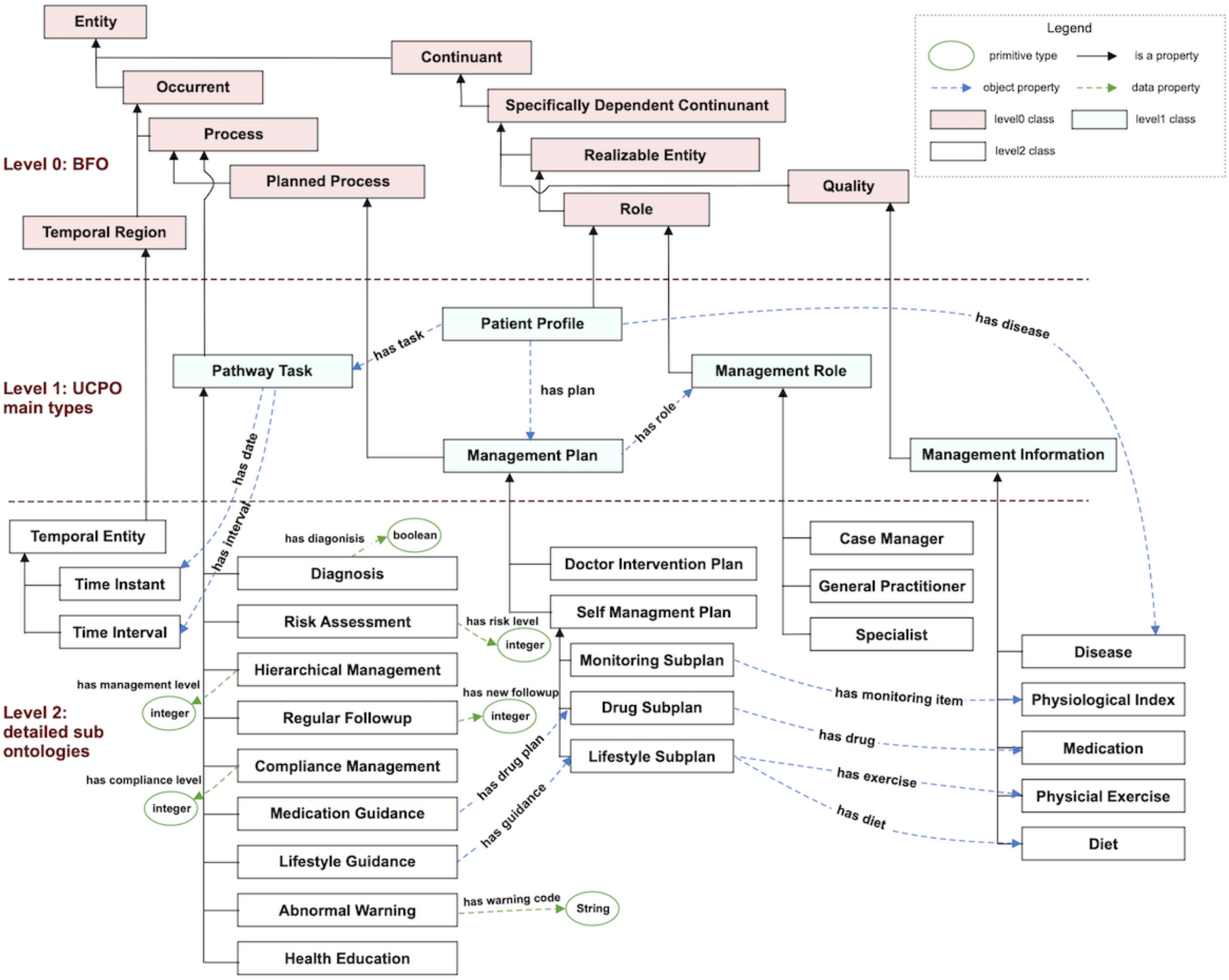
The construction of UCPO was divided into two phases. In the first phase, we represented structural information of the model following a widely used ontology engineering methodology [36]. In short, we first specified the domain and scope of UCPO using competency questions [37], and then defined the classes and class hierarchy of UCPO through a top-down approach based on existing ontologies and all terminologies contained in the model. Subsequently, we defined the properties of classes (including object properties and data properties) as well as property restrictions to describe the internal structure and precise semantics of concepts.

Figure 3 shows the class diagram and properties of the main UCPO core. UCPO was built in three levels of abstraction, inspired by a realistic ontology for diabetes treatment called DMTO [38]. Level 0 included several top-level universals from the most applicable top-level ontology (ie, basic formal ontology

[39]). For all UCPO terms, subclasses of these universals were included to improve the interoperability for future extension and integration. Level 1 includes 5 terms that describe the core concepts in our model: pathway task, management plan, management role, patient profile, and management information. Level 2 includes the detailed elements for each Level 1 class.

Classes are connected via various object properties. Several existing relevant ontologies were reused in UCPO, such as Ontology for General Medical Science [40] (for defining disease-related processes) and Ontology of Adverse Events [41] (for defining adverse events).

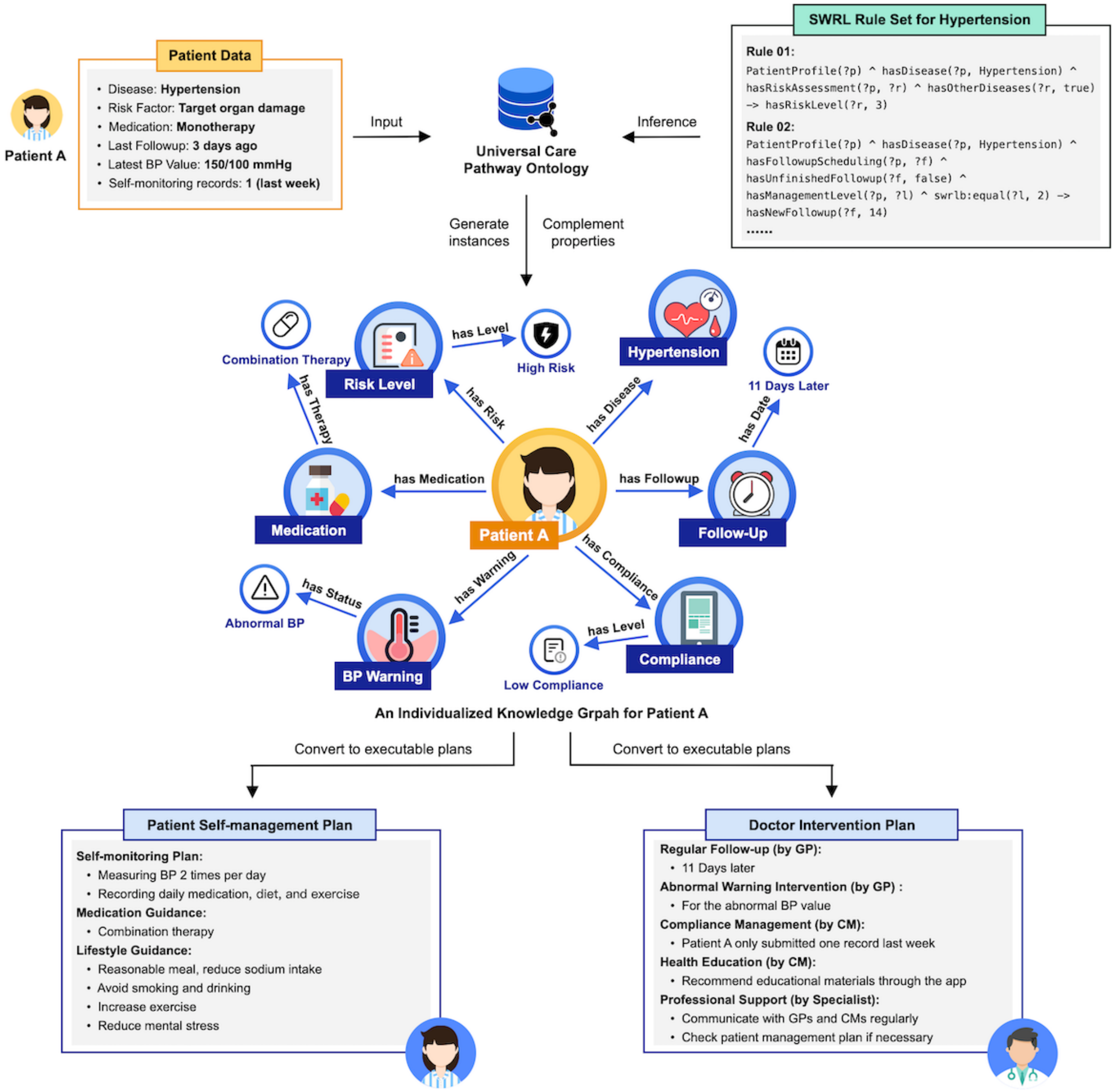
Figure 3. Class diagram of the main core of Universal Care Pathway Ontology (UCPO). BFO: basic formal ontology.



In the second phase, we incorporated medical knowledge of the model using class instances combined with rule-based reasoning. In this study, semantic web rule language (SWRL) [42] rules were utilized to perform the complex deductive inference required for decision support. The detailed description of UCPO is provided in Multimedia Appendix 2. Based on the basic UCPO and predefined SWRL rule set, a small-scale knowledge graph would be generated to incorporate patient data into UCPO for decision support. Figure 4 shows an illustrative example of the decision support process. First, patient data (Patient A in this example) are extracted from the database and then transmitted anonymously to the UCPO to generate instances of classes. Second, according to the disease type of Patient A, the corresponding rule set of the disease would be invoked to make an inference on properties of specific instances. By combining the related instances with the inference results, an individualized knowledge graph for Patient A would be established, which contains various tasks following the corresponding care pathway. Finally, the generated tasks would be converted to executable management plans, including the doctor intervention plan and patient self-management plan, via an independent rule set. The doctor intervention plan mainly involves a follow-up plan as well as intervention reminders for abnormal self-monitoring data and low compliance, whereas the patient self-management plan consists of a self-monitoring plan along with prescriptions for medication and lifestyle.

Several characteristics related to the above decision support process need to be mentioned. First, pathway tasks would be updated regularly at a task-specific frequency according to the patient data. A part of tasks will then be further assigned a valid duration defined by the Time Ontology [43]. Therefore, an incomplete subgraph might be established during a particular decision support process due to the different trigger timing of tasks. Moreover, the generation of one task might serve as a triggering condition for another task generation rule. Second, for patients diagnosed with multiple diseases, the rule set of each single disease would be executed separately. In such a case, an extra rule set for the corresponding MCC would be invoked to merge the management plans generated for different single diseases. Furthermore, the system would automatically deal with the potential redundancies and conflicts of properties in the merged management plan.

Figure 4. Illustrative example of the decision support process. BP: blood pressure; CM: case manager; GP: general practitioner; SWRL: Semantic Web Rule Language.



Patient Data

- Disease: **Hypertension**
- Risk Factor: **Target organ damage**
- Medication: **Monotherapy**
- Last Followup: **3 days ago**
- Latest BP Value: **150/100 mmHg**
- Self-monitoring records: **1 (last week)**

SWRL Rule Set for Hypertension

Rule 01:

```
PatientProfile(?p) ^ hasDisease(?p, Hypertension) ^
hasRiskAssessment(?p, ?r) ^ hasOtherDiseases(?r, true)
-> hasRiskLevel(?r, 3)
```

Rule 02:

```
PatientProfile(?p) ^ hasDisease(?p, Hypertension) ^
hasFollowupScheduling(?p, ?f) ^
hasUnfinishedFollowup(?f, false) ^
hasManagementLevel(?p, ?l) ^ swrlb:equal(?l, 2) ->
hasNewFollowup(?f, 14)
......
```

Patient Self-management Plan

Self-monitoring Plan:
- Measuring BP 2 times per day
- Recording daily medication, diet, and exercise

Medication Guidance:
- Combination therapy

Lifestyle Guidance:
- Reasonable meal, reduce sodium intake
- Avoid smoking and drinking
- Increase exercise
- Reduce mental stress

Doctor Intervention Plan

Regular Follow-up (by GP):
- 11 Days later

Abnormal Warning Intervention (by GP) :
- For the abnormal BP value

Compliance Management (by CM):
- Patient A only submitted one record last week

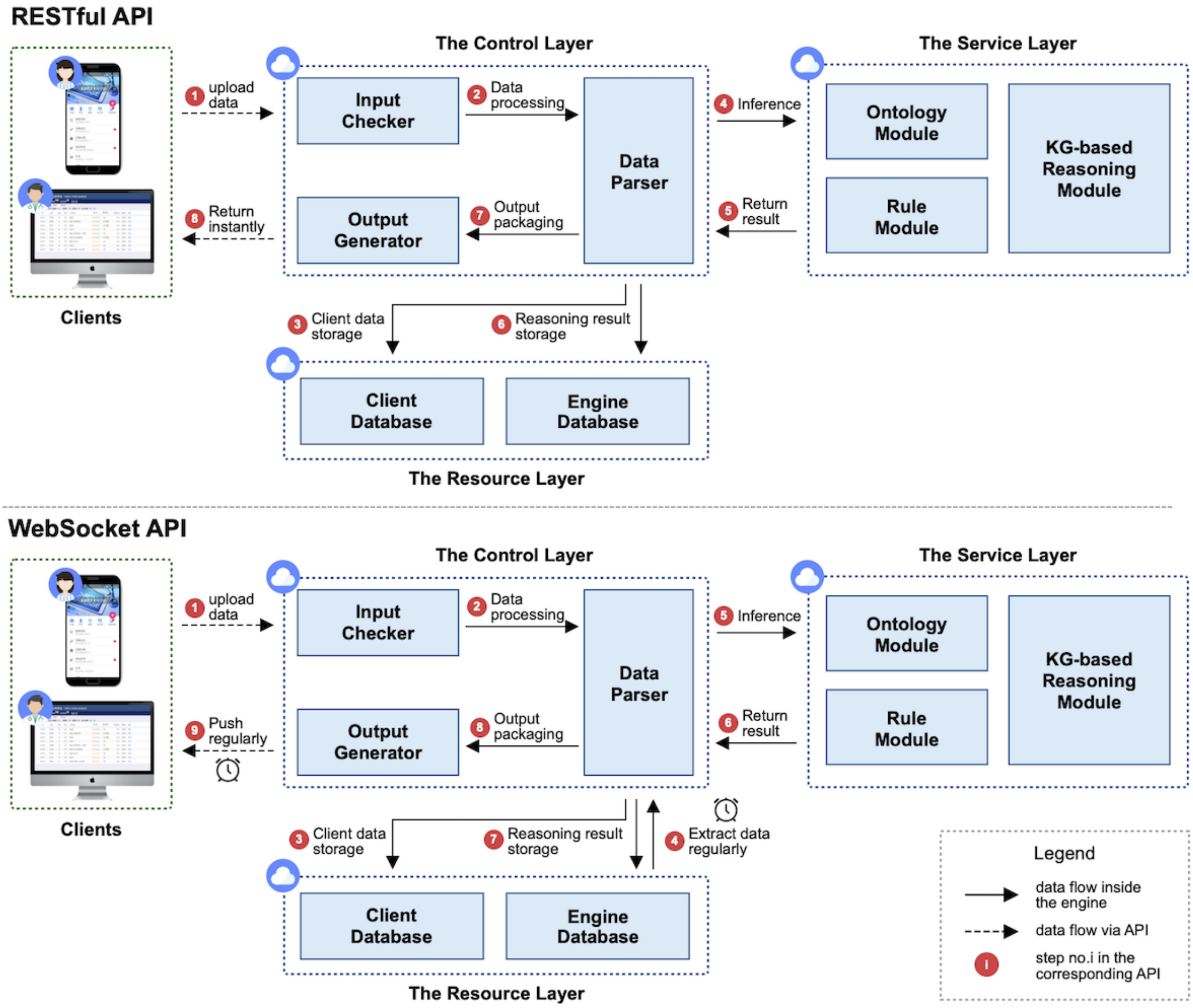
Health Education (by CM):
- Recommend educational materials through the app

Professional Support (by Specialist):
- Communicate with GPs and CMs regularly
- Check patient management plan if necessary

Engine Encapsulation

Finally, we encapsulated the engine based on UCPO to connect to the database and interact with clients. Figure 5 shows a schematic of the encapsulation. According to the different characteristics of pathway tasks, we provided two types of web services for the engine to interact with clients: WebSocket APIs and RESTful APIs. The set of WebSocket APIs deals with the scenario for timing push notifications (eg, patient stratification), whereas the set of RESTful APIs deals with the scenario for immediate feedback (eg, abnormal condition warning). Specifically, the control layer of the engine serves as a transfer station of client data, transmitting the data to the service layer and the resource layer. Client data generally include patients' self-monitoring data as well as intervention records from care providers, which would first be saved to the client database and then input into UCPO in different manners for different types of APIs. For WebSocket APIs, client data would be extracted regularly according to the task-specific frequency, whereas for RESTful APIs, client data would be directly transmitted into the ontology at the uploading time. All of the reasoning results would be saved in the engine database separately, with a portion of significant results also saved in the client database.

Figure 5. Schematic of the engine encapsulation. API: application programming interface. KG: knowledge graph.

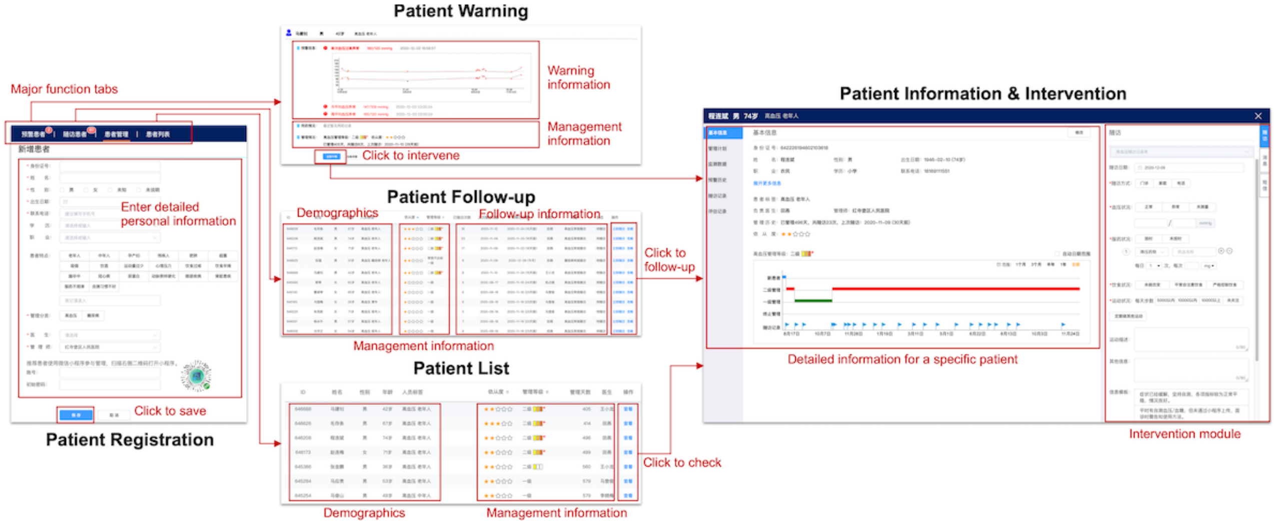


Web Platform for Care Providers

The implementation of the website for care providers followed a widely used agile development methodology [44]. We iteratively added functional modules to the platform according to the proposed pathway-driven model. Figure 6 shows an overview of the main functions in the web platform. The primary users of the website are GPs and CMs. Specialists can also log in to check the status of their patients. We provided four major functional modules via tab-based navigation: patient registration, patient warning, patient follow-up, and patient list. The patient registration module is responsible for including new patients in the management, which is mainly performed by CMs. As shown in Figure 6, patients could formally receive the management after providing several types of information, including basic information (eg, demographics, phone number, ID number), disease information (eg, main disease type, associated symptoms), and the corresponding care provider information. For patients who enter the management period, the timing of intervention is determined by the other three major functional modules: (1) the patient warning module displaying all of the untreated warnings of patients' self-monitoring data, with different types of displayed information for different types of warnings; (2) the patient follow-up module presented as a patient list in order of the next follow-up date; and (3) the patient list module, demonstrating information of patients with different diseases also in the form of a list, mainly focusing on checking compliance and searching for a specific patient.

Care providers could enter the interface of "patient information and intervention" by clicking on the corresponding buttons in the above three major functional modules. In this interface, care providers could check various types of patient information, including demographics, management plans, self-monitoring records, warning history, follow-up records, and assessment records. The platform supports three types of interventions for care providers: complete follow-up via telephone or a clinic visit, short message service text reminders, and message push via the app. Various types of templates and options are provided to simplify and normalize the intervention procedure. Moreover, care providers could edit and push educational materials to patients in another separate interface. The detailed screenshots of the web platform can be found in Multimedia Appendix 3.

Figure 6. Overview of the main functions in the web platform for care providers.

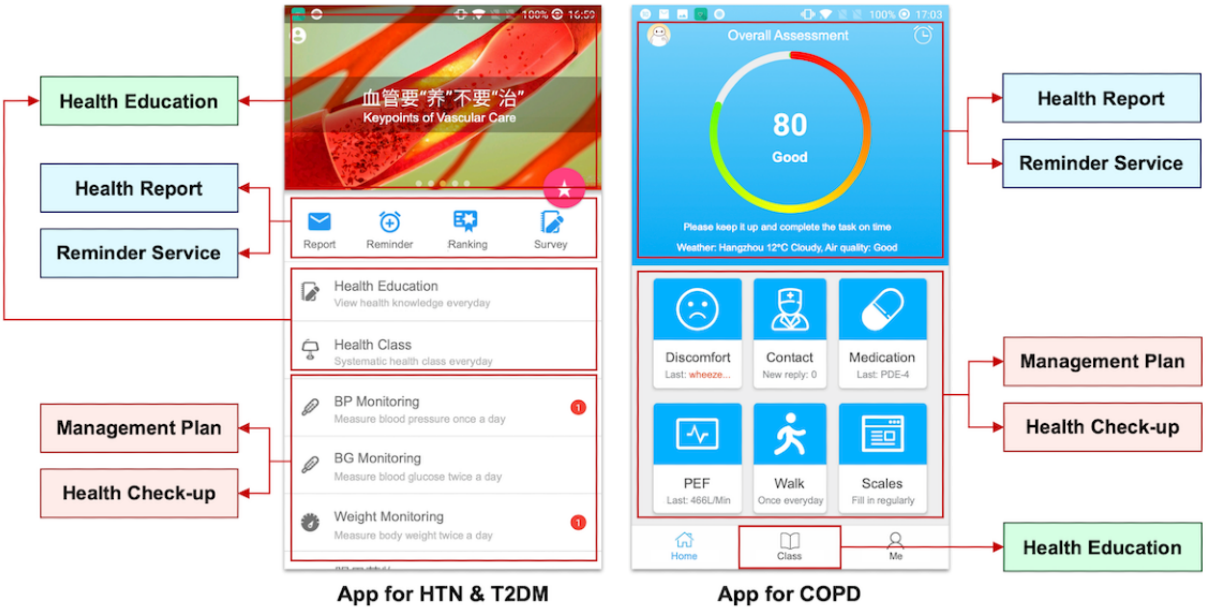


Mobile App for Patients

We utilized a goal-directed design [45] to develop the mobile app for patients. Patients were engaged in the design process to identify their needs. The concrete design process has been described in our previous study [46]; Figure 7 shows an overview of the main functions in the mobile app. Owing to the distinction between the HTN and T2DM care pathway and the COPD care pathway, we applied different user interface designs to the app for HTN/T2DM and the app for COPD. The two apps share the same underlying framework and provide similar functional modules. As shown in Figure 7, the app includes 5 major functional modules that can be accessed from the main interface: management plan, health checkup, health report, reminder service, and health education. The management plan module is the core function of the app that can be checked directly in the main interface. Currently, the initial self-management plan for patients is generated by the engine based on the management level of patients, and is then manually adjusted by care providers according to patients' specific conditions. The management plan on the patient app is shown in the form of daily tasks along with control targets. Each task is required to be accomplished at a designated time during each day. Patients could click on the corresponding task and input the required data in a new interface. The submitted data would then be uploaded to the engine for further analysis.

The other four major functional modules were designed for patients with different needs of self-management, aiming to further improve their compliance. Concretely, the health checkup module would analyze patients' self-monitoring data, and provide immediate and understandable feedback with the aid of the engine; the health report module would summarize the recent completion status of provided tasks and change trends in health data; the reminder service module would set reminders for the execution of daily tasks; and the health education module would display various types of educational materials selected by care providers. The detailed screenshots of the app can be found in Multimedia Appendix 4.

Figure 7. Overview of the main functions in the mobile app for patients. COPD: chronic obstructive pulmonary disease; HTN: hypertension; T2DM: type 2 diabetes mellitus; PEF: peak expiratory flow.



Development Tools

The service engine was developed based on the Spring Boot Framework, an open-source micro service framework for the Java platform. The connected database uses MySQL 8.0, an open-source relational database management system. The clients were developed by a team of experienced programmers collaboratively. All patients, specialists, GPs, and CMs have unique IDs, and their login passwords are encrypted and kept anonymous to the database administrator.

UCPO was constructed using the Protégé 5.5.0 open-source ontology editor in W3C Web Ontology Language (OWL) standard format (second edition). We integrated UCPO into the service engine through the OWL API, a Java API implementation for manipulating OWL ontologies. For the rule-based reasoning, the SWRL Rule Engine Bridge in the SWRL API [47] was used to invoke the execution of SWRL rules through a third-party rule engine. The official implementation currently adopts the Drools rule engine owing to its good execution speed and compatibility with Java programs.

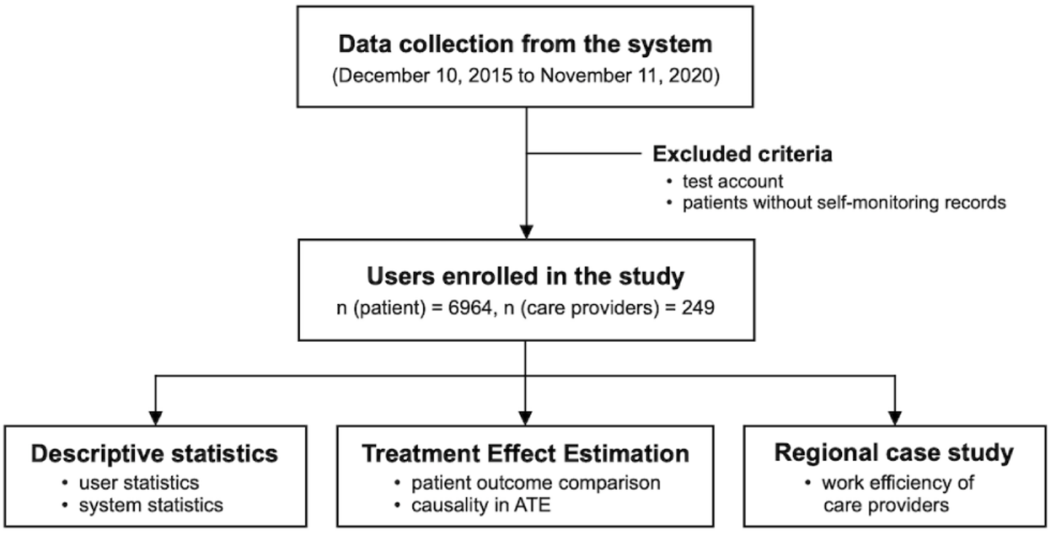
Retrospective Study

Study Design

The first version of our system was deployed in Ningxia Hui Autonomous Region in 2015, which only supported HTN management at that time. Currently, the system supports management of three single chronic diseases (HTN, T2DM, and COPD) and one type of MCC (HTN with T2DM). To investigate the effect of our system, we collected almost all of the data generated through the system since its deployment in 2015. A retrospective analysis was then performed based on the collected data to evaluate system usability, treatment effect, and quality of care. Figure 8 illustrates the overall study design. We first screened a portion of user accounts that had no self-monitoring records or were created for testing purposes. The remaining users for retrospective analysis consisted of 6964 patients with chronic diseases and 249 care providers. We then performed three types of analyses: (1) descriptive statistical analysis for user information and system usage, (2) treatment effect estimation for patient outcome changes after receiving the management, and (3) a regional case study for understanding the work efficiency of care providers.

Specifically, for the descriptive statistics, we mainly analyzed patient usage of the system and decision support abilities of the engine. For the treatment effect estimation, physiological indices from patients' self-monitoring data were selected as patient outcomes to evaluate the treatment effect. We first compared the change of patient outcomes over different time spans from a traditional statistical perspective, and then estimated the average treatment effect (ATE) from a causal perspective. For the case study, we selected several primary care institutions in different districts of Ningxia Hui Autonomous Region to evaluate the work efficiency of care providers over a long-term horizon. Owing to the limitation of data acquisition, we only analyzed the work efficiency of GPs for patients with HTN and diabetes from two aspects: the frequency of follow-ups in one day and the handling time of a follow-up request. The frequency of follow-ups in one day demonstrates how our system reduces the time cost of a single follow-up, whereas the response days of a follow-up request represents the time duration before a generated follow-up request is handled.

Figure 8. Overall retrospective study design. ATE: average treatment effect.


Data collection from the system
(December 10, 2015 to November 11, 2020)
Excluded criteria
• test account
• patients without self-monitoring records
Users enrolled in the study
n (patient) = 6964, n (care providers) = 249
Descriptive statistics
• user statistics
• system statistics
Treatment Effect Estimation
• patient outcome comparison
• causality in ATE
Regional case study
• work efficiency of care providers

Informed Consent and Ethical Considerations

Patients registered in the telehealth system have signed inform consent forms for accessing and using their personal data. The care providers signed informed consent forms as well. All procedures were performed in accordance with the ethical guidelines for biomedical research involving human subjects at Ningxia Medical University.

Data Analysis

Python 3.7 was used for data preprocessing, including data extraction from the database and descriptive analysis. Statistical analysis was performed using SPSS version 23.0. A paired Student *t* test was used for analyzing changes in patient outcomes. All statistical tests are reported at a two-sided significance level of 5%. For the causal inference between pathway-driven intervention and patient outcomes, we adopted DoWhy—an end-to-end Python library—to estimate the causal effect of our intervention [48]. In short, DoWhy follows four key steps to perform causal inference: (1) model the problem as a causal graphical model based on user-defined assumptions; (2) identify a desired causal effect estimand based on the model; (3) estimate the identified causal effect using statistical methods such as matching or stratification; and (4) verify the validity of the estimate using a variety of robustness checks. In this study, we adopted two classic causal inference methods to estimate ATE: propensity score matching (PSM) [49,50] and propensity score stratification (PSS) [51]. Both methods utilize propensity scores to achieve comparability of treatment groups and control groups in terms of their pretreatment covariates, thereby eliminating confounding bias in estimating treatment effects [52].

Results

Descriptive Statistics

User Statistics

As described above, since its deployment in 2015, a total of 6964 patients with chronic diseases and 249 care providers have registered in our system and actually used the system. Table 2 summarizes the demographics of patients and the detailed information of care providers. The average age of the patients was 58 years. Among the 6964 patients, 55.41% (n=3859) reported a relatively low educational level (high school and below), and only approximately 20% had a college degree or above; one-quarter of the patients did not provide their educational attainment at the time of registration. In terms of disease type, a substantial proportion of enrolled patients (81.7%) were diagnosed with HTN, the majority of whom had HTN alone with the remaining patients having coexisting T2DM. The other patients had a clinical diagnosis of T2DM (not with HTN) or COPD. The changing trends in the number of patients with different diseases over time are shown in Figure 9. The care providers consisted of 56 specialists, 107 GPs, and 86 CMs from different departments in different levels of hospitals.

Table 3 provides simple descriptive summary statistics of patients' self-monitoring data. Patients could submit various types of records through the mobile app, mainly including physiological indices, lifestyle records, medication, and discomfort. Physiological indices included blood pressure (BP, together with heart rates) for patients with HTN, blood glucose (BG) for patients with diabetes, and peak expiratory flow (PEF) for patients with COPD. All three indices could be measured at home via different devices [53-55]. Patients who had medication orders were required to record their medications regularly. Patients with HTN and/or T2DM were recommended to record their daily diet and exercise. For patients with COPD, psychological conditions were monitored through several validated scales [56,57]. From the statistical results, medication records and BP records were the most frequently submitted data by patients using the system. Moreover, for patients with different diseases, the emphasis of their records was also different, as shown in Table 4.

Figure 9. Changing trends in the number of patients with different diseases over time. COPD: chronic obstructive pulmonary disease; HTN: hypertension; T2DM: type 2 diabetes mellitus.

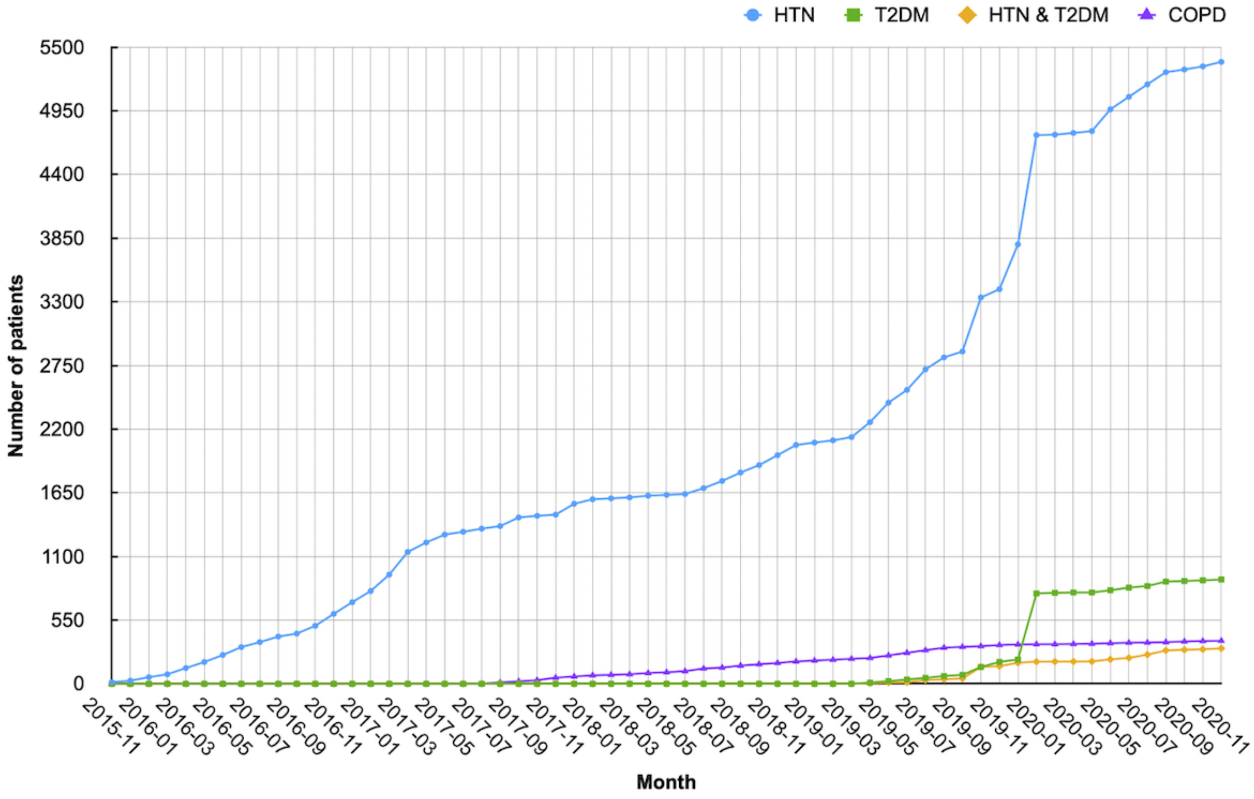


Table 2. Patient demographics and care provider information (N=6964).

Characteristics	Value
Patient demographics	
Sex, n (%)	
Male	3973 (57.1)
Female	2991 (42.9)
Age (years), mean (SD)	58 (12.3)
Educational level, n (%)	
Secondary school and below	2988 (42.9)
High school	871 (12.5)
Graduate and above	1388 (19.9)
Unknown	1717 (24.7)
Disease type, n (%)	
Hypertension (single)	5384 (77.3)
Type 2 diabetes (single)	901 (12.9)
Hypertension with type 2 diabetes	306 (4.4)
COPD[a]	373 (5.4)
Care provider information	
Position, n (%)	
Specialist	56 (22.5)
General practitioner	107 (43)
Case manager	86 (34.5)
Department, n (%)	
Cardiology	28 (11.3)
Endocrinology	15 (6.0)
Pneumology	13 (5.2)
General practice	193 (77.5)

[a] COPD: chronic obstructive pulmonary disease.

Table 3. Descriptive statistics of patients’ self-monitoring data through the system.

Data type	Patients, N	Records, N
Physiological indices		
Blood pressure	6379	139,234
Blood glucose	1195	4599
Peak expiratory flow	119	9511
Lifestyle records		
Diet	316	46,246
Exercise	1430	50,721
Psychological status	274	11,900
Medication	2260	222,055
Discomfort	493	35,332
Total counts	6964	519,598

Table 4. Self-monitoring data for patients with different diseases.

Data type	Records for patients with different diseases, n (%)			
	HTN[a]	T2DM[b]	HM[c]	COPD[d]
Physiological indices				
Blood pressure	135,512 (36.4)	2299 (42.1)	1423 (47.0)	0 (0)
Blood glucose	480 (0.1)	3003 (55)	1116 (36.8)	0 (0)
Peak expiratory flow	0 (0)	0 (0)	0 (0)	9511 (6.8)
Lifestyle records				
Diet	46,189 (12.4)	30 (0.5)	27 (0.9)	0 (0)
Exercise	39,979 (10.8)	14 (0.3)	18 (0.6)	10,710 (7.7)
Psychological status	0 (0)	0 (0)	0 (0)	11,900 (8.5)
Medication	149,291 (40.2)	102 (1.9)	441 (14.6)	72,221 (51.8)
Discomfort	335 (0.1)	8 (0.1)	5 (0.2)	34,984 (25.1)

[a] HTN: hypertension.

[b] T2DM: type 2 diabetes mellitus.

[c] HM: Hypertension with type 2 diabetes mellitus.

[d] COPD: chronic obstructive pulmonary disease.

System Statistics

Table 5 presents an overview of intervention records through the system following the four different care pathways. From the perspective of engine workflow, we classified the records in accordance with the proposed 9 common tasks (diagnosis was not involved in the system) into three categories: automatic evaluation, patient self-management support, and care provider intervention. Automatic evaluation included risk assessment and hierarchical management, which would be automatically calculated by the engine. Patient self-management support included lifestyle guidance and medication guidance, which were initially formulated by the engine and can be adjusted by care providers through the system (ie, the self-management plan). Care provider intervention included regular follow-up and abnormal condition intervention by GPs, as well as compliance management and health education by CMs. The engine would automatically schedule the follow-ups and detect the abnormal condition or low compliance, and then care providers would need to contact patients through the system to deliver the actual intervention. Health education was performed in the form of electronic materials on the patient app.

From the statistical analysis, the engine was able to generate different types of records regularly according to the specific care pathway. The content and frequency of each task were different for different diseases. For example, patients with COPD would directly be classified based on the risk evaluation results without an extra classification task. For medication guidance, we only counted the medication adjustment records generated by the engine. Moreover, medication guidance for patients with COPD have not yet been incorporated into the engine (conducted manually by care providers). Since patient compliance was updated every day by the engine, the number of records was relatively larger than that for other types of records. In terms of health education, we counted the number of articles and videos that can be viewed on the patient app [58]. Notably, several types of interventions for the COPD care pathway only involved a small number of patients due to relatively late deployment of relevant functional modules.

Table 5. Descriptive statistics of intervention records through the system following different care pathways.

Intervention type	Patients receiving intervention, N				Records generated by the engine, N			
	HTN[a]	T2DM[b]	HM[c]	COPD[d]	HTN	T2DM	HM	COPD
Automatic evaluation								
Risk assessment	3372	841	306	74	3933	912	788	4615
Hierarchical management	5383	901	306	NA[e]	301,735	6356	15,287	NA
Patient self-management support								
Lifestyle guidance	5376	901	306	30	20,459	1851	1417	1065
Medication guidance	1381	609	69	NE[f]	5200	2048	490	NE
Care provider intervention								
Regular follow-up	5322	892	304	339	18,217	2983	2064	1621
Abnormal condition intervention	1657	60	94	59	8385	218	318	2317
Compliance management	5379	900	306	30	2,099,894	274,718	175,538	6930
Health education	5384	901	306	373	199	199	199	115

[a] HTN: hypertension.

[b] T2DM: type 2 diabetes mellitus.

[c] HM: Hypertension with type 2 diabetes mellitus.

[d] COPD: chronic obstructive pulmonary disease.

[e] NA: Not applicable in the current pathway.

[f] NE: Not currently incorporated into the engine.

Treatment Effect Estimation

Patient Outcome Comparison Over Multiple Time Spans

Self-measured physiological indices submitted by patients were regarded as patient outcomes to evaluate the effect of our system. Figure 10 shows the monthly records of different patient outcomes during the most recent year. For BP, both systolic BP (SBP) and diastolic BP are presented; for BG, the system provided two options for BG self-monitoring: fasting blood glucose (FBG) and postprandial blood glucose. Compared with BP, the numbers of records for BG and PEF were relatively small due to the large proportion of patients with HTN in our system. From the trends of monthly mean value of these indices, the BP value remained basically stable at a normal level, whereas BG and PEF values fluctuated within a certain range.

We then compared the change of patient outcomes over different time spans (from 1 month to 1 year). The mean value of an outcome for a specific patient before and after a time span was calculated based on the submitted records at the first 2 weeks when the patient was enrolled and the 2 weeks after the specific time span. In terms of BP and BG values, we only analyzed SBP and FBG. From the comparison tests shown in Table 6, there were significant differences in the change of SBP values over all time spans, as well as a change of FBG values over 2 months. The change of FBG values over 1 month to 4 months showed a nonstatistically but clinically considerable decrease. No significant difference was found for the change of PEF values in these comparisons. A detailed subgroup analysis on patients with different diseases and in different age or gender groups is provided in Multimedia Appendix 5.

Figure 10. Monthly records of patient outcomes from October 2019 to October 2020. DBP: diastolic blood pressure; FBG: fasting blood glucose; PBG: plasma blood glucose; PEF: peak expiratory flow; SBP: systolic blood pressure.

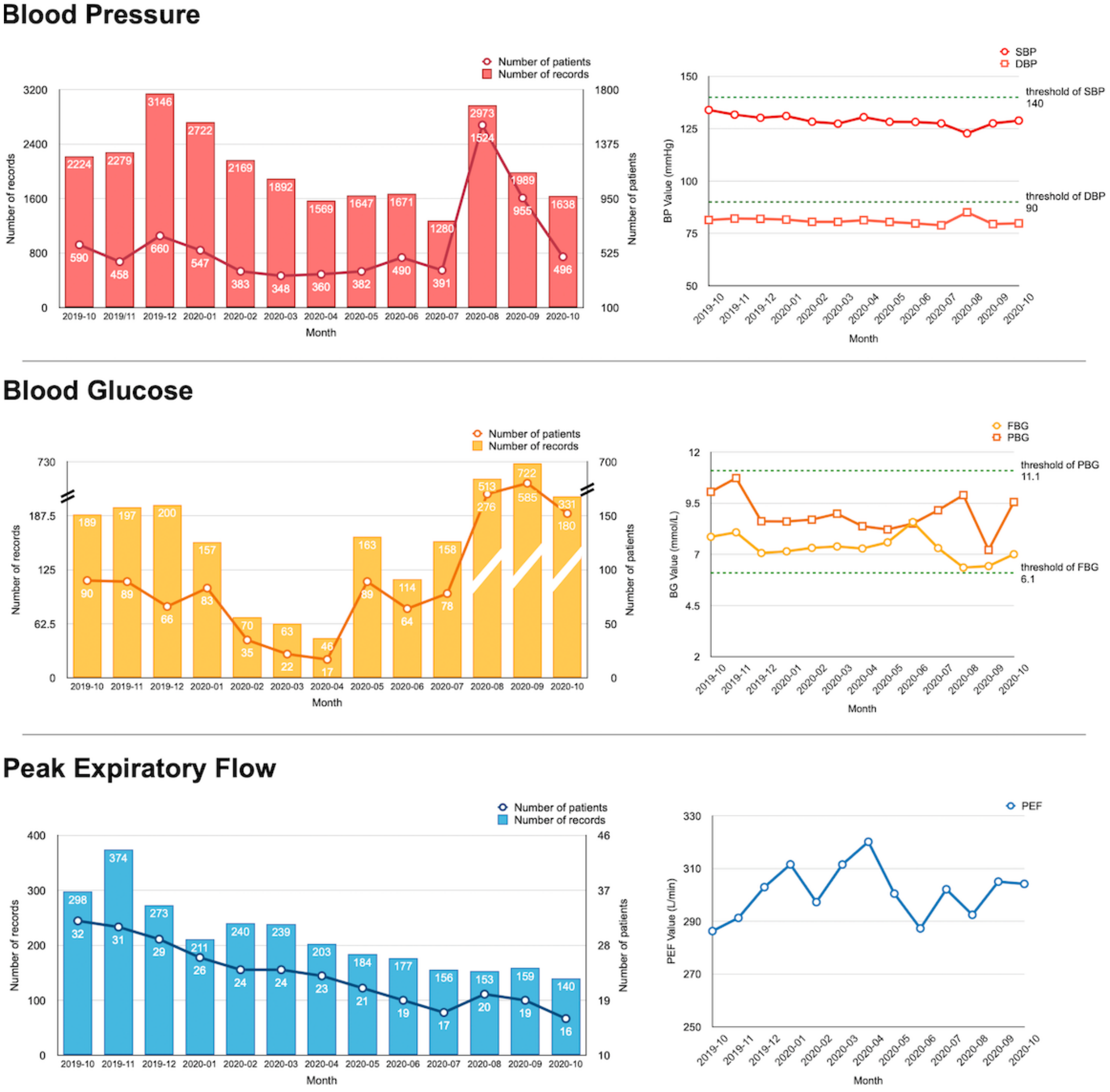


Table 6. Comparison of patient outcomes over different time spans.

Patient outcome	Patients who have records, N	Mean value before	Mean value after	*P* value
SBP[a] (mmHg)				
30 days	1140	132.13	128.43	<.001
60 days	1263	128.77	125.78	<.001
90 days	1008	130.52	128.14	<.001
120 days	725	132.56	128.86	<.001
150 days	446	130.8	127.6	<.001
180 days	403	130.94	128.18	<.001
360 days	214	130.44	128.23	.02
FBG[b] (mmol/L)				
30 days	76	8.34	6.76	.32
60 days	457	5.53	4.92	.045
90 days	58	6.74	6.58	.31
120 days	28	7.17	6.77	.24
150 days	12	6.94	7.38	.52
180 days	10	7.15	6.78	.55
360 days	10	7.08	7.11	.95
PEF[c] (L/min)				
30 days	55	315.78	320.59	.67
60 days	47	320.51	323.13	.84
90 days	44	325.77	310.69	.98
120 days	45	326.61	327.76	.95
150 days	40	316.67	309.55	.73
180 days	41	314.81	311.16	.86
360 days	29	318.3	299.04	.51

[a] SBP: systolic blood pressure.

[b] FBG: fast blood glucose.

[c] PEF: peak expiratory flow.

Causality in ATE

For the observational data, treatment effect estimation may be affected by the potential existing confounders. A confounder is a type of covariate that affects both the treatment assignment and the outcome. Spurious effect and selection bias are two main challenges brought about by confounders [59]. To estimate the true treatment effect behind our intervention, we utilized several causal inference methods to eliminate the influence of confounders. Concretely, we first constructed a causal graphical model for our problem based on prior knowledge (confirmed by physicians), as shown in Figure 11. Four confounders were considered in this study: patient age, management level, abnormal warning, and management time. The treatment variable was the intervention performed by care providers, mainly including regular follow-up and abnormal condition interventions. The outcome variables were physiological indices submitted by patients, including SBP, FBG, and PEF.

Based on the causal graph, we extracted a subdataset specifically for causal inference. For the treatment group (ie, T=True), we selected the mean value of patient outcomes within 1 month after receiving the intervention as the potential outcome (ie, Y), whereas for the control group (ie, T=False), we extracted the records of patients who did not receive any intervention within 2 weeks and regarded the mean value of self-monitoring records as the outcome. For the confounders, “management level” was the latest level of the patient at the initial time point of the record, “abnormal warning” was a Boolean variable demonstrating whether the patient has reported any abnormal condition during the corresponding period of the record, and “management time” was the time since the patient started to receive the management. The sizes of the screened dataset for the three types of patient outcomes are listed in Table 7.

Subsequently, two propensity score-based methods were adopted for evaluating the ATE, namely PSM and PSS. The estimated results are also presented in Table 7. The value of the causal estimate represents the change in the outcome value when performing the intervention (ie, if we change the treatment from "False" to "True," then the outcome value will change by the value of "estimate"). A positive value means that the outcome increases with treatment, whereas a negative value means that the outcome decreases with treatment. As expected, the values of SBP and FBG decreased significantly after receiving the intervention. The value of PEF increased after the intervention, which might be interpreted as an improvement in pulmonary function. Further, we utilized multiple refutation methods to validate the obtained estimates, which confirmed that our assumptions and results were reliable. The detailed results of refutation can be found in Multimedia Appendix 6.

Figure 11. Causal graphical model for average treatment effect.

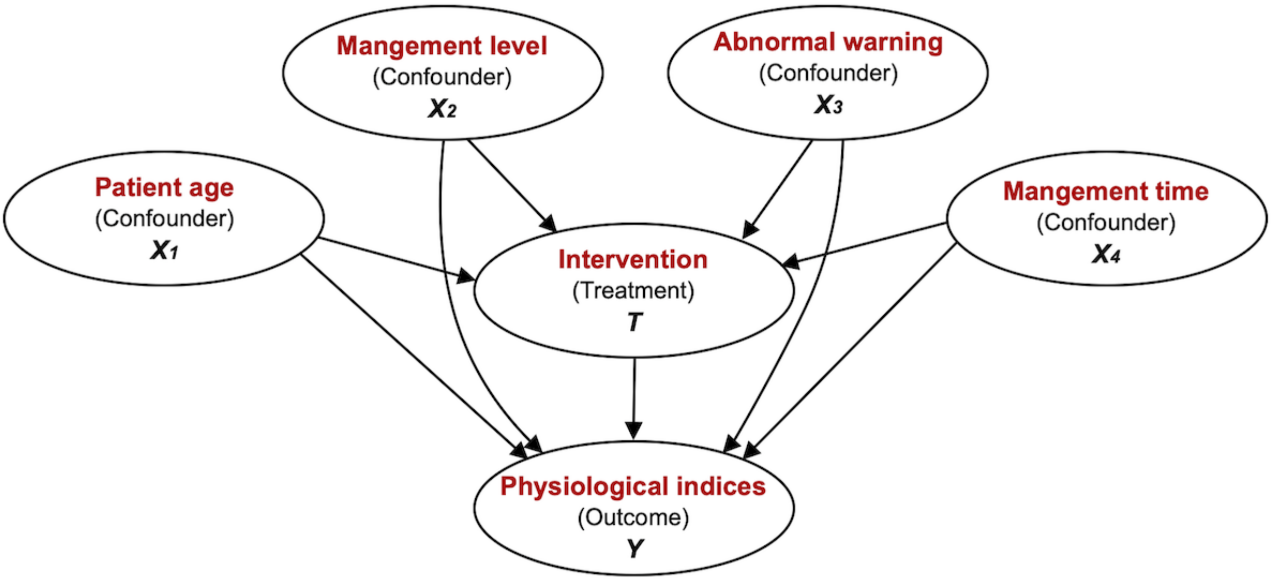


Table 7. Causal inference with propensity score matching (PSM) and propensity score stratification (PSS) for different patient outcomes.

Patient outcome	Records for inference, N	Estimated ATE[a]	
		PSM	PSS
SBP[b] (mmHg)	20,535	–5.24	–5.51
FBG[c] (mmol/L)	1765	–1.82	–1.27
PEF[d] (L/min)	1196	10.08	2.36

[a] ATE: average treatment effect.

[b] SBP: systolic blood pressure.

[c] FBG: fast blood glucose.

[d] PEF: peak expiratory flow.

Regional Case Study

We selected three primary health care facilities that registered in our system at the same time (in the first half of 2019) from different districts of Ningxia Hui Autonomous Region. All three institutions were staffed with one GP and one CM to participate in the management. Moreover, several specialists from the nearest secondary or tertiary hospitals aided with patient diagnosis and recruitment. Figure 12 shows an overview of the selected three facilities. Patients mainly comprised those with HTN, with a small proportion of patients with diabetes. Patients with COPD were not included in these three institutions.

GPs and CMs worked collaboratively to perform pathway-driven management through the system. Table 8 presents the descriptive statistics of the records of interventions performed by GPs and CMs. Further, we calculated the work efficiency of GPs based on their regular follow-up records, as shown in Figure 13. The pie chart demonstrates the percentage of different numbers of follow-up records in a day, whereas the box plot presents the distribution of response days to a follow-up request per month. From the pie charts, all three GPs performed less than 20 follow-ups in over 80% of the follow-up days, and the average number of follow-ups in a day was 16.4, 10, and 5.3, respectively. From the box plots, almost all of the medians of monthly response days were maintained within 5 days, with a couple of outliers in several months. As time progressed, the trends of response days differed for different GPs. Due to the pandemic outbreak of COVID-19 in China in early 2020 [60], for this case study, we only evaluated the intervention records in 2019 to ensure credibility of the results.

Figure 12. Overview of the selected three primary health care facilities in different districts of Ningxia. HTN: hypertension; T2DM: type 2 diabetes mellitus. HM: hypertension and diabetes.

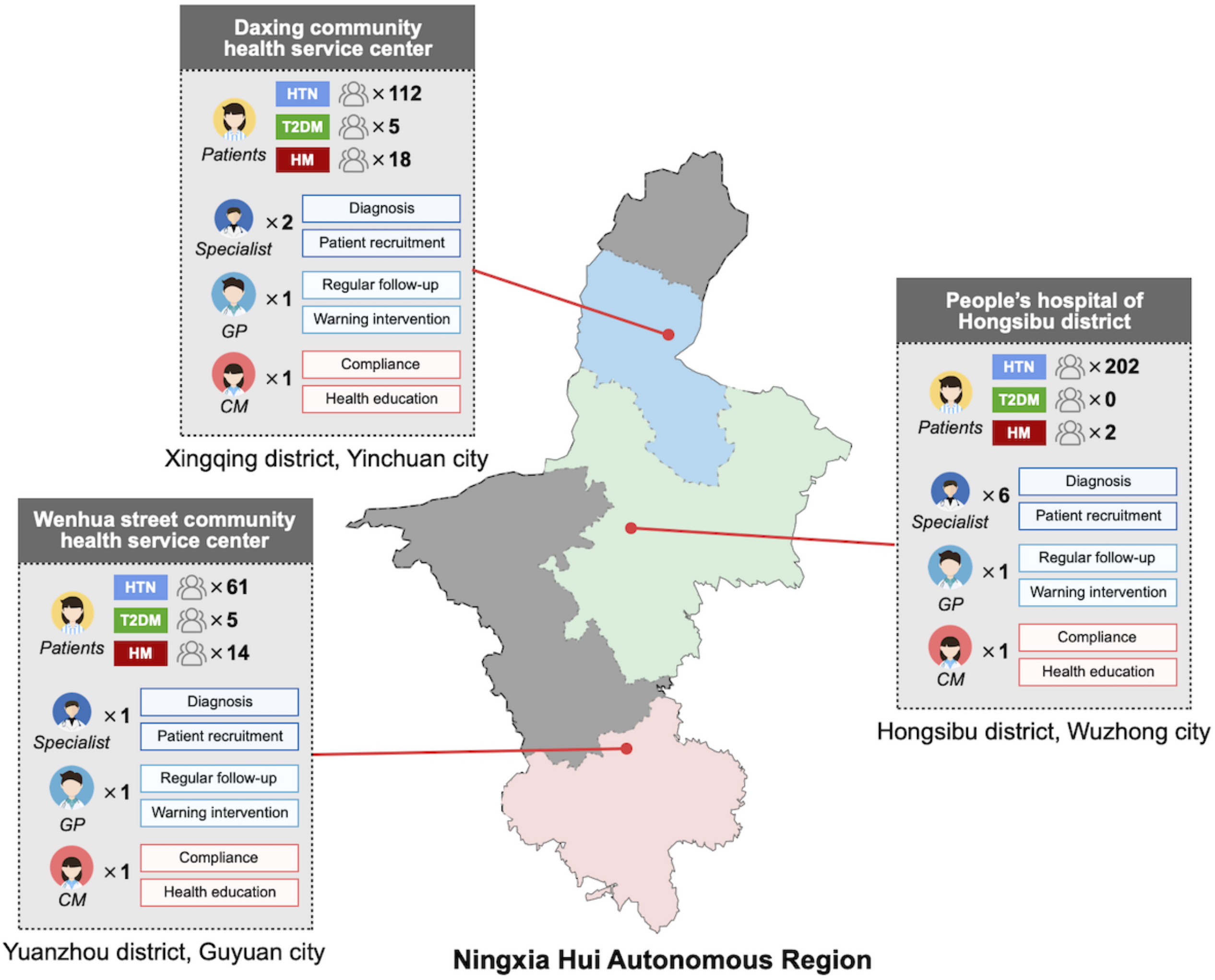


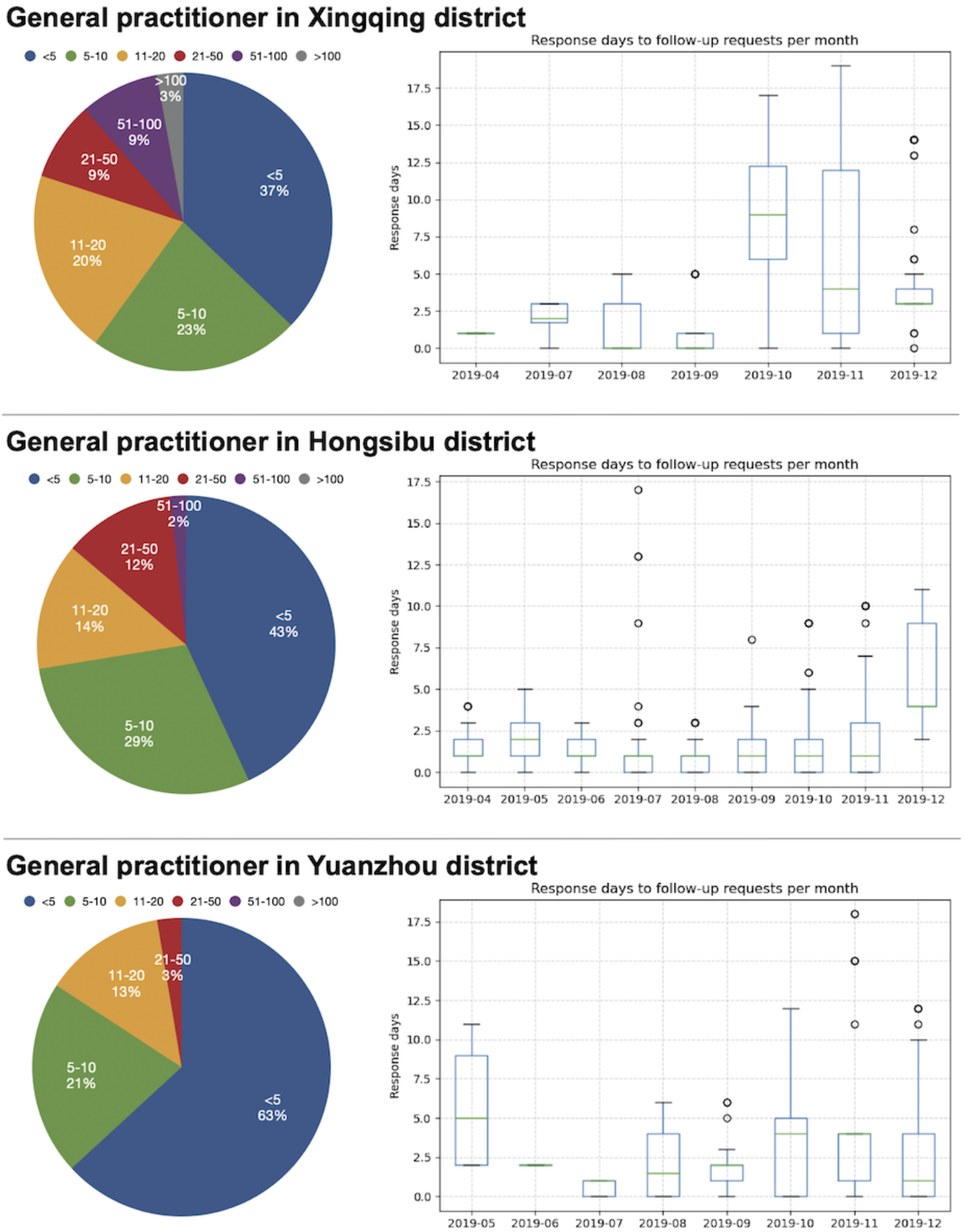
Table 8. Descriptive statistics of records of interventions performed by care providers in the selected three institutions.

Institution location	Regular follow-up (GP[a]), N	Warning intervention (GP), N	Compliance (CM[b]), N
Xingqing district			
Patients	116	114	82
Records	306	148	104
Hongsibu district			
Patients	196	48	103
Records	54	149	118
Yuanzhou district			
Patients	69	73	53
Records	203	301	86

[a] GP: general practitioner.

[b] CM: case manager.

Figure 13. Frequency of follow-ups in one day and response days to a follow-up request for the three general practitioners.



Discussion

Principal Findings

In this study, we designed and implemented a coordinated telehealth system that supports the management of multiple chronic diseases based on a tailored integrated care model for the emerging “three-manager” mode in China. The system could provide pathway-driven decision support throughout the management process via an ontology-based approach. According to the retrospective analyses on long-term system usage data, our system was able to link patients’ self-management to care providers’ interventions through semiautomatic decision support following the predefined care pathway. Furthermore, patient outcomes showed a certain degree of improvement after receiving management through the system.

Several interesting aspects can be found from the system evaluation results. First, in terms of patients’ self-monitoring data, the different emphasis of records for patients with different diseases reflects the focus of out-of-hospital management regimens on different chronic conditions. The management regimen of HTN focuses on lowering BP through medication and lifestyle changes simultaneously, whereas the management regimen of T2DM tends to focus first on lifestyle intervention instead of medication. Moreover, the management regimen of COPD requires persistent medication and close attention to acute exacerbation and mental condition of patients.

Second, in terms of comparison tests, the different levels of statistical significance among patient outcomes could be mainly attributed to the difference in cardinality among these three types of records. Compared to patients with diabetes or COPD, patients with HTN constituted the majority of the registered patients. Moreover, the self-measurement of BG at home was slightly more complicated than that for BP due to its invasiveness [61], which would potentially lower patient compliance. The self-monitoring of PEF relied on having a peak flow meter that has not been massively promoted for patients with COPD in China. Furthermore, from a clinical viewpoint, both the SBP and FBG values showed an expected decrease over most time spans, whereas the PEF value only increased over short time periods (30 days and 60 days). We considered that the relatively large SDs for the PEF value and the irreversibility of pulmonary function for patients with COPD might account for the absence of differences.

Third, in terms of causal effect estimation, the estimated ATE for SBP and FBG was consistent with the comparison tests (decreased after intervention), and the results derived by PSS and PSM were similar. The SBP value showed a relatively greater change than the FBG value under causality assumptions. By contrast, the estimated effect of PEF by the two causal inference methods showed a great degree of dissimilarity (a part of refutation tests for PEF also showed sensitivity). A possible explanation for this is that the distribution of outcomes and selected confounders among patients with COPD had a large discrepancy, which led to different intermediate results when performing stratification and matching. Moreover, certain bias might exist in the extraction strategy itself for causal data. Therefore, the true effect of our intervention for patients with COPD remains a question for further investigation. In another prospective study, we found an improvement in COPD-specific quality of life and mental health status of patients after 6-month pathway-driven management, with no significant difference in the mean PEF value [62].

Fourth, according to the regional case study, the system was able to generate different intervention tasks based on patients' health status, and then assigned them to the corresponding care providers. In this case study, the intervention tasks referred to regular follow-up and abnormal condition intervention performed by GPs, as well as compliance intervention performed by CMs. GPs and CMs were able to work with each other to provide comprehensive management support for patients. In terms of work efficiency, intuitively, the more follow-ups care providers perform in one day and the less handling time a follow-up request costs, the more effective their work will be. However, in practice, considering the workload and work arrangement of care providers, we believe that for one care provider, approximately 10 to 20 follow-ups in a single day and response within one work week to a follow-up request are reasonable, and can guarantee the quality and timeliness of follow-up. According to these criteria, the GP in Hongsibu district did a good job from both aspects, whereas the GP in Xingqing district had a relatively long duration of response to follow-up requests in the last few months of 2019. The GP in Yuanzhou district had a small average number of follow-ups in a single day, which might be attributed to the small cardinality of patients compared with the other two regions.

Comparison With Prior Work

To better delineate the contribution of this study, we compared our work with prior research from multiple aspects. In terms of model construction, the integrated care model proposed in this study can be considered as an individual-level customization of the well-known chronic care model (CCM) [63-66]. Interventions that incorporated one or more elements of the CCM have shown benefits for primary care outcomes, with large effect sizes for self-management support, delivery system design, and decision support [2]. The core of our model is the management plan combined with pathway-driven coordinated intervention, which adequately represent the above three elements. Further, the system itself is an implementation of the clinical information system in the CCM. In addition, several elements of other models can be found in our system, such as a complete eHealth-based feedback loop between patients and care providers mentioned in the eHealth Enhanced Chronic Care Model [10], effective use of health care personnel mentioned in the Innovative Care for Chronic Conditions model [67], and utilization of remote patient monitoring mentioned in the Transitional Care model [68,69].

We then compared our research with 4 prior studies that explored the comanagement of multiple chronic diseases using information technologies. The results are shown in Table 9. Among these studies, three ([16,17] and our study) utilized ontology to provide decision support abilities during the care process. Four of the studies ([16-18] and ours) supported the management of MCC through different mechanisms. Three studies ([18,19] and ours) designed an individual platform for both care providers and patients, respectively. In terms of evaluation, Riaño et al [16], Lasierra et al [17], and Laleci et al [18] only performed a technical evaluation or pilot application on their solutions, whereas Omboni et al [19] and our study deployed the system in a real-world setting for a relatively long period to test the effectiveness. In addition, due to the limitations of labor and time costs, our system currently only supports three types of chronic conditions and one type of MCC.

Table 9. Comparison of recent studies using information technologies on comanagement of multiple chronic diseases.

Study	Country	Target users	Technology for decision support	Disease types	Approach for management of MCC[a]	System implementation	Evaluation
This study	China	Patients and care providers	Ontology-based rule reasoning driven by the care pathway	3	Automatic integration via manually formulated extra rule set	Full-featured telehealth system (with mobile app)	Multidimensional retrospective study
Riaño et al [16]	Italy	Care providers	Case profile ontology combined with SDA[b] diagram	19	Semiautomatic integration of several individual plans	Wrapper system integrated into the K4CARE project	Technical evaluation and ground test involving health care professionals
Lasierra et al [17]	Spain	Patients	Ontology-driven patient profile specification	11	Manual specification of multichronic patient profiles	Semantic autonomic agent prototype	Technical evaluation without end users
Laleci et al [18]	Spain, Sweden, and United Kingdom	Patients and care providers	Decision logic encoded in GDL[c] version 2	4	Manually designed reconciled rules	C3-Cloud web platform for both MDT[d] and patients	Usability studies involving patients and clinicians
Omboni et al [19]	Italy	Patients and care providers	Analysis algorithms for generating a medical report	4	Not mentioned	Web-based telehealth platform in the context of IoMT[e] (with mobile app)	Different observational studies in various settings

[a] MCC: multiple chronic conditions.

[b] SDA: state-decision-action.

[c] GDL: Guideline Definition Language.

[d] MDT: multidisciplinary care team.

[e] IoMT: Internet of medical things.

Compared with these prior studies, our study was innovative from several aspects. To the best of our knowledge, this study is the first to construct a theoretical model for care delivery under the “three-manager” mode in China. The proposed model utilizes the concept to address the challenges of the limited ability of GPs and CMs in the primary care setting. Through refinement of a universal care pathway and specification on different chronic conditions, care providers from primary health care facilities were able to perform effective management following the practice of evidence-based medicine. Further, our model fully embodies the characteristics of coordination and a “closed loop.” Conclusively, the coordination was mainly reflected in the cooperation of different management roles and inherent associations among different pathway tasks (eg, the frequency of regular follow-up is determined by the results of hierarchical management). The “closed-loop” feature was reflected in the feedback mechanism between patients and care providers, which was implemented via dynamically adjustable management plans with the aid of information technologies.

Based on the constructed model, we implemented a telehealth system that is highly applicable for practical deployment in Chinese rural areas. The system has been carefully designed with comprehensive functions and a user-friendly interface. Care providers and patients can easily grasp the operational methods of the system after brief training. Moreover, we evaluated our system through a multidimensional retrospective study. Long-term observational data from the real world were utilized to investigate the effect of our system from several aspects, including system usability, clinical validity, and quality of care.

Strengths and Limitations

Our study has several strengths. First, in terms of system implementation, we provided two forms of mobile apps for patients: the native app and the WeChat mini program. In practical use, we found that compared with the native version, the WeChat mini program did not require installation and was easy to access from WeChat, which is one of the most frequently used apps in China. A majority of enrolled patients (especially elderly patients) tended to use the mini program, which we believe might potentially improve their compliance. Second, benefitting from the design of the universal care pathway, the system can be readily generalizable to other chronic diseases through modification of concrete task contents and definition of the corresponding rule set. The backend service and user interface also need to be updated to complete the full extension of the system. Third, we explored a new approach for evaluating patients’ long-term management effect based on causal inference methods. Although the methods and assumptions adopted in this study were preliminary, we believe that compared with simple comparison tests, the evaluation methods from a causal perspective might be more appropriate for long-term observational data directly extracted from the real world instead of clinical trials. Fourth, to evaluate the quality of care under computer-based management, we proposed a simple assessment method based on the timing and numbers of follow-ups according to our previous study [70]. The proposed method is a type of process measure for care providers from a system usage perspective. The evaluation result was relatively objective and could present a quick understanding of care providers’ work efficiency.

Several potential weaknesses of this study also need to be acknowledged. First, although the ontology-based implementation of the pathway can represent the knowledge in a shareable and elegant way, the adjustment and expansion of ontology need to be completed by knowledge engineers. Care providers encountered some degree of difficulty in understanding the logical rules contained in the ontology. Second, the number of current patients with MCC in our system is relatively small, and we only provide support for one kind of MCC (HTN with T2DM) in the current service engine. The effect of our system on a large scale of patients with diverse MCC requires further exploration. Moreover, several parts of medical knowledge in the guidelines were not integrated in the present management plan, such as proper handling methods for severe acute complications and detailed drug dosage guidance. Third, the long-term compliance remains low in patients. More effective strategies need to be considered to enhance patients' intrinsic motivation. The long-term work efficiency of care providers also needs to be improved through further medical education. Finally, the evaluation on work efficiency of care providers only considered the regular follow-up task of GPs, and the assessment method did not involve analysis on concrete intervention content.

Future Work

In future work, we will keep optimizing the usability of the system and support other common chronic conditions such as asthma, stroke, and chronic kidney disease. More elements concerned with health behavior theory (eg, behavior change technologies [71]) could be incorporated into the system to further improve patient compliance. We also plan to deploy our system in more regions of China and perform the evaluation at a larger scale. The evaluation methods will also be refined to provide more comprehensive and credible evidence, such as cost-effectiveness analysis. Another direction for future work is to explore a more personalized care pathway for a specific patient through advanced artificial intelligence technologies such as using reinforcement learning techniques to schedule the follow-up for patients and generate more precise self-management suggestions based on their self-monitoring data.

Conclusions

This study revealed the commonality in the management of different chronic diseases and explored the feasibility of integrating multiple chronic conditions into a single telehealth system. Management models could be customized for specific policy and challenges in different areas to maximize effectiveness. The tailored closed-loop care pathway proved to be feasible and effective under the "three-manager" mode in China. A part of patient outcomes improved after receiving management through the system, whereas the work efficiency of care providers differed individually. Further research might investigate the effect of such systems in a higher evidence level or introduce state-of-the-art machine learning techniques for a more individualized care pathway.

Acknowledgments

This study was supported by National Key Research and Development Program of China (2020YFC2006400), Key Research and Development Program of Zhejiang Province of China (2021C03111), Key Research and Development Program of Guangxi Zhuang Autonomous of China (2020AB33002), Fundamental Research Funds for the Central Universities (2021FZZX002-05), and Alibaba Cloud.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed description of each disease-specific care pathway.
[DOCX File , 1259 KB - medinform_v9i5e27228_app1.docx]

Multimedia Appendix 2
Detailed information of the Universal Care Pathway Ontology.
[DOCX File , 720 KB - medinform_v9i5e27228_app2.docx]

Multimedia Appendix 3
Detailed screenshots of the web platform for care providers.
[DOCX File , 1772 KB - medinform_v9i5e27228_app3.docx]

Multimedia Appendix 4
Detailed screenshots of the mobile app for patients.
[DOCX File , 1468 KB - medinform_v9i5e27228_app4.docx]

Multimedia Appendix 5

Subgroup analysis on patients with different diseases and in different age or gender for the comparison tests.
[DOCX File , 43 KB - medinform_v9i5e27228_app5.docx]

Multimedia Appendix 6
Detailed results of refutation in causal inference.
[DOCX File , 21 KB - medinform_v9i5e27228_app6.docx]

References

1. Global Status Report on Noncommunicable Diseases 2014. World Health Organization. 2014. URL: http://apps.who.int/iris/bitstream/10665/148114/1/9789241564854_eng.pdf?ua=1 [accessed 2020-12-31]
2. 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]
3. Bodenheimer T, Lorig K, Holman H, Grumbach K. Patient self-management of chronic disease in primary care. JAMA 2002 Nov 20;288(19):2469-2475. [doi: 10.1001/jama.288.19.2469] [Medline: 12435261]
4. 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]
5. Piette JD, Richardson C, Valenstein M. Addressing the needs of patients with multiple chronic illnesses: the case of diabetes and depression. Am J Manag Care 2004 Feb;10(2 Pt 2):152-162 [FREE Full text] [Medline: 15005508]
6. Starfield B, Lemke KW, Herbert R, Pavlovich WD, Anderson G. Comorbidity and the use of primary care and specialist care in the elderly. Ann Fam Med 2005;3(3):215-222 [FREE Full text] [doi: 10.1370/afm.307] [Medline: 15928224]
7. Integrated Care Models: An Overview. WHO Regional Office for Europe. 2016. URL: https://www.euro.who.int/__data/assets/pdf_file/0005/322475/Integrated-care-models-overview.pdf [accessed 2021-01-02]
8. Strengthening People-Centred Health Systems in the WHO European Region: Framework for Action on Integrated Health Services Delivery. World Health Organization Regional Office for Europe. 2016. URL: https://www.euro.who.int/__data/assets/pdf_file/0004/315787/66wd15e_FFA_IHSD_160535.pdf [accessed 2021-01-03]
9. Curry N, Ham C. Clinical and service integration: the route to improved outcomes. London: The King's Fund; 2010. URL: https://www.kingsfund.org.uk/sites/files/kf/Clinical-and-service-integration-Natasha-Curry-Chris-Ham-22-November-2010.pdf [accessed 2021-01-03]
10. Gee PM, Greenwood DA, Paterniti DA, Ward D, Miller LMS. The eHealth Enhanced Chronic Care Model: a theory derivation approach. J Med Internet Res 2015 Apr 01;17(4):e86 [FREE Full text] [doi: 10.2196/jmir.4067] [Medline: 25842005]
11. 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]
12. Quinn CC, Clough SS, Minor JM, Lender D, Okafor MC, Gruber-Baldini A. WellDoc mobile diabetes management randomized controlled trial: change in clinical and behavioral outcomes and patient and physician satisfaction. Diabetes Technol Ther 2008 Jun;10(3):160-168. [doi: 10.1089/dia.2008.0283] [Medline: 18473689]
13. Pfaeffli Dale L, Whittaker R, Jiang Y, Stewart R, Rolleston A, Maddison R. Text message and internet support for coronary heart disease self-management: results from the Text4Heart randomized controlled trial. J Med Internet Res 2015 Oct 21;17(10):e237 [FREE Full text] [doi: 10.2196/jmir.4944] [Medline: 26490012]
14. Fernandez-Granero MA, Sanchez-Morillo D, Leon-Jimenez A. Computerised analysis of telemonitored respiratory sounds for predicting acute exacerbations of COPD. Sensors (Basel) 2015 Oct 23;15(10):26978-26996 [FREE Full text] [doi: 10.3390/s151026978] [Medline: 26512667]
15. Hamine S, Gerth-Guyette E, Faulx D, Green BB, Ginsburg AS. Impact of mHealth chronic disease management on treatment adherence and patient outcomes: a systematic review. J Med Internet Res 2015 Feb 24;17(2):e52 [FREE Full text] [doi: 10.2196/jmir.3951] [Medline: 25803266]
16. Riaño D, Real F, López-Vallverdú JA, Campana F, Ercolani S, Mecocci P, et al. An ontology-based personalization of health-care knowledge to support clinical decisions for chronically ill patients. J Biomed Inform 2012 Jun;45(3):429-446 [FREE Full text] [doi: 10.1016/j.jbi.2011.12.008] [Medline: 22269224]
17. Lasierra N, Alesanco A, Guillén S, García J. A three stage ontology-driven solution to provide personalized care to chronic patients at home. J Biomed Inform 2013 Jun;46(3):516-529 [FREE Full text] [doi: 10.1016/j.jbi.2013.03.006] [Medline: 23567539]
18. Laleci Erturkmen GB, Yuksel M, Sarigul B, Arvanitis TN, Lindman P, Chen R, et al. A collaborative platform for management of chronic diseases via guideline-driven individualized care plans. Comput Struct Biotechnol J 2019;17:869-885 [FREE Full text] [doi: 10.1016/j.csbj.2019.06.003] [Medline: 31333814]
19. Omboni S, Campolo L, Panzeri E. Telehealth in chronic disease management and the role of the Internet-of-Medical-Things: the Tholomeus® experience. Expert Rev Med Devices 2020 Jul;17(7):659-670. [doi: 10.1080/17434440.2020.1782734] [Medline: 32536214]

20. Xu J, Wang W, Li Y, Zhang J, Pavlova M, Liu H, et al. Analysis of factors influencing the outpatient workload at Chinese health centres. BMC Health Serv Res 2010 Jun 05;10:151 [FREE Full text] [doi: 10.1186/1472-6963-10-151] [Medline: 20525381]
21. Zou Y, Zhang X, Hao Y, Shi L, Hu R. General practitioners versus other physicians in the quality of primary care: a cross-sectional study in Guangdong Province, China. BMC Fam Pract 2015 Oct 09;16:134 [FREE Full text] [doi: 10.1186/s12875-015-0349-z] [Medline: 26452648]
22. World Bank Group, World Health Organization, Ministry of Finance, National Family and Health Planning Commission, Ministry of Human Resources and Social Security, China Joint Study Partership. Deepening Health Reform In China: Building High-Quality And Value-Based Service Delivery. Open Knowledge World Bank. 2016. URL: https://openknowledge.worldbank.org/bitstream/handle/10986/24720/HealthReformInChina.pdf [accessed 2021-01-03]
23. Yang S. Introduction of general practitioner education and chronic disease management model in Xiamen. Chinese Gen Pract 2017;20(202):2526-2527. [doi: 10.3969/j.issn.1007-9572.2017.20.020]
24. Bourgueil Y, Marek A, Mousques J. Practice, role and position of nurses in primary care in six European countries, in Ontario and in Quebec. Rech Soins Infirm 2008 Jun(93):94-105. [Medline: 18678084]
25. van Dillen SME, Hiddink GJ. To what extent do primary care practice nurses act as case managers lifestyle counselling regarding weight management? A systematic review. BMC Fam Pract 2014 Dec 10;15:197 [FREE Full text] [doi: 10.1186/s12875-014-0197-2] [Medline: 25491594]
26. Li X, Lu J, Hu S, Cheng KK, De Maeseneer J, Meng Q, et al. The primary health-care system in China. Lancet 2017 Dec 09;390(10112):2584-2594. [doi: 10.1016/S0140-6736(17)33109-4] [Medline: 29231837]
27. Li H. Hypertension management in primary care in China: still a long way to proceed. J Gen Pract 2015;04(02):2-3. [doi: 10.4172/2329-9126.1000238]
28. Li H, Wei X, Wong MC, Yang N, Wong SY, Lao X, et al. A comparison of the quality of hypertension management in primary care between Shanghai and Shenzhen: a cohort study of 3196 patients. Medicine (Baltimore) 2015 Feb;94(5):e455. [doi: 10.1097/MD.0000000000000455] [Medline: 25654383]
29. Joint Committee for Guideline Revision. 2018 Chinese Guidelines for Prevention and Treatment of Hypertension-A report of the Revision Committee of Chinese Guidelines for Prevention and Treatment of Hypertension. J Geriatr Cardiol 2019 Mar;16(3):182-241 [FREE Full text] [doi: 10.11909/j.issn.1671-5411.2019.03.014] [Medline: 31080465]
30. Jia W, Weng J, Zhu D, Ji L, Lu J, Zhou Z, Chinese Diabetes Society. Standards of medical care for type 2 diabetes in China 2019. Diabetes Metab Res Rev 2019 Sep;35(6):e3158. [doi: 10.1002/dmrr.3158] [Medline: 30908791]
31. Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global Strategy for the Diagnosis, Management and Prevention of COPD: 2020 Report. 2020. URL: https://goldcopd.org/wp-content/uploads/2019/12/GOLD-2020-FINAL-ver1.2-03Dec19_WMV.pdf [accessed 2021-01-02]
32. Feng XL, Pang M, Beard J. Health system strengthening and hypertension awareness, treatment and control: data from the China Health and Retirement Longitudinal Study. Bull World Health Organ 2014 Jan 01;92(1):29-41 [FREE Full text] [doi: 10.2471/BLT.13.124495] [Medline: 24391298]
33. Hernandez J, Anderson S. Storied experiences of nurse practitioners managing prehypertension in primary care. J Am Acad Nurse Pract 2012 Feb;24(2):89-96. [doi: 10.1111/j.1745-7599.2011.00663.x] [Medline: 22324864]
34. Zhou XM, Wu C, Zhao L, Gao YZ, Yuan Y, Xiao XX, et al. A cross-sectional survey of the knowledge on chronic obstructive pulmonary disease in physicians of tertiary hospitals in Northern China. Zhonghua Nei Ke Za Zhi 2016 Sep 01;55(9):717-720. [doi: 10.3760/cma.j.issn.0578-1426.2016.09.012] [Medline: 27586981]
35. Wang LM, Chen ZH, Zhang M, Zhao ZP, Huang ZJ, Zhang X, et al. Study of the prevalence and disease burden of chronic disease in the elderly in China. Zhonghua Liu Xing Bing Xue Za Zhi 2019 Mar 10;40(3):277-283. [doi: 10.3760/cma.j.issn.0254-6450.2019.03.005] [Medline: 30884604]
36. Noy NF, McGuinness DL. Ontology Development 101: A Guide to Creating Your First Ontology. Standford University. 2001. URL: https://protege.stanford.edu/publications/ontology_development/ontology101.pdf [accessed 2021-01-03]
37. Grüninger M, Fox M, Gruninger M. Methodology for the design and evaluation of ontologies. 1995 Presented at: International Joint Conference on Artificial Intelligence (IJCAI95); August 20-25, 1995; Montreal, Quebec p. 1-10.
38. El-Sappagh S, Kwak D, Ali F, Kwak K. 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]
39. Arp R, Smith B, Spear A. Building Ontologies with Basic Formal Ontology. Cambridge, MA: MIT Press; 2015.
40. Scheuermann RH, Ceusters W, Smith B. Toward an ontological treatment of disease and diagnosis. Summit Transl Bioinform 2009 Mar 01;2009:116-120 [FREE Full text] [Medline: 21347182]
41. He Y, Sarntivijai S, Lin Y, Xiang Z, Guo A, Zhang S, et al. OAE: The Ontology of Adverse Events. J Biomed Semantics 2014;5:29 [FREE Full text] [doi: 10.1186/2041-1480-5-29] [Medline: 25093068]
42. Horrocks I, Patel-Schneider P, Boley H, Tabet S, Grosof B, Dean M. SWRL: A Semantic Web Rule Language Combining OWL and RuleML. W3C Member Submission. URL: https://www.w3.org/Submission/SWRL/ [accessed 2021-01-03]
43. Hobbs J, Pang F. Time ontology in OWL. W3C Candidate Recommendation. 2006. URL: https://www.w3.org/TR/owl-time/ [accessed 2021-01-03]

44. Dybå T, Dingsøyr T. Empirical studies of agile software development: A systematic review. Inf Softw Technol 2008 Aug;50(9-10):833-859. [doi: 10.1016/j.infsof.2008.01.006]
45. Cooper A, Reimann R, Cronin D, Noessel C. About Face: The Essentials of Interaction Design. Hoboken, NJ: John Wiley & Sons; 2014.
46. Duan H, Wang Z, Ji Y, Ma L, Liu F, Chi M, et al. Using goal-directed design to create a mobile health app to improve patient compliance with hypertension self-management: development and deployment. JMIR Mhealth Uhealth 2020 Feb 25;8(2):e14466 [FREE Full text] [doi: 10.2196/14466] [Medline: 32130161]
47. O'Connor M, Nyulas C, Shankar R, Das A, Musen M. The SWRLAPI: A Development Environment for Working with SWRL Rules. 2008. URL: http://webont.org/owled/2008/papers/owled2008eu_submission_41.pdf [accessed 2021-01-03]
48. Sharma A, Kiciman E. DoWhy: An End-to-End Library for Causal Inference. arxiv. 2020. URL: http://arxiv.org/abs/2011.04216 [accessed 2021-01-03]
49. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70(1):41-55. [doi: 10.1093/biomet/70.1.41]
50. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Statistician 2012 Mar 12;39(1):33-38. [doi: 10.1080/00031305.1985.10479383]
51. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc 1984 Sep;79(387):516-524. [doi: 10.1080/01621459.1984.10478078]
52. Ali MS, Prieto-Alhambra D, Lopes LC, Ramos D, Bispo N, Ichihara MY, et al. Propensity score methods in health technology assessment: principles, extended applications, and recent advances. Front Pharmacol 2019;10:973. [doi: 10.3389/fphar.2019.00973] [Medline: 31619986]
53. Stergiou GS, Kario K, Kollias A, McManus RJ, Ohkubo T, Parati G, et al. Home blood pressure monitoring in the 21st century. J Clin Hypertens (Greenwich) 2018 Jul;20(7):1116-1121. [doi: 10.1111/jch.13284] [Medline: 30003694]
54. Yoo E, Lee S. Glucose biosensors: an overview of use in clinical practice. Sensors (Basel) 2010;10(5):4558-4576 [FREE Full text] [doi: 10.3390/s100504558] [Medline: 22399892]
55. Jackson H, Hubbard R. Detecting chronic obstructive pulmonary disease using peak flow rate: cross sectional survey. BMJ 2003 Sep 20;327(7416):653-654 [FREE Full text] [doi: 10.1136/bmj.327.7416.653] [Medline: 14500437]
56. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
57. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
58. 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]
59. Yao L, Chu Z, Li S, Li Y, Gao J, Zhang A. A survey on causal inference. arxiv. 2020. URL: http://arxiv.org/abs/2002.02770 [accessed 2021-01-03]
60. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, China Novel Coronavirus Investigating Research Team. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020 Feb 20;382(8):727-733 [FREE Full text] [doi: 10.1056/NEJMoa2001017] [Medline: 31978945]
61. Villena Gonzales W, Mobashsher AT, Abbosh A. The progress of glucose monitoring-a review of invasive to minimally and non-invasive techniques, devices and sensors. Sensors (Basel) 2019 Feb 15;19(4):800 [FREE Full text] [doi: 10.3390/s19040800] [Medline: 30781431]
62. Deng N, Chen J, Liu Y, Wei S, Sheng L, Lu R, et al. Using mobile health technology to deliver a community-based closed-loop management system for chronic obstructive pulmonary disease patients in remote areas of China: Development and prospective observational study. JMIR Mhealth Uhealth 2020 Nov 25;8(11):e15978 [FREE Full text] [doi: 10.2196/15978] [Medline: 33237036]
63. Wagner EH, Austin BT, Von Korff M. Organizing care for patients with chronic illness. Milbank Q 1996;74(4):511-544. [Medline: 8941260]
64. Wagner EH, Austin BT, Davis C, Hindmarsh M, Schaefer J, Bonomi A. Improving chronic illness care: translating evidence into action. Health Aff (Millwood) 2001 Nov;20(6):64-78. [doi: 10.1377/hlthaff.20.6.64] [Medline: 11816692]
65. Bodenheimer T, Wagner EH, Grumbach K. Improving primary care for patients with chronic illness. JAMA 2002 Oct 09;288(14):1775-1779. [doi: 10.1001/jama.288.14.1775] [Medline: 12365965]
66. Bodenheimer T, Wagner EH, Grumbach K. Improving primary care for patients with chronic illness: the chronic care model, Part 2. JAMA 2002 Oct 16;288(15):1909-1914. [doi: 10.1001/jama.288.15.1909] [Medline: 12377092]
67. Innovative Care for Chronic Conditions: Building Blocks for Action. Noncommunicable Diseases and Mental Health World Health Organization. 2002. URL: https://www.who.int/chp/knowledge/publications/icccglobalreport.pdf?ua=1 [accessed 2021-01-03]
68. Hirschman KB, Shaid E, McCauley K, Pauly MV, Naylor MD. Continuity of Care: The Transitional Care Model. Online J Issues Nurs 2015 Sep 30;20(3):1 [FREE Full text] [Medline: 26882510]

69. Williams G, Akroyd K, Burke L. Evaluation of the transitional care model in chronic heart failure. Br J Nurs 2010;19(22):1402-1407. [doi: 10.12968/bjon.2010.19.22.1402] [Medline: 21139521]

70. 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]

71. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]

Abbreviations

API: application programming interface
ATE: average treatment effect
BG: blood glucose
BP: blood pressure
CCM: chronic care model
CM: case manager
COPD: chronic obstructive pulmonary disease
FBG: fasting blood glucose
GP: general practitioner
HTN: hypertension
MCC: multiple chronic conditions
OWL: W3C Web Ontology Language
PEF: peak expiratory flow
PSM: propensity score matching
PSS: propensity score stratification
SBP: systolic blood pressure
SWRL: semantic web rule language
T2DM: type 2 diabetes mellitus
UCPO: Universal Care Pathway Ontology

Edited by R Kukafka, G Eysenbach; submitted 17.01.21; peer-reviewed by LC Gradim, X Li, Z Feng; comments to author 10.02.21; revised version received 22.02.21; accepted 16.04.21; published 17.05.21.

Please cite as:
Wang Z, An J, Lin H, Zhou J, Liu F, Chen J, Duan H, Deng N
Pathway-Driven Coordinated Telehealth System for Management of Patients With Single or Multiple Chronic Diseases in China: System Development and Retrospective Study
JMIR Med Inform 2021;9(5):e27228
URL: https://medinform.jmir.org/2021/5/e27228
doi:10.2196/27228
PMID:33998999

Original Paper

Telemedicine for Follow-up Management of Patients After Liver Transplantation: Cohort Study

Min Tian[1], PhD, MD; Bo Wang[1], PhD, MD; Zhao Xue[2], PhD; Dinghui Dong[1], MD; Xuemin Liu[1], MD; Rongqian Wu[2], PhD; Liang Yu[1], MD; Junxi Xiang[1], PhD; Xiaogang Zhang[1], PhD, MD; Xufeng Zhang[1], PhD, MD; Yi Lv[1], PhD, MD

[1]Department of Hepatobiliary Surgery, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China

[2]National Local Joint Engineering Research Center for Precision Surgery & Regenerative Medicine, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China

Corresponding Author:
Yi Lv, PhD, MD
Department of Hepatobiliary Surgery
The First Affiliated Hospital of Xi'an Jiaotong University
No 277 West Yan-ta Road
Xi'an, 710061
China
Phone: 86 02985323900
Email: luyi169@126.com

Abstract

Background: Technical capabilities for performing liver transplantation have developed rapidly; however, the lack of available livers has prompted the utilization of edge donor grafts, including those donated after circulatory death, older donors, and hepatic steatosis, thereby rendering it difficult to define optimal clinical outcomes.

Objective: We aimed to investigate the efficacy of telemedicine for follow-up management after liver transplantation.

Methods: To determine the efficacy of telemedicine for follow-up after liver transplantation, we performed a clinical observation cohort study to evaluate the rate of recovery, readmission rate within 30 days after discharge, mortality, and morbidity. Patients (n=110) who underwent liver transplantation (with livers from organ donation after citizen's death) were randomly assigned to receive either telemedicine-based follow-up management for 2 weeks in addition to the usual care or usual care follow-up only. Patients in the telemedicine group were given a robot free-of-charge for 2 weeks of follow-up. Using the robot, patients interacted daily, for approximately 20 minutes, with transplant specialists who assessed respiratory rate, electrocardiogram, blood pressure, oxygen saturation, and blood glucose level; asked patients about immunosuppressant medication use, diet, sleep, gastrointestinal function, exercise, and T-tube drainage; and recommended rehabilitation exercises.

Results: No differences were detected between patients in the telemedicine group (n=52) and those in the usual care group (n=50) regarding age (*P*=.17), the model for end-stage liver disease score (MELD, *P*=.14), operation time (*P*=.51), blood loss (*P*=.07), and transfusion volume (*P*=.13). The length and expenses of the initial hospitalization (*P*=.03 and *P*=.049) were lower in the telemedicine group than they were in the usual care follow-up group. The number of patients with MELD score ≥30 before liver transplantation was greater in the usual care follow-up group than that in the telemedicine group. Furthermore, the readmission rate within 30 days after discharge was markedly lower in the telemedicine group than in the usual care follow-up group (*P*=.02). The postoperative survival rates at 12 months in the telemedicine group and the usual care follow-up group were 94.2% and 90.0% (*P*=.65), respectively. Warning signs of complications were detected early and treated in time in the telemedicine group. Furthermore, no significant difference was detected in the long-term visit cumulative survival rate between the two groups (*P*=.50).

Conclusions: Rapid recovery and markedly lower readmission rates within 30 days after discharge were evident for telemedicine follow-up management of patients post–liver transplantation, which might be due to high-efficiency in perioperative and follow-up management. Moreover, telemedicine follow-up management promotes the self-management and medication adherence, which improves patients' health-related quality of life and facilitates achieving optimal clinical outcomes in post–liver transplantation.

(JMIR Med Inform 2021;9(5):e27175) doi:10.2196/27175

KEYWORDS

liver transplantation; optimal clinical outcomes; telemedicine; telemedical follow-up management system; telerobotics; rapid recovery; health-related quality of life; remote health monitoring; telehealth; cohort; mortality; morbidity

Introduction

In 1967, Thomas Starzl performed the first successful liver transplantation [1]. Nearly half a century later, it has become a widely accepted treatment for end-stage liver disease and selected liver malignancies. Improvements in multiple dimensions, including refinement of explanting and organ preservation techniques, surgical techniques, perioperative care, and the development of potent immunosuppressive drugs have improved the outcomes of liver transplantation with 1-year survival rates >85% [2,3] and the 5-year survival rate approaching 75% [4]. The success of liver transplantation has led to an expansion of indications [5-7]; however, the lack of availability of the critical organ has prompted the use of edge donor grafts [8], such as those donated after circulatory death, from older donors, and from with hepatic steatosis [9,10]. In the past 20 years, the capabilities for liver transplantation have made remarkable progress in China. The perioperative mortality rate has been reduced to <5%, and the postoperative survival rates at 1, 5, and 10 years have reached 90%, 80%, and 70%, respectively. In 2006, the liver transplantation team led by Shu-sen Zheng at the First Affiliated Hospital, School of Medicine, Zhejiang University proposed the Hangzhou criteria [11]. Comparison of the 1-, 3-, and 5-year survival rates between Milan criteria and Hangzhou criteria groups did not reveal any statistical differences [12]. The technical capabilities for liver transplantation in China, the postoperative graft survival rate, and recipient survival rate are on par with those of the global level [13].

The increasing complexities in the liver transplantation process make it difficult to determine optimal clinical outcomes. *Textbook outcome* is an emerging concept within multiple surgical domains that defines a standardized composite quality benchmark based on multiple endpoints perioperatively, representing the ideal textbook hospitalization [14]. Although the definition of textbook outcome varies, it frequently includes the evaluation of morbidity, mortality, length of stay, and hospital readmission. Moris et al [15] defined textbook outcome as a metric of an ideal outcome in liver transplantation. The textbook outcome for liver transplantation is based on the exclusion of the following parameters: mortality within 90 days, primary allograft nonfunction, early allograft dysfunction, rejection of the graft within 30 days, readmission with 30 days, readmission to the intensive care unit during hospitalization, hospital length of stay >75th percentile of all liver transplantation, red blood cell transfusion requirement >75th percentile for all liver transplantation complications (reintervention), and major intraoperative complications. We speculate that the achievement of textbook outcome in liver transplantation is a composite metric reflecting the quality of perioperative care and cost-effective practice. Therefore, the perioperative management and follow-up system in liver transplantation are under intensive focus.

Telemedicine is the dissemination of health services over long distances by health care providers using information and communication technology [16]. eHealth is an efficient and cost-effective alternative to traditional health care that can be used to improve patients' health-related quality of life and satisfaction [17]. Telemedicine is driven by rapid developments in medicine, information, and communication technology. It has been used for many diseases (chronic obstructive pulmonary disease, asthma, heart failure) because it facilitates real-time consultation between caregivers and patients to provide timely and improved personalized care. Telemedicine also facilitates diagnosis and treatment options when medical evacuation is impossible due to acute medical emergencies, mass casualty disasters, and public health measures (such as during COVID-19 pandemic restrictions) [18,19]. From a global health perspective, telemedicine increases the availability and quality of health care in remote areas and reduces medical inequalities between remote and urban areas [20-24]. Changes in the medical field have prompted concerns—how to achieve the optimal clinical outcome (ie, textbook outcome) in liver transplantation? What is required to establish a new model to meet the challenge of the new era?

The greatest strength of telemedicine is to provide face-to-face communication in over long distances for specialized health care services, thereby eliminating the need for both the physician and patient being in the same location. We aimed to investigate the efficacy of a telemedicine follow-up management intervention after liver transplantation on recovery, hospital readmission, mortality, and morbidity.

Methods

Study Design and Participants

We conducted a clinical observation study. Between January 1, 2015 and September 30, 2018, a total of 340 patients underwent orthotopic liver transplantation in the First Affiliated Hospital of Xi'an Jiao Tong University, Shaanxi, China. The livers were donated after citizen's death. The patients were eligible for inclusion in the study if they fulfilled the discharge conditions for orthotopic liver transplantation (stable liver function and immunosuppressant blood concentration, improved diet and exercise), were willing to participate telemedicine-based follow-up management, and provided written informed consent. Patients were excluded from the study if they did not have a wireless network at home.

Patients who were enrolled in this study were randomly assigned after hospital discharge to either telemedicine-based follow-up management for 2 weeks in addition to the usual care or usual care follow-up only. All patients were followed up for 12 months, and long-term survival follow-up data were recorded until December 31, 2020.

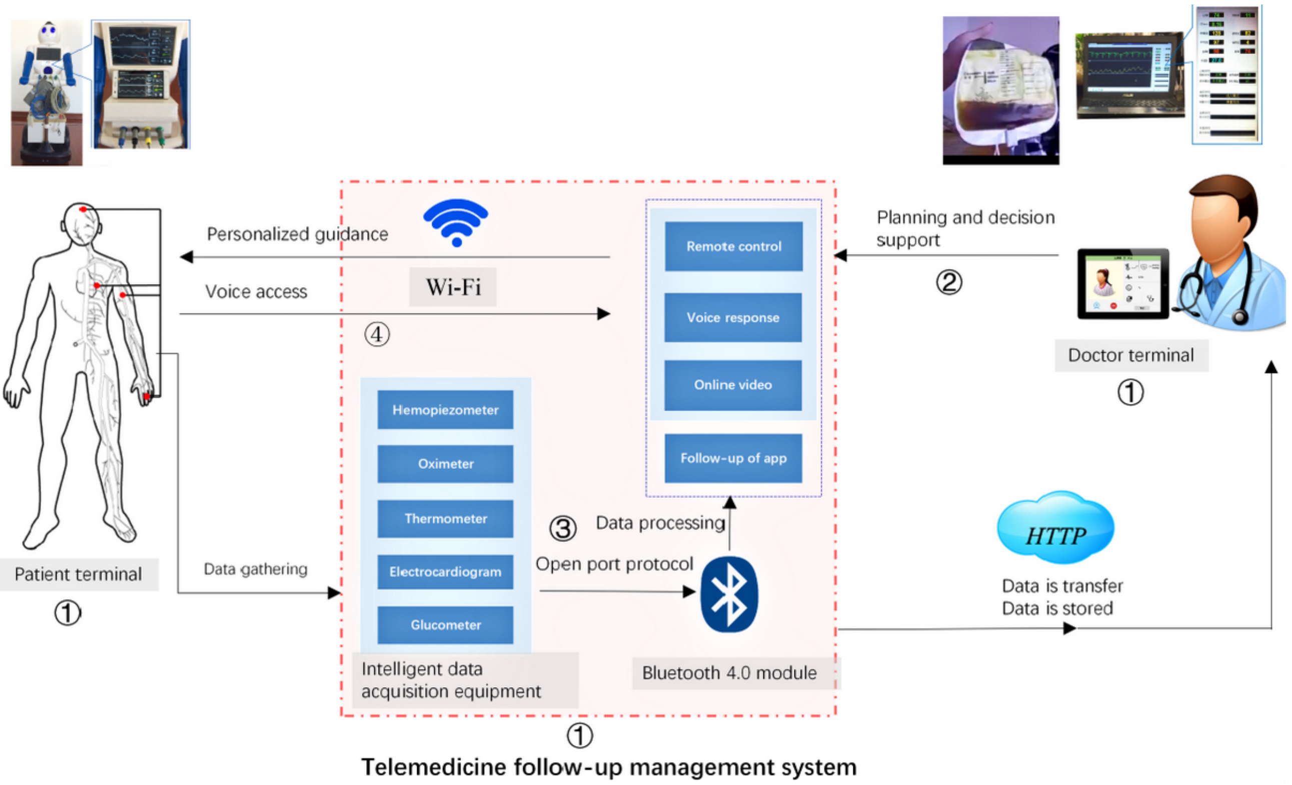
This study was approved by the First Affiliated Hospital of Xi'an Jiao Tong University Ethics Committee

(XJTU1AF2020LSK-171) and conducted in compliance with the Declaration of Helsinki and the International Code of Medical Ethics.

Patients, Health Care Professionals, and Facilities Involved in the Telemedicine Follow-up Management System

The telemedicine follow-up management system (Figure 1) included doctor terminal app, patient terminal app, and management platforms. The data acquisition equipment and intelligent service robot were utilized to acquire blood pressure, blood oxygen, temperature, and electrocardiography (ECG) data. The data transmission between the monitoring equipment and the robot used an Android Bluetooth interface. Internet-based telecommunication with health care professionals [25] used video or telephone links in real-time, and store-and-forward technology was applied [26]. The transplant specialist remotely controlled the intelligent robot face-to-face communication with liver transplantation recipients using a computer, mobile phone, or iPad. Patients' physiological parameters, such as respiratory rate, ECG, blood pressure, oxygen saturation, blood glucose level, and feedback were telemonitored via the wireless equipment [27]. The rehabilitation programs were administered after liver transplantation with home-based video conference supervised exercise, and counseling by transplant professionals.

Figure 1. Schematic of the telemedicine follow-up management system. ① The telemedicine follow-up management system includes doctor-terminal, patient-terminal, and management platform. ② The transplant specialist remotely controlled the intelligent robot "face-to-face" communication with patients by a computer, mobile phone, and tablet from anywhere, such as monitoring vital signs and T tube drainage. ③ Based on Internet-based telecommunication systems, the physiological parameters, such as respiratory rate, ECG, blood pressure, oxygen saturations processed, blood sugar or authorized by transplant specialists with feedback to the patients, were telemonitored by wireless equipment. ④ The patients could communicate with the transplant specialists about the examination results through the telemedicine follow-up management system in real-time or using store-and-forward technology.



User Training

Before initiating this study, we piloted the telemedicine follow-up management system in healthy volunteers to evaluate the feasibility of the system. Both remote transplant specialists and patients were trained to use this system. The average training time for patients was 1 hour. The acceptance of this model was based on the response to a yes or no questionnaire given to the specialists and patients, and a criterion was defined that acceptance should reach >95%.

Telemedicine Follow-up Management Intervention

Patients in this group received telemedicine follow-up management in the first 2 weeks after hospital discharge. Patients were discharged, and the telemedicine follow-up robot was given to them to take home free-of-charge.

The transplant specialists called the patients to turn on the telemedicine follow-up management robot at a specific time every morning. The transplant specialist remotely controlled the intelligent robot via face-to-face communication with liver transplantation recipients using a computer, mobile phone, or iPad. The patient used the equipment of the telemedicine follow-up robot to capture their vital signs (respiratory rate, ECG, blood pressure, oxygen saturation) and blood glucose level.

While monitoring patients' vital data, the transplant specialists inquired about the medication of the immunosuppressive agents after discharge, daily diet, sleep, relief of the bowels, exercise, and drainage of the T tube; provided guidance, and initiated rehabilitation programs for the patients. Each daily session lasted approximately 20 minutes.

The patient visited the outpatient service weekly during the 2 weeks for examination of immunosuppressant blood concentration and biochemical indexes (such as liver function) and for color doppler ultrasonography of the graft. The patients could communicate with the transplant specialists about examination results and drug adjustments through the telemedicine follow-up management system. After the end of the 2-week period, patients returned the telemedicine follow-up robot to the hospital and continued routine outpatient follow-up.

Usual Care Follow-up

The patients in the usual care follow-up group attended outpatient follow-up visits each week in the first month after hospital discharge for examination of immunosuppressant blood concentration and biochemical indexes (such as liver function) and for color doppler ultrasonography of the graft. Outpatient follow-up visits occurred every 2 weeks after the first month, then every month in the first half-year, and thereafter, every 2 to 3 months.

Statistical Analysis

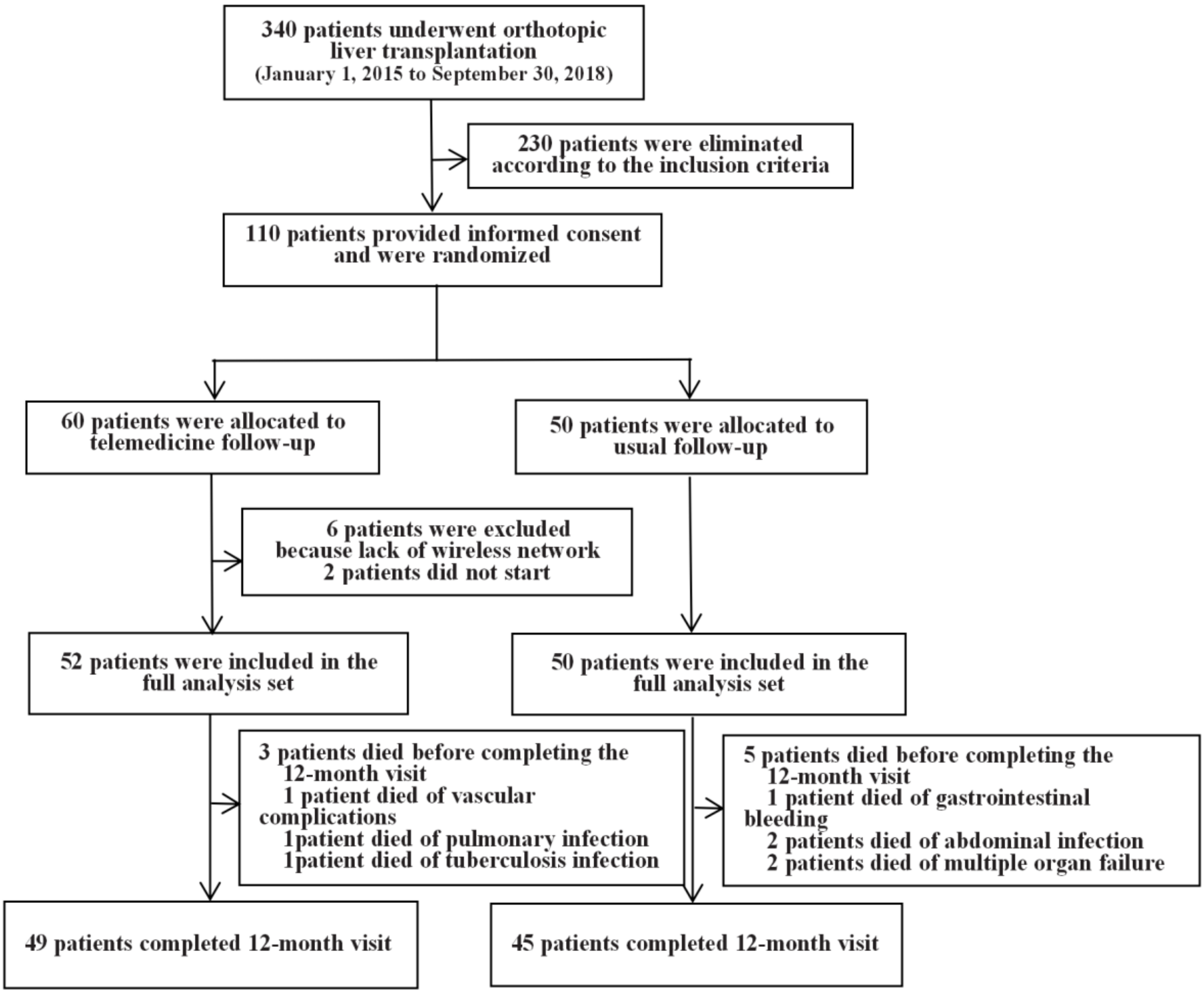
Continuous variables are reported as mean and standard deviation. Categorical variables are presented as frequency and percentages and were compared using one-way analysis of variance. Survival was evaluated using Kaplan-Meier curves. A *P* value <.05 was considered statistically significant. All statistical analyses were performed using SPSS statistical software (version 20; IBM Corp).

Results

Participants

A total of 340 patients underwent liver transplantation between January 1, 2015 and September 30, 2018; 110 patients were included in this study. A total of 60 patients were eligible for inclusion in the telemedicine group, but 6 patients were excluded from the study because they did not have a wireless network at home, 2 patients did not start the program because they could not use the telemedicine follow-up management system, and the other 52 patients were included in the full analysis set; 50 patients in the usual care follow-up group were included in the full analysis set. All patients were followed up for 12 months, and long-term follow-up data were recorded (Figure 2).

Figure 2. Procedures and participants in the telemedicine follow-up management clinical observation study.


340 patients underwent orthotopic liver transplantation (January 1, 2015 to September 30, 2018)
230 patients were eliminated according to the inclusion criteria
110 patients provided informed consent and were randomized
60 patients were allocated to telemedicine follow-up
50 patients were allocated to usual follow-up
6 patients were excluded because lack of wireless network 2 patients did not start
52 patients were included in the full analysis set
50 patients were included in the full analysis set
3 patients died before completing the 12-month visit 1 patient died of vascular complications 1patient died of pulmonary infection 1patient died of tuberculosis infection
5 patients died before completing the 12-month visit 1 patient died of gastrointestinal bleeding 2 patients died of abdominal infection 2 patients died of multiple organ failure
49 patients completed 12-month visit
45 patients completed 12-month visit

Baseline Characteristics

Patient characteristics are reported in Table 1. Of the 102 patients, the mean age was 46.65 (SD 9.66) years, and 72 (70.6%) patients were male. Of the 52 patients in the telemedicine group, the mean age was 45.35 (SD 10.44) years, and 40 (76.9%) patients were male. Of the 50 patients in the usual care follow-up group, the mean age was 48.00 (SD 8.68) years, and 32 (64.0%) patients were male. No significant differences were found for age (*P*=.17) and sex (*P*=.16) between the two groups. Malignant tumor disease before liver transplantation was observed in 20/52 (38.5%) patients in the telemedicine group and in 19/50 (38.0%) patients in the usual care follow-up group (*P*=.96). The model for end-stage liver disease (MELD) score before liver transplantation in the telemedicine group and the usual care follow-up group did not differ significantly (*P*=.14). In further analysis, 38 (73.1%) patients, 10 (19.2%) patients, and 4 (7.7%) patients in the telemedicine group and 31 (62.0%) patients, 9 (18.0%) patients, and 10 (20.0%) patients in usual care follow-up group had MELD scores <20, 20-30, and ≥30, respectively, before liver transplantation.

Table 1. Baseline characteristics.

	Total (N=102)	Telemedicine management intervention (n=52)	Usual care (n=50)	*P* value
Age (years), mean (SD)	46.65 (9.66)	45.35 (10.44)	48.00 (8.68)	.17
Sex, n (%)				.16
Male	72 (70.6)	40 (76.9)	32 (64.0)	
Female	30 (29.4)	12 (23.1)	18 (36.0)	
Diagnosis, n (%)				.96
Malignant diseases	39 (38.2)	20 (38.5)	19 (38.0)	
Benign disease	63 (61.8)	32 (61.5)	31 (62.0)	
MELD[a] score, mean (SD)	18.03 (8.78)	16.77 (7.86)	19.34 (9.55)	.14
MELD score, n (%)				.13
<20	69 (67.7)	38 (73.1)	31 (62.0)	
20-30	19 (19.6)	10 (19.2)	9 (18.0)	
≥30	14 (12.7)	4 (7.7)	10 (20.0)	
Donor age (years), mean (SD)	47.21 (14.66)	43.44 (14.51)	51.12 (13.91)	.008
Donor age (years), n (%)				.003
<18	3 (2.9)	3 (5.8)	0 (0)	
18-65	87 (85.3)	47 (90.4)	40 (80.0)	
≥65	12 (11.8)	2 (3.8)	10 (20.0)	
Orthotopic liver transplantation operation time (hours), mean (SD)	6.33 (1.00)	6.27 (1.01)	6.40 (1.00)	.51
Blood loss (mL), mean (SD)	1462.26 (1280.54)	1234.62 (945.55)	1699.00 (1528.79)	.07
Transfusion volume (mL), mean (SD)	5948.92 (1733.48)	5694.17 (1457.13)	6213.84 (1960.49)	.13
Length of initial hospitalization (days), mean(SD)	17.69 (6.56)	16.31 (3.57)	19.12 (8.45)	.03
Expense of initial hospitalization (Yuan[b]), mean (SD)	395094 (66101.04)	382502.36 (35115.42)	408190.11 (85904.13)	.049
Readmission rate within 30 days after discharge, mean (SD)	0.16 (0.37)	0.08 (0.27)	0.24 (0.43)	.02
Survival rate (%) at 12-month visit, mean (SD)	94 (92.2)	49 (94.2)	45 (90.0)	.65

[a] MELD: Model for End-Stage Liver Disease

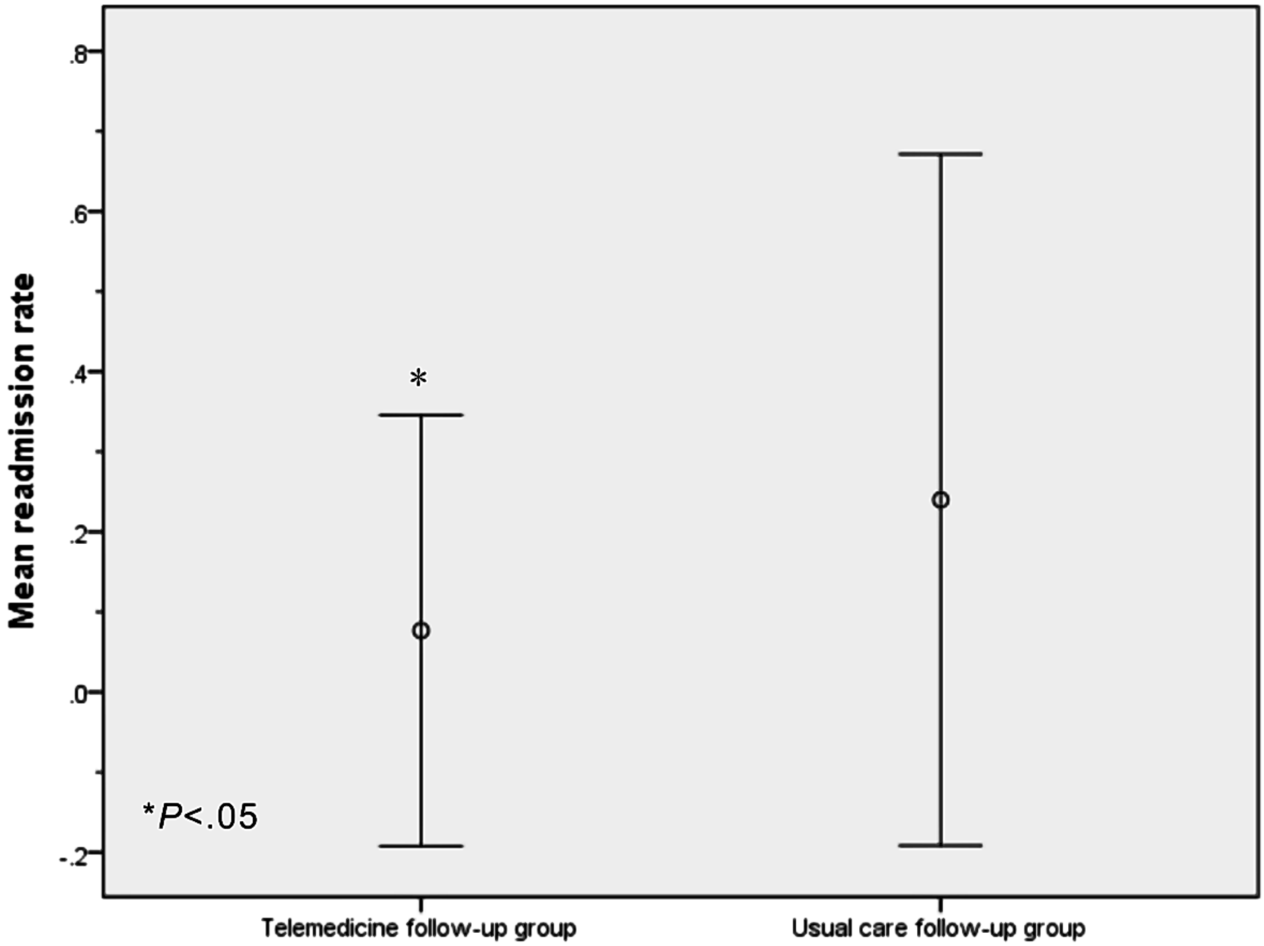
[b] An approximate exchange rate of 6.48 Yuan=US $1 was applicable at the time of publication.

Livers donation after citizen's death are currently the primary source of donors in China [28]. The donor age in the telemedicine group was lower than that in the usual care follow-up group (*P*=.008). Further analysis revealed that 2/52 (3.85%) in the telemedicine group, while 10/50 (20%) patients in the usual care follow-up group were older adult (>65 years old) donors.

Primary and Key Secondary Outcomes

No difference were found between the telemedicine and the usual care follow-up group with respect to operation time (*P*=.51), blood loss (*P*=.07), and intraoperative transfusion volume (*P*=.13); the operation quality parameters of liver transplantation in the two groups were similar. Nevertheless, statistically significant differences were found in the length of initial hospitalization (telemedicine: mean 16.31, SD 3.57; usual care: mean 19.12, SD 8.45; *P*=.03) and initial hospitalization expense (telemedicine: mean 382502.36 Yuan, SD 35115.42; usual care: mean 408190.11 Yuan SD 85904.13, an approximate exchange rate of 6.48 Yuan=US $1 was applicable at the time of publication; *P*=.049). The number of patients with MELD score ≥30 before liver transplantation was greater in the usual care follow-up group than that in telemedicine follow-up group. Furthermore, the readmission rate within 30 days after discharge was markedly lower in the telemedicine group than that in the usual care follow-up group (telemedicine: mean 0.08, SD 0.27; usual care: mean 0.24, SD 0.43; *P*=.02) (Figure 3).

Figure 3. The mean readmission rate within 30 days after discharge in the two groups. Readmission rate within 30 days after discharge in the telemedicine follow-up group was markedly lower than that in the usual care follow-up group (telemedicine: mean 0.08, SD 0.27; usual care: mean 0.24, SD 0.43; *P*=.02).



In the telemedicine group, 3 patients died before the 12-month visit (vascular complications: n=1, pulmonary infection: n=1, and tuberculosis infection: n=1); the postoperative survival rate at 12 months was 94.2%. In the usual care follow-up group, 5 patients died (portal vein thrombosis that led to gastrointestinal bleeding: n=1, severe abdominal infection: n=2, multiple organ failure: n=2); the postoperative survival rate at 12 months was 90.0% (Figure 2). There was no significant difference in the 12-month cumulative survival rate between the two groups (*P*=.65).

Major Complications After Liver Transplantation

Occurrences of significant complications, such as primary graft failure, primary graft dysfunction, acute rejection reaction, vascular complications, biliary complications, tumor recurrence, and severe infection, after liver transplantation of patients at the 12-month follow-up did not differ significantly between the two groups (Table 2).

One patient in the telemedicine group (male; 37 years old; acute-on-chronic liver failure, hepatitis B, and cirrhosis) underwent liver transplantation on August 12, 2016. He had severe postoperative complications, such as primary graft dysfunction. The patient was treated with methylprednisolone combined with multiple plasmapheresis, as well as anti-infection and liver protection. The patient recovered, was discharged after 39 days of hospitalization, and enrolled in the telemedicine group to gain guidance for postoperative rehabilitation and follow-up. At the end of the study, he was alive and healthy.

Three (6.0%) patients in the follow-up group had portal vein thrombosis, and underwent interventional thrombolysis and portal vein stents immediately; however, these were not effective, and 1 patient died of gastrointestinal bleeding. Although portal vein thrombosis did not occur in any patients in the telemedicine group, 3 patients exhibited portal vein stenosis in the telemedicine group; thus, it was recommended by the transplant specialists of the telemedicine follow-up management that these patients be readmitted; they were readmitted for portal vein angiography and portal vein stent implantation and survived.

Table 2. Major complications after liver transplantation of patients at the 12-month follow-up visit in the two groups.

Groups	All (N=102), n	Telemedicine management intervention (n=52), n (%)	Usual care (n=50), n (%)	*P* value
Primary graft failure	0	0 (0)	0 (0)	N/A[a]
Primary graft dysfunction	1	1 (1.9)	0 (0)	.33
Acute rejection reaction	7	4 (7.7)	3 (6.0)	.74
Hepatic artery thrombosis	5	3 (5.8)	2 (4.0)	.68
Portal vein thrombosis	3	0 (0)	3 (6.0)	.07
Severe biliary complications	11	4 (7.7)	7 (14.0)	.31
Tumor recurrence	7	3 (5.8)	4 (8.0)	.66
Serious infection	4	2 (3.9)	2 (4.0)	.97

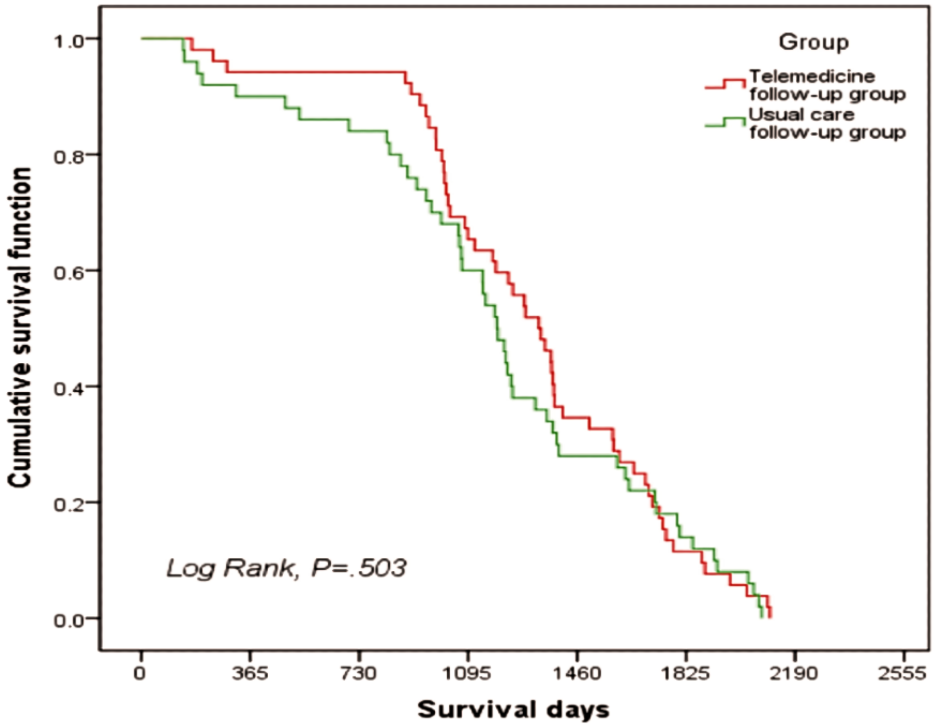
[a] N/A: not applicable.

The biliary complications were common complications of liver transplantation and required repeated endoscopic retrograde cholangiopancreatography procedures or preoperative biliary drainage: 4 (7.7%) patients in the telemedicine group and 7 (14.0%) patients in the usual care follow-up group with benign biliary stricture or bile leakage. We found that magnetic compression anastomosis was a minimally invasive method of performing choledochostomy for benign biliary stricture. One patient in the telemedicine group had benign biliary stricture, and hence, we attempted a variety of conventional treatments that failed, following which, the patient underwent preoperative biliary drainage before magnetic compression anastomosis. The device consisted of a parent and a daughter magnet. The daughter magnet was delivered via the preoperative biliary drainage route to the proximal end of the obstruction, and the parent magnet was delivered via endoscopic retrograde cholangiopancreatography to the distal end of the obstruction. After recanalization, the magnetic compression anastomosis device was removed, and biliary stenting was performed for at least 6 months with complete resolution of the condition [29,30]. Additionally, 2 patients with bile leakage detected at the telemedicine follow-up management were admitted immediately for endoscopic retrograde cholangiopancreatography and a biliary stent implanted under the guidance of transplant specialists.

Long-term Survival Analysis

Since the patients who encountered liver transplantation were followed up for life, the two groups of patients in this study were monitored continually. Long-term survival follow-up data have been recorded up until December 31, 2020. 4 patients have died in the telemedicine group, and 10 patients have died in the usual care follow-up group. The majority of these patients exhibited tumor recurrence and included other post–liver transplantation complications, such as lymphoma and cholestatic cirrhosis. None died in the perioperative period. The postoperative survival rates in the telemedicine group at 1, 2, and 3 years were 94.2%, 94.2%, and 65.4%, respectively. The postoperative survival rates in the usual care follow-up group at 1, 2, and 3 years were 90.0%, 84.0%, and 60.0%, respectively; however, no significant differences were detected between the cumulative survival curves of the two groups (*P*=.50) (Figure 4).

Figure 4. The cumulative survival curves for both groups in the long-term follow-up. No significant difference was detected in the cumulative survival rate between the two groups (*P*=.503).



Number in each group						
Telemedicine follow-up	52	49	49	34	18	6
Usual care follow-up	50	45	42	30	14	7

Discussion

Principal Findings

Rapid recovery and lower readmission rate within 30 days after discharge were evident for telemedicine follow-up management of patients after liver transplantation. Furthermore, the warning signs of complications (such as portal vein stenosis, bile leakage) were discovered earlier in the telemedicine group, and the patients received professional treatment in timely. There was no significant difference in the cumulative survival curves; however, there was a 2-year period of stability post–liver transplantation in the telemedicine follow-up group, and the cumulative survival rate was high (Figure 4). It might be associated with enhanced patient self-management and medication adherence through the telemedicine follow-up management system. Thus, the telemedicine follow-up management system could improve the patients' health-related quality of life and facilitate achieving long-term outcomes in patients.

The influencing factors for long-term survival post–liver transplantation are numerous, complicated, and frequently associated with patient-specific risk factors (age, preoperative complications, disease severity, and donor conditions). Previously, being an older adult was considered to be a contraindication for being a donor due to the increased risk of poor graft function; however, subsequent studies [31] have indicated that liver grafts from donors ≥70 years old have outcomes similar to those of younger donors. Cumulative experiences with advanced age donors report excellent outcomes in this era of organ shortage and aging population. Moreover, the study of ex vivo machine perfusion of the liver is under investigation. Improvements in donor management, organ preservation, and mitigation of ischemia and reperfusion injury hold promise in allowing safe expansion of the donor pool and improvement of outcomes in the liver transplantation [32,33]. Our study indicated that telemedicine follow-up management system is closer to achieving textbook outcomes in liver transplantation. In the modern era of rapidly developing liver transplantation capabilities, we speculate that the textbook outcome in liver transplantation is cost-effective and useful as a composite metric to reflect the quality of perioperative care. Patients with challenging perioperative courses can be helped and might experience positive long-term outcomes. The telemedicine follow-up management in liver transplantation improved the quality of perioperative care and significantly reduced the readmission rate within 30 days after discharge; therefore, post–liver transplantation medical expenses were lower.

Patients in the telemedicine group in our study were satisfied with the telemedicine follow-up management system stating that it enhanced the sense of security and medication compliance after liver transplantation. It also saved costs and time in outpatient follow-up. Furthermore, the telemedicine follow-up management system saves time for transplant specialists, optimizes the allocation of medical resources, and promotes the early and rapid recovery of patients after liver transplantation. The telemedicine follow-up management system is highly beneficial to patients with poor recovery from severe complications post–liver transplantation by helping transplant specialists to closely monitor the patients' condition after discharge and guide recovery.

Although no significant difference was detected in the diagnosis and treatment of postoperative complications between the telemedicine group and the usual care follow-up group, a large number of patients in the telemedicine group showed improved self-management and medication adherence. Additionally, early warning signs of complications were detected, and the patients received timely professional treatment. For example, a change was detected in the drainage fluid through remote video follow-up, the warning signs of portal vein stenosis were detected early in the telemedicine group, and the patients received professional treatment in a timely manner and improved the quality of life. Portal vein stenosis occurs in approximately 3% of liver transplantations but occurs in approximately 3.4% to 14% of split liver transplantations; early detection and treatment are essential for long-term graft survival [34,35]. Recently, some studies [36-38] highlighted the key role of interventional radiology in treating the stenosis safely and successfully with balloon angioplasty with stenting. In addition, patients could actively learn self-management and healthy exercise after liver transplantation. Robust physical activity after liver transplantation is a critical determinant of long-term health, similar that of pretransplant activity, for withstanding the immediate stress of transplantation [39].

Digital technology currently plays a major role in various fields. Digitization in medicine has been implemented for remote health monitoring, visual interactions between patient and doctor, and visual interactions between doctors from different hospitals and countries [40-42]. Increasing attention has been focused on the sustainability of health care systems; telemedicine allows health care providers to remotely diagnose and treat patients using telecommunications as either an alternative to or along with clinical visits [43,44]. Self-management support is one of the mechanisms by which telemedicine interventions have been proposed to facilitate the management of long-term conditions. In the last decade, telemedicine supported self-management of heart failure, asthma, chronic obstructive pulmonary disease, and cancer [45]. The most prominent examples within telehealth are related to pulmonary care: telemedicine with diagnosis at a distance based on spirometry tracing, teleconsultation, telemonitoring of biological signals, decision support systems, telecare, telerehabilitation, and second-opinion calls [16]. While telemedicine-mediated self-management was not consistently superior to that of usual care in several studies [45], none of the reviews reported negative effects, suggesting that it is a safe option for the delivery of self-management support. The key to optimizing the use of telemedicine is to correctly identify the ideal candidates, durations, and time points for a specific need [46].

In our study, the telemedicine follow-up management system was customized for patient post–liver transplantation, and the intervention administered for a short time after hospital discharge, which has not previously been done. We also emphasized the interaction between patient and transplant specialists, and rehabilitation guidance was provided according to the individual's recovery early post–liver transplantation. The increasing number of patients requiring organ transplants, the complex landscape of liver transplantation, long distances, and poor road infrastructure between doctors and patients create barriers for the delivery of health care services, especially rural regions, some of which can be addressed by telemedicine. The telemedicine follow-up management system for liver transplantation promoted innovative treatment by accelerated exchange of patient data, and faster patient recovery is beneficial to both doctors and patients.

The development of telemedicine has some limitations. The most relevant factors in assessing the quality of telemedicine management are correct imaging, correct medical history, and the clinical skills of the physician. A 92% to 98% diagnostic conformity was detected between telemedicine assessment and a face-to-face clinical assessment in a prospective pilot study [47]. Second, the misuse of personal data and information from patients' medical documents is a significant issue. Unfair access to such personal and confidential information can be potentially dangerous [41]. Therefore, it is necessary to strengthen digital information security and formulate a relevant management system.

Limitations

This study has several limitations. The follow-up intervention duration was only 2 weeks, and the number of patients was small. The generalizability of our results requires verification. Additionally, we could not determine whether telemedicine follow-up management differed between younger and older patients. However, our telemedicine follow-up management was customized in post–liver transplantation with emphasis on the interaction between patient and transplant specialists. In order to promote and apply to other fields, additional specific components of follow-up are essential. The telemedicine follow-up robot was inconvenient to carry; hence, a wireless network is required; however, some patients may not have access to a wireless network to be able to implement the program. Therefore, further improvement is required (for example, using 5G networks) to make it flexible and convenient.

Conclusion

We demonstrated that rapid recovery and low readmission rate within 30 days after discharge were evident for telemedicine follow-up management of patients in the early stage, post–liver transplantation, which might be due to more efficient perioperative follow-up management. Furthermore, warning signs of complications were discovered early in the telemedicine group, and the patients received professional and timely treatment. The survival rate of patients in the telemedicine follow-up group was high in the first 2 years post–liver transplantation, which could be attributed to better patient self-management and medication adherence through the telemedicine follow-up management system. The telemedical management system is crucial in improving the patients' health-related quality of life and achieving long-term outcomes in patients. Therefore, the intervention of the telemedicine follow-up management system is beneficial to achieving optimal clinical outcomes in liver transplantation.

Acknowledgments

This work was supported by the basic scientific research operating expenses of central universities (XJJ2018271) and Military Logistics Research Project (BZZ19J004).

The authors acknowledge the engineers of Xi’ an Langang Science and Technology Co, who contributed to developing and debugging the telemedicine follow-up management system. The authors are also grateful to the doctors and nurses of the First Affiliated Hospital of Xi’an Jiao Tong University who took part in the study.

Authors' Contributions

YL participated in research design and contributed to developing and debugging the telemedicine follow-up management system. MT participated in research design, participated in writing the paper, finished the study, collected data, and analyzed data. RW, ZX, and DD participated in developing the telemedicine follow-up management system, collected data, and analyzed data. LY, BW, XL, and XZ participated in the developing and debugging telemedicine follow-up management system. XZ and JX revised the manuscript. All authors approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Starzl TE, Demetris AJ, Van Thiel D. Liver transplantation (1). N Engl J Med 1989 Oct 12;321(15):1014-1022 [FREE Full text] [doi: 10.1056/NEJM198910123211505] [Medline: 2674716]
2. Adam R, Karam V, Delvart V, O'Grady J, Mirza D, Klempnauer J, All contributing centers (www.eltr.org), European Liver and Intestine Transplant Association (ELITA). Evolution of indications and results of liver transplantation in Europe. a report from the European Liver Transplant registry (ELTR). J Hepatol 2012 Sep;57(3):675-688 [FREE Full text] [doi: 10.1016/j.jhep.2012.04.015] [Medline: 22609307]
3. Agopian V, Petrowsky H, Kaldas F, Zarrinpar A, Farmer D, Yersiz H, et al. The evolution of liver transplantation during 3 decades: analysis of 5347 consecutive liver transplants at a single center. Ann Surg 2013 Sep;258(3):409-421. [doi: 10.1097/SLA.0b013e3182a15db4] [Medline: 24022434]
4. Kim WR, Lake JR, Smith JM, Schladt DP, Skeans MA, Harper AM, et al. OPTN/SRTR 2016 annual data report: liver. Am J Transplant 2018 Jan;18 Suppl 1:172-253 [FREE Full text] [doi: 10.1111/ajt.14559] [Medline: 29292603]
5. Moris D, Tsilimigras DI, Ntanasis-Stathopoulos I, Beal EW, Felekouras E, Vernadakis S, et al. Liver transplantation in patients with liver metastases from neuroendocrine tumors: a systematic review. Surgery 2017 Sep;162(3):525-536. [doi: 10.1016/j.surg.2017.05.006] [Medline: 28624178]
6. Moris D, Kostakis ID, Machairas N, Prodromidou A, Tsilimigras DI, Ravindra KV, et al. Comparison between liver transplantation and resection for hilar cholangiocarcinoma: a systematic review and meta-analysis. PLoS One 2019;14(7):e0220527 [FREE Full text] [doi: 10.1371/journal.pone.0220527] [Medline: 31365594]
7. Shah VH, Rao MK. Changing landscape of solid organ transplantation for older adults: trends and post-transplant age-related outcomes. Curr Transplant Rep. 2020. URL: https://doi.org/10.1007/s40472-020-00275-1 [accessed 2020-12-26]
8. Dutkowski P, Linecker M, DeOliveira ML, Müllhaupt B, Clavien P. Challenges to liver transplantation and strategies to improve outcomes. Gastroenterology 2015 Feb;148(2):307-323. [doi: 10.1053/j.gastro.2014.08.045] [Medline: 25224524]
9. Gao Q, Mulvihill MS, Scheuermann U, Davis RP, Yerxa J, Yerokun BA, et al. Improvement in liver transplant outcomes from older donors: a US national analysis. Ann Surg 2019 Aug;270(2):333-339. [doi: 10.1097/SLA.0000000000002876] [Medline: 29958229]
10. Taylor R, Allen E, Richards JA, Goh MA, Neuberger J, Collett D, Liver Advisory Group to NHS Blood and Transplant. Survival advantage for patients accepting the offer of a circulatory death liver transplant. J Hepatol 2019 May;70(5):855-865. [doi: 10.1016/j.jhep.2018.12.033] [Medline: 30639505]
11. Zheng SS, Xu X, Wu J, Chen J, Wang WL, Zhang M, et al. Liver transplantation for hepatocellular carcinoma: Hangzhou experiences. Transplantation 2008 Jun 27;85(12):1726-1732. [doi: 10.1097/TP.0b013e31816b67e4] [Medline: 18580463]
12. Wang LY, Zheng SS. Advances in predicting the prognosis of hepatocellular carcinoma recipients after liver transplantation. J Zhejiang Univ Sci B 2018 Jul;19(7):497-504 [FREE Full text] [doi: 10.1631/jzus.B1700156] [Medline: 29971988]
13. Wang FS, Fan JG, Zhang Z, Gao B, Wang HY. The global burden of liver disease: the major impact of China. Hepatology 2014 Dec;60(6):2099-2108 [FREE Full text] [doi: 10.1002/hep.27406] [Medline: 25164003]
14. Kolfschoten NE, Kievit J, Gooiker GA, van Leersum NJ, Snijders HS, Eddes EH, et al. Focusing on desired outcomes of care after colon cancer resections; hospital variations in 'textbook outcome'. Eur J Surg Oncol 2013 Feb;39(2):156-163. [doi: 10.1016/j.ejso.2012.10.007] [Medline: 23102705]

15. Moris D, Shaw BI, Gloria J, Kesseli SJ, Samoylova ML, Schmitz R, et al. Textbook outcomes in liver transplantation. World J Surg 2020 Oct;44(10):3470-3477. [doi: 10.1007/s00268-020-05625-9] [Medline: 32488663]
16. Ambrosino N, Vitacca M, Dreher M, Isetta V, Montserrat JM, Tonia T, ERS Tele-Monitoring of Ventilator-Dependent Patients Task Force. Tele-monitoring of ventilator-dependent patients: a European Respiratory Society statement. Eur Respir J 2016 Sep;48(3):648-663 [FREE Full text] [doi: 10.1183/13993003.01721-2015] [Medline: 27390283]
17. Daniel H, Sulmasy LS, HealthPublic Policy Committee of the American College of Physicians. Policy recommendations to guide the use of telemedicine in primary care settings: an American College of Physicians position paper. Ann Intern Med 2015 Nov 17;163(10):787-789. [doi: 10.7326/M15-0498] [Medline: 26344925]
18. Cohen JM, Bunick CG, Perkins SH. The new normal: an approach to optimizing and combining in-person and telemedicine visits to maximize patient care. J Am Acad Dermatol 2020 Nov;83(5):e361-e362 [FREE Full text] [doi: 10.1016/j.jaad.2020.06.075] [Medline: 32593636]
19. Yen YH, Tsai YF, Su VY, Chan SY, Yu WR, Ho H, et al. Use and cost-effectiveness of a telehealth service at a centralized COVID-19 quarantine center in Taiwan: cohort study. J Med Internet Res 2020 Dec 11;22(12):e22703 [FREE Full text] [doi: 10.2196/22703] [Medline: 33259324]
20. Ambrosino N, Vagheggini G, Mazzoleni S, Vitacca M. Telemedicine in chronic obstructive pulmonary disease. Breathe (Sheff) 2016 Dec;12(4):350-356 [FREE Full text] [doi: 10.1183/20734735.014616] [Medline: 28210321]
21. Culmer N, Smith T, Stager C, Wright A, Burgess K, Johns S, et al. Telemedical asthma education and health care outcomes for school-age children: a systematic review. J Allergy Clin Immunol Pract 2020 Jun;8(6):1908-1918. [doi: 10.1016/j.jaip.2020.02.005] [Medline: 32084596]
22. Koehler F, Koehler K, Deckwart O, Prescher S, Wegscheider K, Kirwan B, et al. Efficacy of telemedical interventional management in patients with heart failure (TIM-HF2): a randomised, controlled, parallel-group, unmasked trial. The Lancet 2018 Sep 22;392(10152):1047-1057. [doi: 10.1016/S0140-6736(18)31880-4] [Medline: 30153985]
23. Penninga L, Lorentzen AK, Davis C. A telemedicine case series for acute medical emergencies in Greenland: a model for austere environments. Telemed J E Health 2020 Aug;26(8):1066-1070. [doi: 10.1089/tmj.2019.0123] [Medline: 31804895]
24. Gregory ME, Sonesh SC, Hughes AM, Marttos A, Schulman CI, Salas E. Using telemedicine in mass casualty disasters. Disaster Med Public Health Prep 2020 Feb 05:1-8. [doi: 10.1017/dmp.2019.156] [Medline: 32019620]
25. Yardley L, Joseph J, Michie S, Weal M, Wills G, Little P. Evaluation of a web-based intervention providing tailored advice for self-management of minor respiratory symptoms: exploratory randomized controlled trial. J Med Internet Res 2010 Dec 15;12(4):e66 [FREE Full text] [doi: 10.2196/jmir.1599] [Medline: 21159599]
26. Liu WT, Huang CD, Wang CH, Lee KY, Lin SM, Kuo HP. A mobile telephone-based interactive self-care system improves asthma control. Eur Respir J 2011 Feb;37(2):310-317 [FREE Full text] [doi: 10.1183/09031936.00000810] [Medline: 20562122]
27. Burgos F, Disdier C, de Santamaria EL, Galdiz B, Roger N, Rivera ML, e-Spir@p Group. Telemedicine enhances quality of forced spirometry in primary care. Eur Respir J 2012 Jun;39(6):1313-1318 [FREE Full text] [doi: 10.1183/09031936.00168010] [Medline: 22075488]
28. Xu X, Chen J, Wei Q, Liu ZK, Yang Z, Zhang M, et al. Clinical practice guidelines on liver transplantation for hepatocellular carcinoma in China (2018 edition). Hepatobiliary Pancreat Dis Int 2019 Aug;18(4):307-312. [doi: 10.1016/j.hbpd.2019.06.010] [Medline: 31279679]
29. Li Y, Sun H, Yan X, Wang S, Dong D, Liu X, et al. Magnetic compression anastomosis for the treatment of benign biliary strictures: a clinical study from China. Surg Endosc 2020 Jun;34(6):2541-2550. [doi: 10.1007/s00464-019-07063-8] [Medline: 31399950]
30. Zhao G, Yan X, Ma L, Liu W, Zhang J, Guo H, et al. Biomechanical and performance evaluation of magnetic elliptical-ring compressive anastomoses. J Surg Res 2019 Jul;239:52-59. [doi: 10.1016/j.jss.2019.01.063] [Medline: 30802705]
31. Emre S, Schwartz ME, Altaca G, Sethi P, Fiel MI, Guy SR, et al. Safe use of hepatic allografts from donors older than 70 years. Transplantation 1996 Jul 15;62(1):62-65. [doi: 10.1097/00007890-199607150-00013] [Medline: 8693547]
32. Boteon YL, Afford SC. Machine perfusion of the liver: Which is the best technique to mitigate ischaemia-reperfusion injury? World J Transplant 2019 Jan 16;9(1):14-20 [FREE Full text] [doi: 10.5500/wjt.v9.i1.14] [Medline: 30697517]
33. Rijkse E, IJzermans JN, Minnee RC. Machine perfusion in abdominal organ transplantation: Current use in the Netherlands. World J Transplant 2020 Jan 18;10(1):15-28 [FREE Full text] [doi: 10.5500/wjt.v10.i1.15] [Medline: 32110511]
34. Orons PD, Zajko AB, Bron KM, Trecha GT, Selby RR, Fung JJ. Hepatic artery angioplasty after liver transplantation: experience in 21 allografts. J Vasc Interv Radiol 1995;6(4):523-529. [doi: 10.1016/s1051-0443(95)71128-9] [Medline: 7579858]
35. Cheng YF, Ou HY, Tsang LL, Yu CY, Huang TL, Chen TY, et al. Vascular stents in the management of portal venous complications in living donor liver transplantation. Am J Transplant 2010 May;10(5):1276-1283 [FREE Full text] [doi: 10.1111/j.1600-6143.2010.03076.x] [Medline: 20353467]
36. Yabuta M, Shibata T, Shibata T, Shinozuka K, Isoda H, Okamoto S, et al. Long-term outcome of percutaneous transhepatic balloon angioplasty for portal vein stenosis after pediatric living donor liver transplantation: a single institute's experience. J Vasc Interv Radiol 2014 Sep;25(9):1406-1412. [doi: 10.1016/j.jvir.2014.03.034] [Medline: 24854391]

37. Thornburg B, Katariya N, Riaz A, Desai K, Hickey R, Lewandowski R, et al. Interventional radiology in the management of the liver transplant patient. Liver Transpl 2017 Oct;23(10):1328-1341. [doi: 10.1002/lt.24828] [Medline: 28741309]
38. Prasad R, Yadav RR, Israrahmed A, Mittal SR. Endovascular management in post liver transplant recipients with venous anastomotic site stenosis and an associated iatrogenic arterio-portal fistula: case series and review of literature. J Clin Diagn Res 2020 May:TR01-TR04 [FREE Full text] [doi: 10.7860/JCDR/2020/44144.13711]
39. Dunn MA, Rogal SS, Duarte-Rojo A, Lai JC. Physical function, physical activity, and quality of life after liver transplantation. Liver Transpl 2020 May;26(5):702-708. [doi: 10.1002/lt.25742] [Medline: 32128971]
40. Raeesi Vanani I, Amirhosseini M. IoT-based diseases prediction and diagnosis system for healthcare. In: Chakraborty C, Banerjee A, Kolekar M, Garg L, Chakraborty B, editors. Internet of Things for Healthcare Technologies Studies in Big Data, vol 73. Singapore: Springer; Jun 9, 2020:21-48.
41. Mirskikh I, Mingaleva Z, Kuranov V, Matseeva S. Digitization of medicine in Russia: mainstream development and potential. In: Antipova T, editor. Integrated Science in Digital Age Lecture Notes in Networks and Systems, vol 136. Cham: Springer; May 27, 2020:2020.
42. Ye J, Zuo Y, Xie T, Wu M, Ni P, Kang Y, et al. A telemedicine wound care model using 4G with smart phones or smart glasses: a pilot study. Medicine (Baltimore) 2016 Aug;95(31):e4198 [FREE Full text] [doi: 10.1097/MD.0000000000004198] [Medline: 27495023]
43. Cowie MR, Bax J, Bruining N, Cleland JGF, Koehler F, Malik M, et al. e-Health: a position statement of the European Society of Cardiology. Eur Heart J 2016 Jan 01;37(1):63-66 [FREE Full text] [doi: 10.1093/eurheartj/ehv416] [Medline: 26303835]
44. Marx G, Beckers R, Brokmann JC, Deisz R, Pape HC. Tele-cooperation for innovative care using the example of the University Hospital Aachen. Telematics in intensive care medicine, emergency medicine, and telemedical intersectoral rehabilitation planning in geriatric trauma. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2015 Oct;58(10):1056-1061. [doi: 10.1007/s00103-015-2224-4] [Medline: 26281718]
45. Hanlon P, Daines L, Campbell C, McKinstry B, Weller D, Pinnock H. Telehealth interventions to support self-management of long-term conditions: a systematic metareview of diabetes, heart failure, asthma, chronic obstructive pulmonary disease, and cancer. J Med Internet Res 2017 May 17;19(5):e172 [FREE Full text] [doi: 10.2196/jmir.6688] [Medline: 28526671]
46. Vitacca M, Montini A, Comini L. How will telemedicine change clinical practice in chronic obstructive pulmonary disease? Ther Adv Respir Dis 2018;12:1753465818754778 [FREE Full text] [doi: 10.1177/1753465818754778] [Medline: 29411700]
47. Villa L, Matz O, Olaciregui Dague K, Kluwig D, Rossaint R, Brokmann JC. The assessment of dermatological emergencies in the emergency department via telemedicine is safe: a prospective pilot study. Intern Emerg Med 2020 Oct;15(7):1275-1279. [doi: 10.1007/s11739-020-02323-1] [Medline: 32248403]

Abbreviations

COVID-19: coronavirus disease 2019
ECG: electrocardiography
MELD: Model for End-Stage Liver Disease

Edited by G Eysenbach; submitted 17.01.21; peer-reviewed by H Mehdizadeh, T Menser; comments to author 05.02.21; revised version received 11.04.21; accepted 15.04.21; published 17.05.21.

Please cite as:
Tian M, Wang B, Xue Z, Dong D, Liu X, Wu R, Yu L, Xiang J, Zhang X, Zhang X, Lv Y
Telemedicine for Follow-up Management of Patients After Liver Transplantation: Cohort Study
JMIR Med Inform 2021;9(5):e27175
URL: https://medinform.jmir.org/2021/5/e27175
doi:10.2196/27175
PMID:33999008

Original Paper

Neural Network–Based Retinal Nerve Fiber Layer Profile Compensation for Glaucoma Diagnosis in Myopia: Model Development and Validation

Lei Li[1,2], MS; Haogang Zhu[1,3*], PhD; Zhenyu Zhang[1], BS; Liang Zhao[2], MD; Liang Xu[2], MD; Rahul A Jonas[4], MD; David F Garway-Heath[3], MD; Jost B Jonas[2,5], MD; Ya Xing Wang[2*], MD

[1]State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing, China

[2]Beijing Institute of Ophthalmology, Beijing Tongren Hospital, Capital University of Medical Science, Beijing Ophthalmology and Visual Sciences Key Laboratory, Beijing, China

[3]NIHR Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital NHS Foundation Trust, UCL Institute of Ophthalmology, London, United Kingdom

[4]Department of Ophthalmology, Faculty of Medicine and University Hospital, University of Cologne, Cologne, Germany

[5]Department of Ophthalmology, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[*]these authors contributed equally

Corresponding Author:
Ya Xing Wang, MD
Beijing Institute of Ophthalmology
Beijing Tongren Hospital, Capital University of Medical Science
Beijing Ophthalmology and Visual Sciences Key Laboratory
17 Hougou Lane
Beijing, 100005
China
Phone: 86 18600059315
Email: yaxingw@gmail.com

Abstract

Background: Due to the axial elongation–associated changes in the optic nerve and retina in high myopia, traditional methods like optic disc evaluation and visual field are not able to correctly differentiate glaucomatous lesions. It has been clinically challenging to detect glaucoma in highly myopic eyes.

Objective: This study aimed to develop a neural network to adjust for the dependence of the peripapillary retinal nerve fiber layer (RNFL) thickness (RNFLT) profile on age, gender, and ocular biometric parameters and to evaluate the network's performance for glaucoma diagnosis, especially in high myopia.

Methods: RNFLT with 768 points on the circumferential 3.4-mm scan was measured using spectral-domain optical coherence tomography. A fully connected network and a radial basis function network were trained for vertical (scaling) and horizontal (shift) transformation of the RNFLT profile with adjustment for age, axial length (AL), disc-fovea angle, and distance in a test group of 2223 nonglaucomatous eyes. The performance of RNFLT compensation was evaluated in an independent group of 254 glaucoma patients and 254 nonglaucomatous participants.

Results: By applying the RNFL compensation algorithm, the area under the receiver operating characteristic curve for detecting glaucoma increased from 0.70 to 0.84, from 0.75 to 0.89, from 0.77 to 0.89, and from 0.78 to 0.87 for eyes in the highest 10% percentile subgroup of the AL distribution (mean 26.0, SD 0.9 mm), highest 20% percentile subgroup of the AL distribution (mean 25.3, SD 1.0 mm), highest 30% percentile subgroup of the AL distribution (mean 24.9, SD 1.0 mm), and any AL (mean 23.5, SD 1.2 mm), respectively, in comparison with unadjusted RNFLT. The difference between uncompensated and compensated RNFLT values increased with longer axial length, with enlargement of 19.8%, 18.9%, 16.2%, and 11.3% in the highest 10% percentile subgroup, highest 20% percentile subgroup, highest 30% percentile subgroup, and all eyes, respectively.

Conclusions: In a population-based study sample, an algorithm-based adjustment for age, gender, and ocular biometric parameters improved the diagnostic precision of the RNFLT profile for glaucoma detection particularly in myopic and highly myopic eyes.

(JMIR Med Inform 2021;9(5):e22664) doi:10.2196/22664

KEYWORDS

retinal nerve fiber layer thickness; radial basis neural network; neural network; glaucoma; optic nerve head; optical coherence tomography; myopia; optic nerve

Introduction

Glaucoma, as one of the most common causes of irreversible vision impairment and blindness, is diagnosed by the morphometric analysis of the optic nerve head including the peripapillary retinal nerve fiber layer (RNFL) and by psychophysical techniques such as perimetry [1-3]. These routinely applied techniques decrease in their diagnostic precision in myopic eyes and in particular, in highly myopic globes [4,5]. Due to irregularities in the refractive error and shape of the posterior part of the globe and due to high myopia-associated morphological changes in the macular region, perimetric defects lose their specificity for glaucoma and can have a multitude of causes, in addition to glaucomatous optic nerve damage [6]. Similarly, morphometric methods such as assessment of the neuroretinal rim of the optic disc and measurement of the peripapillary RNFL thickness (RNFLT) become more limited with a greater axial length of the eyes [7-11]. Furthermore, the prevalence of glaucomatous or glaucoma-like optic neuropathy increases with longer axial length, especially beyond an axial length of 26.5 mm, with odds ratios ranging from 1.6 to 3.75 for all myopic eyes and from 3.3 to 4.6 for highly myopic eyes [12-14]. These findings show the need to further improve the available methods to refine the diagnosis of glaucomatous optic neuropathy in myopic eyes.

Previous studies have shown that the thickness profile of peripapillary RNFL depends on systemic and ocular biometric parameters [15-18]. The investigations revealed that the RNFLT decreases with older age, parallel to a histomorphometrically examined loss of retinal ganglion cell axons of 0.3% per year of life, and that the peripapillary distribution of the RNFLT depends on gender, axial length, the optic disc-fovea distance, and the angle between the disc-fovea line and the horizontal (“disc-fovea angle”). In recent years, the neural network technique has been intensively studied and widely applied in computer science, including artificial intelligence in the fields of bioscience and clinical medicine [19-25]. Assuming that a neural network can transform the RNFL profile and make it comparable in eyes that differ in parameters influencing the RNFL profile, in this study, we examined whether such transformation of the RNFLT profile could improve the diagnosis of glaucoma, with special emphasis on myopic and highly myopic eyes.

Methods

Data Collection

Participants were randomly selected from the population-based Beijing Eye Study 2011, in which 3468 participants with an age ≥50 years were enrolled. The Medical Ethics Committee of the Beijing Tongren Hospital approved the study protocol, and all study participants gave their written informed consent. The study population and study design were described in detail previously [26,27].

Due to the relatively small number of glaucoma patients in the Beijing Eye Study, we additionally included another group of glaucoma patients who were randomly selected from the study population of the community-based Kailuan Study, which was a prospective cohort study conducted in the industrial city of Tangshan located 200 kilometers from Beijing [28]. The study was approved by the Ethics Committees of Kailuan General Hospital and followed the guidelines outlined in the Declaration of Helsinki. All participants signed a written informed consent form. Between June 2006 and October 2007, a total of 101,510 individuals (81,110 men) aged 18-98 years were recruited to participate in the study, and the participants were re-examined biannually [28]. In the re-examination period of 2014-2016, a randomly selected group of 14,400 participants from the Kailuan Study additionally underwent an ophthalmological examination including fundus photography and optical coherence tomography (OCT) of the peripapillary RNFL.

Glaucomatous optic neuropathy was defined by absolute criteria, each of which was sufficient for the diagnosis of glaucoma, and by relative criteria. The absolute criteria included a notch in the neuroretinal rim in the temporal inferior region and/or the temporal superior region, so that the inferior-superior-nasal-temporal-rule of the neuroretinal rim shape was not fulfilled; localized RNFL defects that could not be explained by any other cause than glaucoma; and an abnormally large cup in relation to the size of the optic disc. Relative criteria for the diagnosis included a markedly thinner neuroretinal rim in the inferior disc region; a diffuse decrease in the visibility of the RNFL; a marked diffuse and/or focal thinning of the retinal arteries if there was no other reason than glaucoma for retinal vessel thinning; or an optic disc hemorrhage, if there was no other reason for disc bleeding such as retinal vessel occlusions. If none of the absolute glaucoma criteria was fulfilled, the diagnosis of glaucoma required that at least 2 relative criteria had to be fulfilled, among them had to be a suspicious neuroretinal rim shape in eyes with an optic cup large enough for the assessment of the rim shape or at least 2 relative criteria had to be positive including the occurrence of an optic cup in a small optic disc, which usually would not show cupping [29]. These criteria were similar to those suggested by Foster and colleagues [30]. Using digital fundus photographs, the assessment of glaucomatous optic neuropathy was carried out by two senior graders (YXW, JBJ). In case of disagreement, the optic disc photographs were re-assessed up to 3 times, until eventually both graders agreed upon the diagnosis.

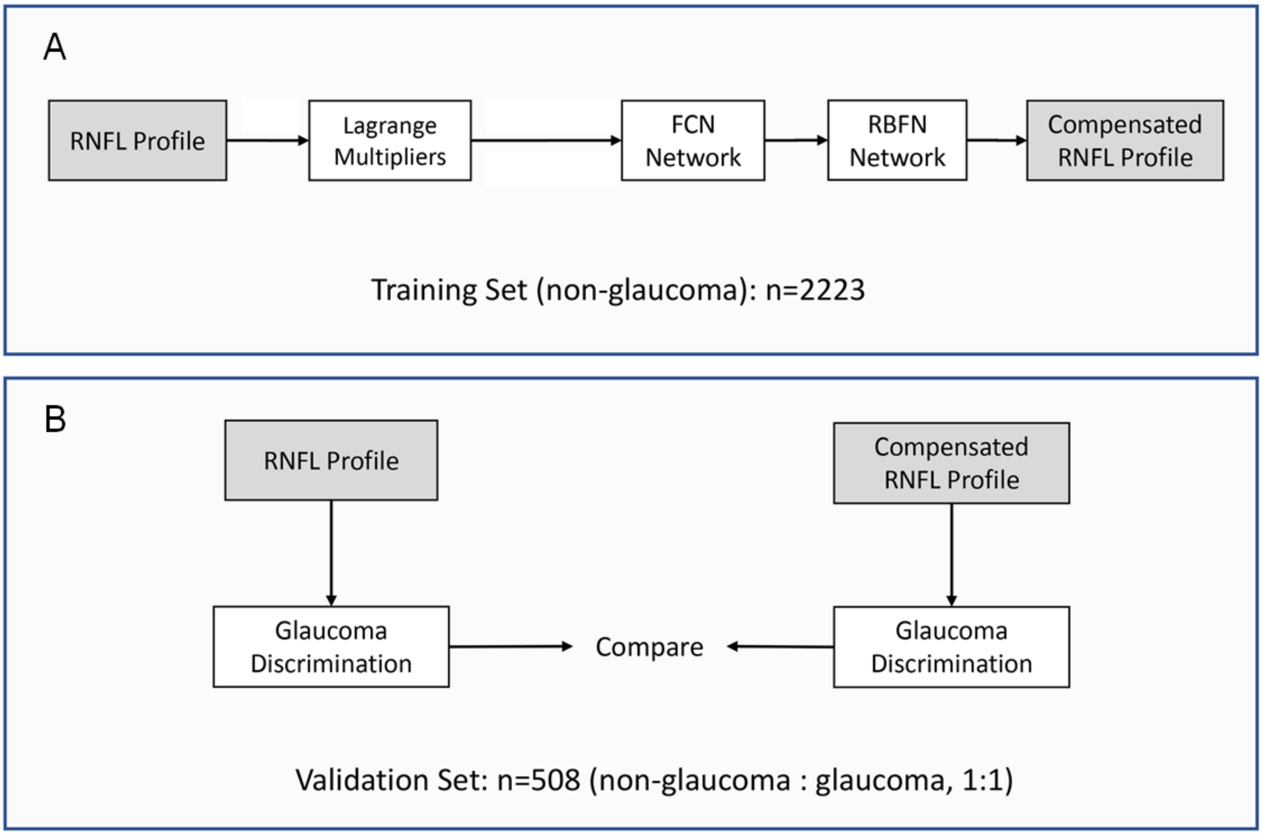
All study participants (Beijing Eye Study and Kailuan Study) underwent spectral domain OCT (Spectralis OCT; Heidelberg Engineering, Heidelberg, Germany) including a circular B-scan centered on the optic disc center with a diameter of 3.4 mm. Fundus photographs of the macula and optic disc were additionally taken (CR6-45NM Camera; Canon Inc, Ota, Tokyo, Japan). Using optical low-coherence reflectometry (Lenstar 900 Optical Biometer; Haag-Streit, Koeniz, Switzerland), biometry of the right eyes was performed for measurement of the anterior corneal curvature, central corneal thickness, anterior chamber depth, lens thickness, and axial length. The disc-fovea distance and the angle between the disc-fovea line and the horizontal ("disc-fovea angle") were measured on fundus photographs by one grader (RAJ) [30,31]. The magnification was corrected using the Littmann-Bennett method [31,32].

We used the Heidelberg Explorer (HEE, version 5.3; Heidelberg Engineering, Heidelberg, Germany) for the automatic segmentation of the RNFL and to calculate the RNFLT. The upper border and the lower border of the RNFL were automatically outlined and generated. In rare cases with obvious misalignment, the RNFL were manually re-adjusted by trained examiners (LZ). The data of 768 RNFLT measurements equally spaced on the 360° circle were extracted, and the RNFLT profile was composed. RNFL scans with a quality score less than 15 were excluded. The data for 1 eye per individual were used for the statistical analysis.

Training of RNFL Profile Compensation

Based on the findings obtained in previous investigations, 5 parameters shown to be associated with the RNFLT profile were chosen to be included in the present study: age, gender, axial length, the disc-fovea distance, and the disc-fovea angle. These parameters were used for the training of the RNFL profile compensation [16,31-34]. The training was performed with the images obtained from 2223 eyes from 2223 participants randomly chosen from the control group. Due to the positive correlation between older age and longer axial length, 2 independent phases were carried out. In the first phase, the parameter of age was inputted as the only factor to compensate the RNFLT vertically. Lagrange multiplier methods were applied to optimize the variance between the compensated RNFLT and the initial RNFLT, depending on the fact that each point in the RNFL profile was interassociated with neighboring points. In the second phase, the parameters of axial length, disc-fovea distance, disc-fovea angle, and gender were included in a fully connected network (FCN) for the RNFLT compensation in both the vertical and horizontal directions. The output from the FCN was further trained by a radial basis function network (RBFN) embedded with a spatial correlation, to optimize the variance between the compensated RNFLT data (Figure 1). Details of the 2-phase compensation are described in Multimedia Appendix 1.

Figure 1. Overview of the 2-phased process in retinal nerve fiber layer (RNFL) profile compensation and its validation in discriminating glaucoma, which consisted of (A) applying the Lagrange multiplier, fully connected network (FCN), and radial basis function network (RBFN) to the training set, composed of 2223 eyes from 2223 nonglaucomatous participants, for RNFL thickness (RNFLT) compensation based on the impact of axial length (AL), age, disc-fovea angle (DFA), and disc fovea distance (DFD) and (B) evaluation of the performance of compensated RNFLT for glaucoma discrimination by comparing with the performance of the original RNFL profile.



Validation

The validation was performed in a separate dataset containing both glaucomatous and nonglaucomatous eyes in a relationship of 1:1. The compensation algorithm was applied, and discrimination between glaucoma versus no glaucoma was carried out using either the original RNFLT profile or the compensated RNFLT profile. An eye was marked as glaucomatous if the thickness values of continuous points in the original RNFL profile or in the compensated RNFL profile were located below the single-sided 95% confidence interval of the original RNFL profile of the nonglaucomatous eyes or the compensated RNFL profile of the nonglaucomatous eyes, respectively. A receiver operating characteristic (ROC) curve including the area under the ROC curve (AUROC) was calculated to evaluate the performance of RNFLT data, in their original form and in their compensated form, for the detection of glaucoma. The accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were additionally analyzed.

Results

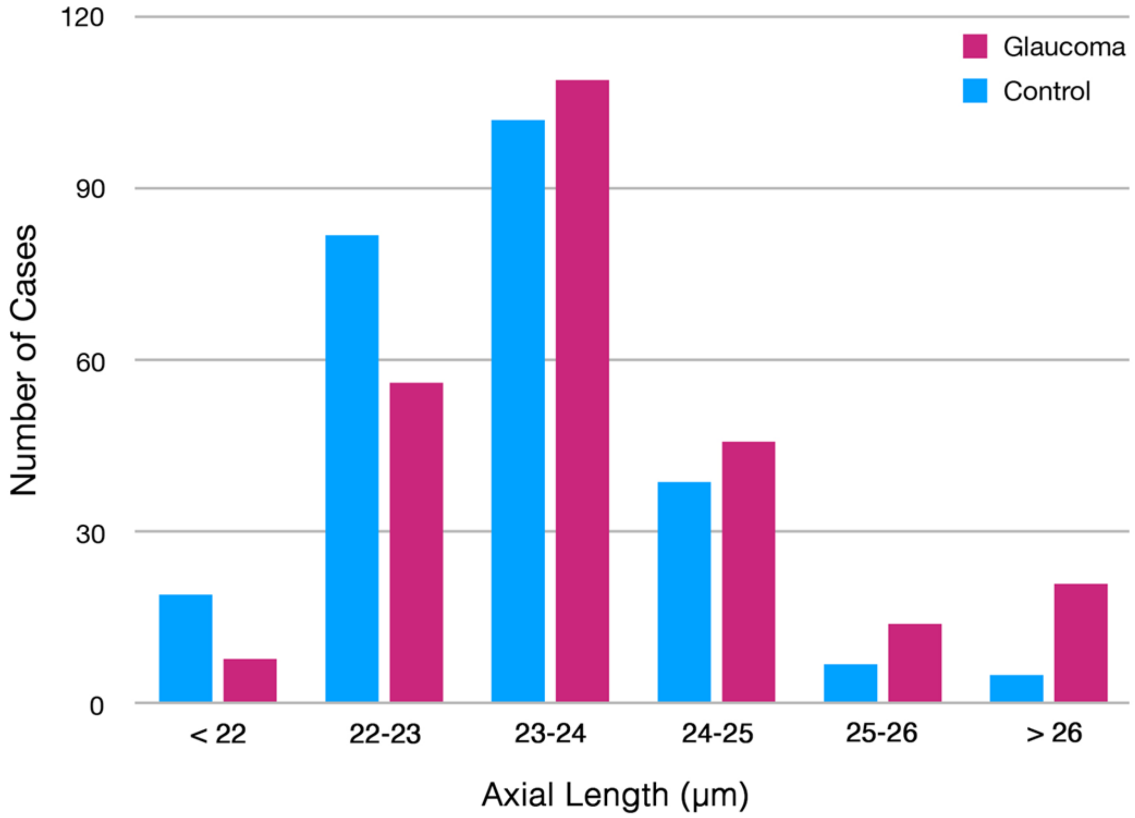
Among the 3654 participants in the Beijing Eye Study 2011, 2622 eyes from 2622 participants were randomly chosen, including 2477 individuals for the control group and 145 patients with glaucoma for the study group. After adding 109 glaucomatous eyes from 109 randomly selected patients from the Kailuan Study, a total of 2731 eyes from 2731 participants (2477 control and 254 glaucoma; men: 1214/2731, 44.5%) were included, with a mean age of 63.0 (SD 9.2; range: 50-91) years. Due to an insufficient scan quality, we excluded 26 eyes (26/2731, 0.9%) from the analysis, so that the training data were eventually composed of 2223 randomly selected control eyes, and the validation group included 254 individuals in the validation control group and 254 patients with glaucoma (Table 1). The glaucomatous eyes had a longer axial length (mean 23.77, SD 1.28 mm) as compared with the nonglaucomatous eyes (mean 23.30, SD 0.96 mm) in the validation set (*P*<.001; Figure 2).

Table 1. Demographic and ocular parameters of the study population.

Eye sets	Age (years), mean (SD, range)	Gender (male), n (%)	Axial length (mm), mean (SD, range)	Disc-fovea distance (mm), mean (SD, range)	Disc-fovea angle (°), mean (SD, range)	Mean RNFLT[a] (μm), mean (SD, range)
All (n=2731)	63.0 (9.2, 50 to 91)	1214 (44.5)	23.23 (1.01, 18.96 to 28.87)	4.93 (0.39, 3.68 to 7.63)	7.64 (3.51, –16.64 to 23.25)	101 (12, 32 to 147)
Normal (n=2477)	62.3 (8.9, 50 to 91)	1076 (43.4)	23.18 (0.96, 18.96 to 28.87)	4.88 (0.27, 3.68 to 5.99)	7.59 (3.4, –16.64 to 23.25)	102 (11, 43 to 147)
Glaucoma (n=254)	69.4 (9.3, 50 to 90)	138 (54.3)	23.77 (1.27, 19.59 to 28.84)	5.41 (0.82, 3.8 to 7.63)	8.16 (4.36, –13.14 to 22.59)	85 (17, 33 to 122)
Training set						
All eyes (n=2223)	62.2 (8.9, 50 to 91)	960 (43.2)	23.16 (0.96, 18.96 to 28.87)	4.88 (0.27, 3.68 to 5.99)	7.67 (3.42, –16.64 to 23.25)	102 (11, 43 to 141)
10% longest eyes (n=222)	63.5 (8.7, 50 to 85)	145 (65.3)	25.08 (0.77, 24.32 to 28.78)	4.86 (0.3, 3.86 to 5.99)	7.49 (3.86, –6.3 to 23.25)	96 (10, 60 to 119)
20% longest eyes (n=444)	63.41 (8.77, 50 to 85)	290 (65.3)	24.56 (0.76, 23.83 to 28.87)	4.83 (0.28, 3.68 to 5.99)	7.64 (3.55, –6.3 to 23.25)	98 (11, 43 to 124)
30% longest eyes (n=666)	63.3 (8.97, 50 to 90)	408 (61.2)	24.26 (0.75, 23.53 to 28.87)	4.84 (0.27, 3.68 to 5.99)	7.62 (3.4, –6.3 to 23.25)	100 (11, 43 to 139)
Validation set						
All eyes n=508)	66.4 (9.6, 50 to 90)	254 (50. 0)	23.53 (1.15, 19.59 to 28.84)	5.14 (0.67, 3.8 to 7.63)	7.54 (3.87, –13.14 to 22.59)	94 (17, 32 to 147)
Glaucoma in all eyes (n=254)	69.4 (9.3, 50 to 90)	138 (54.3)	23.77 (1.28, 19.59 to 28.84)	5.41 (0.82, 3.8 to 7.63)	8.16 (4.37, –13.14 to 22.59)	86 (17, 32 to 122)
10% longest eyes (n=51)	68.3 (9.3, 50 to 90)	32 (62.7)	26.01 (0.89, 24.95 to 28.84)	5.3 (0.9, 4.05 to 7.63)	8.44 (4.39, –2.05 to 22.59)	82 (16, 40 to 122)
Glaucoma in 10% longest eyes (n=37)	68.78 (9.8, 50 to 90)	23 (62.2)	26.2 (0.93, 24.95 to 28.84)	5.46 (0.99, 4.05 to 7.63)	8.88 (4.92, –2.05 to 22.59)	80 (17, 40 to 122)
20% longest eyes (n=102)	67.8 (9.2, 50 to 90)	66 (64.7)	25.26 (0.99, 24.21 to 28.84)	5.34 (0.85, (4.05 to 7.63)	7.92 (3.99, –2.05 to 22.59)	88 (17, 38 to 122)
Glaucoma in 20% longest eyes (n=66)	68.1 (9.2, 50 to 90)	41 (62.1)	25.44 (1.08, 24.21 to 28.84)	5.58 (0.95, 4.05 to 7.63)	8.5 (4.43, –2.05 to 22.59)	83 (17, 38 to 122)
30% longest eyes (n=153)	67.2 (9.0, 50 to 90)	99 (64.7)	24.86 (0.99, 23.92 to 28.84)	5.27 (0.77, 3.8 to 7.63)	7.51 (4.18 (–13.14 to 22.59)	90 (17, 38 to 147)
Glaucoma in 30% longest eyes (n=91)	67.8 (9.0, 50 to 90)	57 (62.6)	25.06 (1.11, 23.92 to 28.84)	5.5 (0.9, 3.8 to 7.63)	7.91 (4.79, –13.14 to 22.59)	85 (17, 38 to 122)

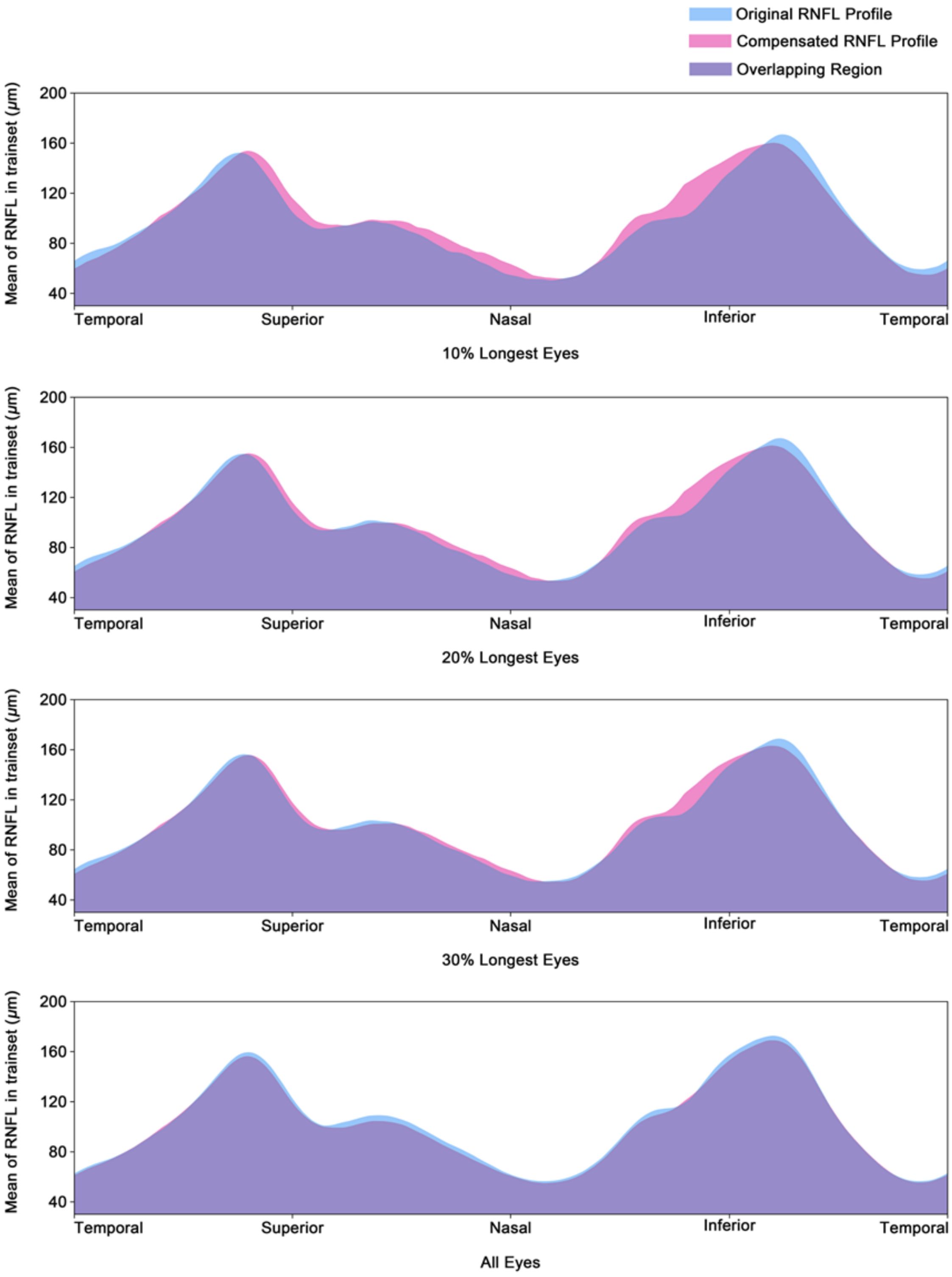
[a] RNFLT: retinal nerve fiber layer thickness.

Figure 2. Distribution of axial length of the glaucomatous eyes and control eyes in the validation group.



The compensation-induced change in the RNFLT values in height (vertical) and in location (horizontal) increased with longer axial length (Figure 3). It was most marked in the eyes with the longest axial length: The subgroup of eyes in the highest 10% percentile of the axial length distribution (mean axial length 25.08, SD 0.77 mm) had the highest compensation, followed by the subgroup of eyes in the highest 20% percentile of the axial length distribution (mean 24.56, SD 0.76 mm) and the subgroup of eyes in the highest 30% percentile of the axial length distribution (mean 24.26, SD 0.75 mm). The mean difference between the uncompensated RNFLT values and the compensated values was negligible in the eyes with an axial length outside of the 30% percentile of the longest axial length (mean axial length 23.16, SD 0.96 mm; Figure 3).

Figure 3. The mean original retinal nerve fiber layer (RNFL) profile (blue) and the mean compensated RNFL profile (pink) of the 10% longest eyes, 20% longest eyes, 30% longest eyes, and all eyes.



Comparing the compensated RNFLT values with the uncompensated RNFLT values revealed that the AUROC for the detection of glaucoma increased from 0.70 to 0.84, from 0.75 to 0.89, from 0.77 to 0.89, and from 0.78 to 0.87, for eyes within the 10% highest length percentile, eyes within the 20% highest length percentile, eyes within the 30% highest axial length percentile, and all eyes, respectively (Figure 4). The relative increase was more pronounced in eyes with longer axial length, with an increase by 19.8%, 18.9%, 16.2%, and 11.3% in the highest 10% percentile subgroup, highest 20% percentile subgroup, highest 30% percentile subgroup, and all eyes, respectively. The accuracy, sensitivity, specificity, PPV, and NPV of the original and compensated RNFL in subgroups are shown in Table 2.

Figure 4. Area under receiver operation curve (AUROC) for the detection of glaucoma in the validation data set before (blue line) and after (pink line) the transformation, in eyes of the 10% longest axial length (mean 26.01 mm), 20% longest axial length (mean 25.26 mm), 30% longest axial length (mean 24.86 mm), and all eyes (mean 23.53 mm).

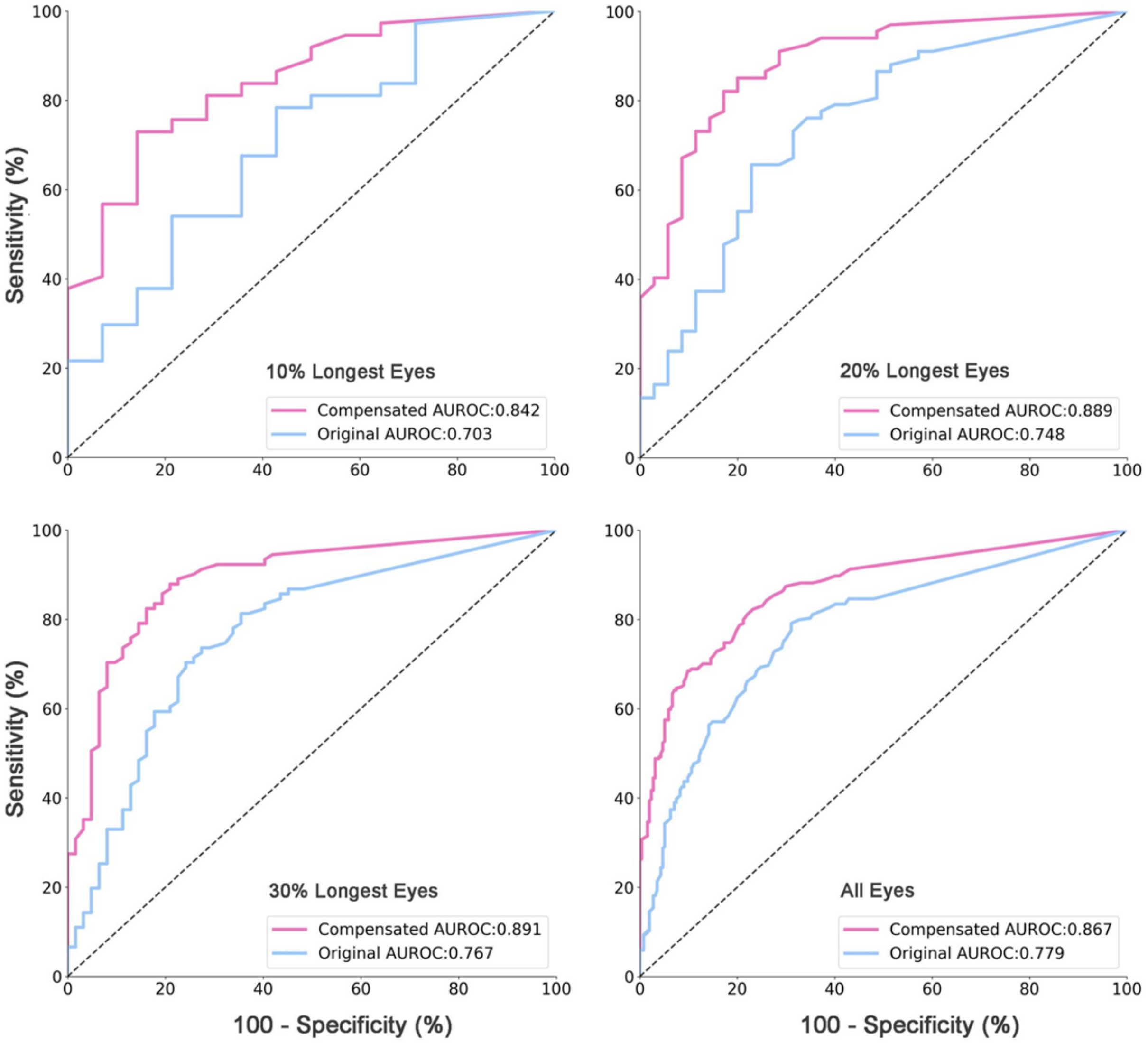


Table 2. Performance of the original retinal nerve fiber layer (RNFL) and the compensated RNFL to detect glaucoma in subgroups and all eyes of the validating dataset.

Eye groups	Accuracy	Sensitivity	Specificity	Positive predictive value	Negative predictive value
10% longest eyes					
Original	0.073	0.784	0.571	0.829	0.500
Compensated	0.077	0.730	0.857	0.931	0.545
20% longest eyes					
Original	0.140	0.657	0.771	0.846	0.540
Compensated	0.165	0.821	0.829	0.902	0.707
30% longest eyes					
Original	0.220	0.736	0.726	0.798	0.652
Compensated	0.250	0.824	0.839	0.882	0.765
All eyes					
Original	0.740	0.791	0.689	0.718	0.768
Compensated	0.795	0.811	0.779	0.786	0.805

Discussion

Principal Findings

In our population-based study, the diagnostic precision of the peripapillary RNFLT profile for the detection of glaucoma increased when the dependence of the RNFLT profile on age and the ocular biometric parameters of axial length, disc-fovea distance, and disc-fovea angle were taken into account by applying 2 neural networks. These networks, FCN and RBFN, developed an algorithm by which the RNFLT profile was transformed either horizontally or vertically. Applying the algorithm increased the diagnostic performance of the RNFLT profile, which was markedly better with longer axial length. The improvement in relative percentage points as measured by the AUROC was 19.8% in the subgroup of eyes within the highest 10% percentile group, 16.2% in the highest 30% percentile subgroup, and 11.3% in all eyes of the study population.

Myopia-related changes in the appearance of the optic nerve head can make the detection of additional changes caused by glaucomatous optic neuropathy in myopic eyes difficult [5]. The parapapillary gamma zone and delta zone in myopic eyes increase the brightness of the background so that the visibility of the retinal nerve fiber layer upon ophthalmoscopy is reduced due to a physical-optical effect. The presence of a gamma and delta zone additionally leads to irregularities in the profile of the tissues underlying the RNFL, so that the automatic delineation of the inner retinal layer containing the retinal nerve fibers from the subsequent layer gets more difficult. The axial elongation–associated increase in the parapapillary region by the development of the gamma and delta zone can lead to a thinning of the RNFL due to geometric reasons. In moderate myopia, the Bruch's membrane opening as the inner opening layer of the optic nerve head usually shifts temporally in the direction of the fovea, leading to an overhanging of the Bruch's membrane at the nasal optic disc border and a lack of the Bruch's membrane at the temporal disc border (ie, gamma zone) [35]. The resulting oblique course of the retinal ganglion cell axons through the myopic optic nerve head canal as compared to a perpendicular course in emmetropic eyes leads to a change in the configuration of the neuroretinal rim in myopic eyes, rendering the detection of glaucomatous rim changes more difficult. The axial elongation–associated enlargement of the optic disc is associated with a stretching of the lamina cribrosa so that the depth of the optic cup may be reduced. It leads to decreased spatial contrast between the height of the neuroretinal rim and the depth of the optic cup and thus renders the delineation of the rim from the cup more difficult. Simultaneously, the color of the rim changes from pink in direction to yellow, so that the color contrast between the rim and optic cup decreases in myopic eyes, again rendering the differentiation of the rim from the optic cup more difficult. As also pointed out earlier in the paragraph, perimetric changes also lose their specificity for glaucomatous optic nerve damage as their cause. The axial elongation–associated changes can also present with perimetric defects that mimic or cover a glaucoma-related visual field defect. These changes might include diffuse peripapillary and macular chorioretinal atrophy, macular Bruch's membrane defects, and scleral staphylomas. Furthermore, the intraocular pressure in myopic eyes with glaucomatous can be within the normal range since the myopia-associated stretching and thinning of the lamina cribrosa and peripapillary scleral flange may increase the pressure susceptibility of the optic nerve fibers when passing through the lamina cribrosa. These examples may demonstrate the need for improved morphometric glaucoma diagnosis in myopic eyes [4,5].

Previous studies showed that the thickness profile of the RNFL depended on other morphologic parameters such as axial length, the disc-fovea distance, and the disc-fovea angle [31,32]. The longer the axial length and disc-fovea distance were, the smaller the angle kappa between the temporal superior and temporal inferior vascular arcade, which accompanies the RNFL branches. The disc-fovea angle was a surrogate for sagittal rotation of the optic nerve head, also influencing the location of the RNFLT profile. By taking these associations of the RNFLT profile into consideration and adjusting for them using a compensation algorithm, there was an improvement in the diagnostic precision of the RNFLT profile for the detection of glaucoma (Figures 3 and 4). The improvement was more marked with more myopic eyes.

The AUROC values found in our study population are roughly comparable to those of previous investigations. To cite examples, Shoji and colleagues [9,35] examined 31 patients with high myopia and 51 patients with high myopia and glaucoma and found that the peripapillary RNFLT had an AUROC of 0.83 in the discrimination of normal eyes from glaucomatous eyes. Kim and associates [36] reported that the ability to detect glaucomatous changes in a highly myopic group (n=45) by RNFL examination had an AUROC of 0.83. When comparing the various studies, one may consider that they markedly differed in the size and composition of their study population. In particular, our study population was recruited in a population-based manner. Subsequently, the glaucoma patients showed all stages, including early stages, of glaucomatous optic neuropathy. In addition, the nonglaucomatous group in our study population included eyes with nonglaucomatous optic nerve damage in addition to other pathological conditions like retinal diseases, nonglaucomatous neuropathies, and cataract. If we had included only eyes without any (nonglaucomatous) optic nerve damage and without any retinal disease in the control group, separating the glaucomatous study and control group would have been easier, and the AUROC would have been higher.

The findings that the height and profile of the peripapillary RNFLT were associated with various ocular and systemic parameters were also found in other investigations. Yamashita and colleagues [37] noted that the position of the superior-temporal RNFLT peak was associated with the location of the papillomacular position, optic disc tilt, and body height, while the inferior-temporal RNFL peak position was correlated with corneal thickness and axial length. Leung et al [8] investigated 189 myopic eyes and reported that the angle between the superotemporal and inferotemporal RNFL bundles decreased with longer axial length. Fujino et al [38] found that a RNFLT profile correction based on the retinal vessel position in all twelve 30° sectors was able to improve the structure-function relationship in all sectors. Rho et al [39] adjusted the 1% reference line of the RNFLT profile according to the retinal vessel position, by which they obtained better agreement with the standard diagnosis of glaucoma. These previous studies on the dependence of the RNFLT profile on other ocular parameters revealed, however, that these associations with the RNFLT profile change were not linear and that the effect of a correction by a linear mathematical method was limited. In this study, the FCN and RBFN were used to compensate the RNFLT profile in both the horizontal direction (position shift) and vertical direction (thickness change). Decreasing the systemic variability of the RNFLT profile resulted in an improvement of the diagnostic performance for glaucoma detection, especially in highly myopic patients.

Limitations

When discussing the results of our study, its limitations should be taken into account. First, compensation of the RNFLT profile was based on data from participants with an age ≥50 years, and the performance of glaucoma detection was not validated in younger participants. Second, the composition of the validation dataset included a 1:1 ratio of glaucomatous eyes to nonglaucomatous eyes. However, since the prevalence of glaucoma increases with axial length, a relatively high proportion of glaucomatous eyes in the validation set may reflect the higher prevalence of glaucoma in eyes with myopia and high myopia. The strengths of the study included that the large population-based dataset offered an opportunity to observe the diverse patterns of the nonlinear relationship between the RNFLT profile and axial length. An RBFN with the advantages of good generalization, strong tolerance to input noise, and online learning ability made it possible to interpret the patterns to a reliable compensation. Due to the population-based recruitment of the study population, the validation group included glaucomatous eyes of all glaucoma stages, so that the results are more generalizable than in hospital-based studies with a preponderance of advanced glaucoma stages in the study groups.

Conclusion

Applying an algorithm to adjust the nonlinear dependence of the RNFLT profile on age, axial length, disc-fovea distance, and disc-fovea angle resulted in improved diagnostic precision of the peripapillary RNFLT profile for the detection of glaucoma in a population-based study population. The improvement in the diagnostic precision of the compensated versus uncompensated RNFLT profile data increased in relative terms with longer axial length. With an increase of 20%, it was most marked in the highly myopic group. The application of this neural network–based RNFLT profile compensation may also be helpful to improve glaucoma diagnosis in myopic eyes in clinical practice.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (number 81570835) and the State Key Laboratory of Software Development Environment, Beihang University, Beijing, China.

Conflicts of Interest

DFG-H is an unpaid consultant for Carl Zeiss Meditec, and offers reserach support for Carl Zeiss Meditec and Topcon. The other authors have no conflicts to declare.

Multimedia Appendix 1

Details of the two-phase-compensation of retinal nerve fiber layer.

[PDF File (Adobe PDF File), 328 KB - medinform_v9i5e22664_app1.pdf]

References

1. Flaxman SR, Bourne RRA, Resnikoff S, Ackland P, Braithwaite T, Cicinelli MV, Vision Loss Expert Group of the Global Burden of Disease Study. Global causes of blindness and distance vision impairment 1990-2020: a systematic review and meta-analysis. Lancet Glob Health 2017 Dec;5(12):e1221-e1234 [FREE Full text] [doi: 10.1016/S2214-109X(17)30393-5] [Medline: 29032195]
2. Bourne RRA, Flaxman SR, Braithwaite T, Cicinelli MV, Das A, Jonas JB, Vision Loss Expert Group. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. Lancet Glob Health 2017 Sep;5(9):e888-e897 [FREE Full text] [doi: 10.1016/S2214-109X(17)30293-0] [Medline: 28779882]
3. Weinreb RN, Khaw PT. Primary open-angle glaucoma. The Lancet 2004 May 22;363(9422):1711-1720. [doi: 10.1016/S0140-6736(04)16257-0] [Medline: 15158634]
4. Chang RT, Singh K. Myopia and glaucoma: diagnostic and therapeutic challenges. Curr Opin Ophthalmol 2013 Mar;24(2):96-101. [doi: 10.1097/ICU.0b013e32835cef31] [Medline: 23542349]
5. Tan NYQ, Sng CCA, Jonas JB, Wong TY, Jansonius NM, Ang M. Glaucoma in myopia: diagnostic dilemmas. Br J Ophthalmol 2019 Oct;103(10):1347-1355. [doi: 10.1136/bjophthalmol-2018-313530] [Medline: 31040131]
6. Weinreb RN, Leung CKS, Crowston JG, Medeiros FA, Friedman DS, Wiggs JL, et al. Primary open-angle glaucoma. Nat Rev Dis Primers 2016 Sep 22;2:16067. [doi: 10.1038/nrdp.2016.67] [Medline: 27654570]
7. Kim NR, Lim H, Kim JH, Rho SS, Seong GJ, Kim CY. Factors associated with false positives in retinal nerve fiber layer color codes from spectral-domain optical coherence tomography. Ophthalmology 2011 Sep;118(9):1774-1781. [doi: 10.1016/j.ophtha.2011.01.058] [Medline: 21550120]
8. Leung CK, Yu M, Weinreb RN, Mak HK, Lai G, Ye C, et al. Retinal nerve fiber layer imaging with spectral-domain optical coherence tomography: interpreting the RNFL maps in healthy myopic eyes. Invest Ophthalmol Vis Sci 2012 Oct 17;53(11):7194-7200. [doi: 10.1167/iovs.12-9726] [Medline: 22997288]
9. Shoji T, Nagaoka Y, Sato H, Chihara E. Impact of high myopia on the performance of SD-OCT parameters to detect glaucoma. Graefes Arch Clin Exp Ophthalmol 2012 Dec;250(12):1843-1849. [doi: 10.1007/s00417-012-1994-8] [Medline: 22555896]
10. Suwan Y, Rettig S, Park SC, Tantraworasin A, Geyman LS, Effert K, et al. Effects of Circumpapillary Retinal Nerve Fiber Layer Segmentation Error Correction on Glaucoma Diagnosis in Myopic Eyes. J Glaucoma 2018 Nov;27(11):971-975. [doi: 10.1097/IJG.0000000000001054] [Medline: 30113513]
11. Qiu K, Zhang M, Wu Z, Nevalainen J, Schiefer U, Huang Y, et al. Retinal nerve fiber bundle trajectories in Chinese myopic eyes: Comparison with a Caucasian based mathematical model. Exp Eye Res 2018 Nov;176:103-109. [doi: 10.1016/j.exer.2018.07.002] [Medline: 30008388]
12. Xu L, Wang Y, Wang S, Wang Y, Jonas JB. High myopia and glaucoma susceptibility the Beijing Eye Study. Ophthalmology 2007 Mar;114(2):216-220. [doi: 10.1016/j.ophtha.2006.06.050] [Medline: 17123613]
13. Perera SA, Wong TY, Tay W, Foster PJ, Saw S, Aung T. Refractive error, axial dimensions, and primary open-angle glaucoma: the Singapore Malay Eye Study. Arch Ophthalmol 2010 Jul;128(7):900-905. [doi: 10.1001/archophthalmol.2010.125] [Medline: 20625053]
14. Marcus MW, de Vries MM, Junoy Montolio FG, Jansonius NM. Myopia as a risk factor for open-angle glaucoma: a systematic review and meta-analysis. Ophthalmology 2011 Oct;118(10):1989-1994.e2. [doi: 10.1016/j.ophtha.2011.03.012] [Medline: 21684603]
15. Hoh S, Lim MCC, Seah SKL, Lim ATH, Chew S, Foster PJ, et al. Peripapillary retinal nerve fiber layer thickness variations with myopia. Ophthalmology 2006 May;113(5):773-777. [doi: 10.1016/j.ophtha.2006.01.058] [Medline: 16650672]
16. Wang YX, Pan Z, Zhao L, You QS, Xu L, Jonas JB. Retinal nerve fiber layer thickness. The Beijing Eye Study 2011. PLoS One 2013;8(6):e66763 [FREE Full text] [doi: 10.1371/journal.pone.0066763] [Medline: 23826129]
17. Knight OJ, Girkin CA, Budenz DL, Durbin MK, Feuer WJ, Cirrus OCT Normative Database Study Group. Effect of race, age, and axial length on optic nerve head parameters and retinal nerve fiber layer thickness measured by Cirrus HD-OCT. Arch Ophthalmol 2012 Mar;130(3):312-318 [FREE Full text] [doi: 10.1001/archopthalmol.2011.1576] [Medline: 22411660]
18. Yamashita T, Asaoka R, Kii Y, Terasaki H, Murata H, Sakamoto T. Structural parameters associated with location of peaks of peripapillary retinal nerve fiber layer thickness in young healthy eyes. PLoS One 2017;12(5):e0177247 [FREE Full text] [doi: 10.1371/journal.pone.0177247] [Medline: 28542289]
19. 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]

20. Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. JAMA 2017 Dec 12;318(22):2211-2223 [FREE Full text] [doi: 10.1001/jama.2017.18152] [Medline: 29234807]
21. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018 Mar;2(3):158-164. [doi: 10.1038/s41551-018-0195-0] [Medline: 31015713]
22. Shibata N, Tanito M, Mitsuhashi K, Fujino Y, Matsuura M, Murata H, et al. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. Sci Rep 2018 Oct 02;8(1):14665 [FREE Full text] [doi: 10.1038/s41598-018-33013-w] [Medline: 30279554]
23. Wisse RPL, Muijzer MB, Cassano F, Godefrooij DA, Prevoo YFDM, Soeters N. Validation of an Independent Web-Based Tool for Measuring Visual Acuity and Refractive Error (the Manifest versus Online Refractive Evaluation Trial): Prospective Open-Label Noninferiority Clinical Trial. J Med Internet Res 2019 Nov 08;21(11):e14808 [FREE Full text] [doi: 10.2196/14808] [Medline: 31702560]
24. Nam SM, Peterson TA, Butte AJ, Seo KY, Han HW. Explanatory Model of Dry Eye Disease Using Health and Nutrition Examinations: Machine Learning and Network-Based Factor Analysis From a National Survey. JMIR Med Inform 2020 Mar 20;8(2):e16153 [FREE Full text] [doi: 10.2196/16153] [Medline: 32130150]
25. Guo Y, Hao Z, Zhao S, Gong J, Yang F. Artificial Intelligence in Health Care: Bibliometric Analysis. J Med Internet Res 2020 Jul 29;22(7):e18228 [FREE Full text] [doi: 10.2196/18228] [Medline: 32723713]
26. Jonas JB, Xu L, Wang YX. The Beijing Eye Study. Acta Ophthalmol 2009 May;87(3):247-261 [FREE Full text] [doi: 10.1111/j.1755-3768.2008.01385.x] [Medline: 19426355]
27. Yan YN, Wang YX, Xu L, Xu J, Wei WB, Jonas JB. Fundus Tessellation: Prevalence and Associated Factors: The Beijing Eye Study 2011. Ophthalmology 2015 Sep;122(9):1873-1880. [doi: 10.1016/j.ophtha.2015.05.031] [Medline: 26119000]
28. Wang L, Cui L, Wang Y, Vaidya A, Chen S, Zhang C, et al. Resting heart rate and the risk of developing impaired fasting glucose and diabetes: the Kailuan prospective study. Int J Epidemiol 2015 Apr;44(2):689-699 [FREE Full text] [doi: 10.1093/ije/dyv079] [Medline: 26002923]
29. Wang YX, Xu L, Yang H, Jonas JB. Prevalence of glaucoma in North China: the Beijing Eye Study. Am J Ophthalmol 2010 Dec;150(6):917-924. [doi: 10.1016/j.ajo.2010.06.037] [Medline: 20970107]
30. Foster PJ, Buhrmann R, Quigley HA, Johnson GJ. The definition and classification of glaucoma in prevalence surveys. Br J Ophthalmol 2002 Mar;86(2):238-242 [FREE Full text] [doi: 10.1136/bjo.86.2.238] [Medline: 11815354]
31. Jonas RA, Wang YX, Yang H, Li JJ, Xu L, Panda-Jonas S, et al. Optic Disc-Fovea Distance, Axial Length and Parapapillary Zones. The Beijing Eye Study 2011. PLoS One 2015;10(9):e0138701 [FREE Full text] [doi: 10.1371/journal.pone.0138701] [Medline: 26390438]
32. Jonas RA, Wang YX, Yang H, Li JJ, Xu L, Panda-Jonas S, et al. Optic Disc - Fovea Angle: The Beijing Eye Study 2011. PLoS One 2015;10(11):e0141771 [FREE Full text] [doi: 10.1371/journal.pone.0141771] [Medline: 26545259]
33. Mwanza J, Lee G, Budenz DL. Effect of Adjusting Retinal Nerve Fiber Layer Profile to Fovea-Disc Angle Axis on the Thickness and Glaucoma Diagnostic Performance. Am J Ophthalmol 2016 Jan;161:12-21.e1. [doi: 10.1016/j.ajo.2015.09.019] [Medline: 26387935]
34. Zhang Q, Xu L, Wei WB, Wang YX, Jonas JB. Size and Shape of Bruch's Membrane Opening in Relationship to Axial Length, Gamma Zone, and Macular Bruch's Membrane Defects. Invest Ophthalmol Vis Sci 2019 Jun 03;60(7):2591-2598. [doi: 10.1167/iovs.19-27331] [Medline: 31219533]
35. Shoji T, Sato H, Ishida M, Takeuchi M, Chihara E. Assessment of glaucomatous changes in subjects with high myopia using spectral domain optical coherence tomography. Invest Ophthalmol Vis Sci 2011 Mar 25;52(2):1098-1102. [doi: 10.1167/iovs.10-5922] [Medline: 21051712]
36. Kim NR, Lee ES, Seong GJ, Kang SY, Kim JH, Hong S, et al. Comparing the ganglion cell complex and retinal nerve fibre layer measurements by Fourier domain OCT to detect glaucoma in high myopia. Br J Ophthalmol 2011 Aug;95(8):1115-1121. [doi: 10.1136/bjo.2010.182493] [Medline: 20805125]
37. Yamashita T, Sakamoto T, Yoshihara N, Terasaki H, Tanaka M, Kii Y, et al. Correlations between local peripapillary choroidal thickness and axial length, optic disc tilt, and papillo-macular position in young healthy eyes. PLoS One 2017;12(10):e0186453 [FREE Full text] [doi: 10.1371/journal.pone.0186453] [Medline: 29023585]
38. Fujino Y, Yamashita T, Murata H, Asaoka R. Adjusting Circumpapillary Retinal Nerve Fiber Layer Profile Using Retinal Artery Position Improves the Structure-Function Relationship in Glaucoma. Invest Ophthalmol Vis Sci 2016 Jun 01;57(7):3152-3158. [doi: 10.1167/iovs.16-19461] [Medline: 27309619]
39. Rho S, Sung Y, Kang T, Kim NR, Kim CY. Improvement of diagnostic performance regarding retinal nerve fiber layer defect using shifting of the normative database according to vessel position. Invest Ophthalmol Vis Sci 2014 Jul 29;55(8):5116-5124. [doi: 10.1167/iovs.14-14630] [Medline: 25074779]

Abbreviations

AUROC: area under the receiver operating characteristic curve

FCN: fully connected network
NPV: negative predictive value
OCT: optical coherence tomography
PPV: positive predictive value
RBFN: radial basis function network
RNFL: retinal nerve fiber layer
RNFLT: retinal nerve fiber layer thickness
ROC: receiver operating characteristic

Edited by G Eysenbach; submitted 21.07.20; peer-reviewed by H Demir, K Koshechkin; comments to author 03.09.20; revised version received 17.09.20; accepted 13.04.21; published 18.05.21.

Please cite as:
Li L, Zhu H, Zhang Z, Zhao L, Xu L, Jonas RA, Garway-Heath DF, Jonas JB, Wang YX
Neural Network–Based Retinal Nerve Fiber Layer Profile Compensation for Glaucoma Diagnosis in Myopia: Model Development and Validation
JMIR Med Inform 2021;9(5):e22664
URL: https://medinform.jmir.org/2021/5/e22664
doi:10.2196/22664
PMID:34003137

Original Paper

Transforming a Patient Registry Into a Customized Data Set for the Advanced Statistical Analysis of Health Risk Factors and for Medication-Related Hospitalization Research: Retrospective Hospital Patient Registry Study

Zhivko Taushanov[1,2*], PhD; Henk Verloo[3,4*], MSc, PhD; Boris Wernli[5*], PhD; Saviana Di Giovanni[3,6], PhD; Armin von Gunten[4], MD; Filipa Pereira[3,7], BSc, MSc

[1]Faculty of Social and Political Sciences, University of Lausanne, Lausanne, Switzerland

[2]Faculty of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland

[3]School of Health Sciences, HES-SO Valais-Wallis, Sion, Switzerland

[4]Service of Old Age Psychiatry, Lausanne University Hospital, Lausanne, Switzerland

[5]FORS, Swiss Centre of Expertise in the Social Sciences, University of Lausanne, Lausanne, Switzerland

[6]Pharmacy Benu Tavil-Chatton, Morges, Switzerland

[7]Institute of Biomedical Sciences Abel Salazar, University of Porto, Porto, Portugal

[*]these authors contributed equally

Corresponding Author:
Filipa Pereira, BSc, MSc
School of Health Sciences
HES-SO Valais-Wallis
Chemin de l'Agasse 5
Sion, 1950
Switzerland
Phone: 41 058 606 84 4
Email: filipa.pereira@hevs.ch

Abstract

Background: Hospital patient registries provide substantial longitudinal data sets describing the clinical and medical health statuses of inpatients and their pharmacological prescriptions. Despite the multiple advantages of routinely collecting multidimensional longitudinal data, those data sets are rarely suitable for advanced statistical analysis and they require customization and synthesis.

Objective: The aim of this study was to describe the methods used to transform and synthesize a raw, multidimensional, hospital patient registry data set into an exploitable database for the further investigation of risk profiles and predictive and survival health outcomes among polymorbid, polymedicated, older inpatients in relation to their medicine prescriptions at hospital discharge.

Methods: A raw, multidimensional data set from a public hospital was extracted from the hospital registry in a CSV (.csv) file and imported into the R statistical package for cleaning, customization, and synthesis. Patients fulfilling the criteria for inclusion were home-dwelling, polymedicated, older adults with multiple chronic conditions aged ≥65 who became hospitalized. The patient data set covered 140 variables from 20,422 hospitalizations of polymedicated, home-dwelling older adults from 2015 to 2018. Each variable, according to type, was explored and computed to describe distributions, missing values, and associations. Different clustering methods, expert opinion, recoding, and missing-value techniques were used to customize and synthesize these multidimensional data sets.

Results: Sociodemographic data showed no missing values. Average age, hospital length of stay, and frequency of hospitalization were computed. Discharge details were recoded and summarized. Clinical data were cleaned up and best practices for managing missing values were applied. Seven clusters of medical diagnoses, surgical interventions, somatic, cognitive, and medicines data were extracted using empirical and statistical best practices, with each presenting the health status of the patients included in it as accurately as possible. Medical, comorbidity, and drug data were recoded and summarized.

Conclusions: A cleaner, better-structured data set was obtained, combining empirical and best-practice statistical approaches. The overall strategy delivered an exploitable, population-based database suitable for an advanced analysis of the descriptive, predictive, and survival statistics relating to polymedicated, home-dwelling older adults admitted as inpatients. More research is needed to develop best practices for customizing and synthesizing large, multidimensional, population-based registries.

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2019-030030

(JMIR Med Inform 2021;9(5):e24205) doi:10.2196/24205

KEYWORDS

cluster analysis; hierarchical 2-step clustering; registry; raw data; hospital; retrospective; population based; multidimensional

Introduction

The transition from paper-based patient records to electronic health records has provided unprecedented access to vast amounts of diverse clinical and health data at the point of care [1]. Undoubtedly, this transition offers a huge opportunity to exploit patient registries for scientific, clinical, and health-policy purposes. An electronic health record is the systematized collection of patients' digitally stored health information. The term *patient registry* is generally used to distinguish registries focused on health information from other data sets, but there is currently no consistent definition in use [2]. The World Health Organization (WHO) describes registries in health information systems as "a file of documents containing uniform health information about individual persons, collected in a systematic and comprehensive way, in order to serve a predetermined purpose" [3]. Properly designed and executed patient registries can provide a real-world view of clinical practice, patient outcomes, safety, and comparative effectiveness [4,5]. Several national registries (eg, the National Committee on Vital and Health Statistics, or the Agency for Healthcare Research and Quality, both in the United States) are used for a broad range of purposes in public health and medicine as part of "an organized system for the collection, storage, retrieval, analysis, and dissemination of information on individual persons who have either a particular disease, a condition (eg, a risk factor) that predisposes the occurrence of a health-related event, or prior exposure to substances (or circumstances) known or suspected to cause adverse health effects" [1]. Other terms used to refer to patient registries are clinical registries, clinical data registries, disease registries, and outcomes registries [5,6]. A patient registry can be a powerful tool for observing the course of a disease, understanding variations in treatment and outcomes, examining factors that influence prognosis, describing care patterns, including the appropriateness of care and disparities in its delivery, assessing effectiveness, monitoring safety and harm, and measuring some aspects of the quality of care [1,6].

National and international statistics document elevated rates of hospitalization and emergency department admissions among polymedicated, home-dwelling older adults with multiple chronic conditions, and these are often caused by medication-related problems (MRPs) [7-10]. However, the determining factors of medication-related hospitalizations are poorly understood and require more investigations based on existing patient data [11]. The associations between age, comorbidities, polypharmacy, and adverse effects on health outcomes and health care consumption have been reported in multiple studies of emergency departments and hospitals, but the underlying mechanisms have often been unclear [12-14]. Several studies have demonstrated that one-quarter of the emergency department admissions for polymedicated, home-dwelling older adults are related to the inappropriate prescription of medicines or unsatisfactory medication management [15,16]. Poor medication management, inappropriate medicine prescription, and drug–drug interactions are frequent causes of admission [17,18]. The risk of MRPs increases not only with old age and comorbidities but also with the number of medications prescribed and with certain classes of medicines, such as medicines for cardiovascular diseases and diabetes [9,19]. The mechanisms behind those high rates of hospitalization in relation to MRPs deserve more attention. More knowledge and understanding of the factors predisposing and precipitating hospitalization and MRPs among polymedicated, home-dwelling older adults are needed too.

This paper aims to describe the method used to transform and synthesize a raw, multidimensional, patient registry data set to prepare it for exploitation as a database with which to examine predictive and survival analysis among hospitalized older inpatients.

Methods

Study Design

This multidimensional, retrospective, patient registry–based study explored the methods required to transform and synthesize a raw data set into a suitable database for further analysis of descriptive, predictive, and survival statistics to identify the risk factors that might induce MRPs among discharged, polymedicated older inpatients.

Population and Sample

The multidimensional patient registry included 140 variables routinely collected during hospital stays by older adult inpatients aged 65 years old or more, living at home before hospitalization, with at least five prescribed medicines at discharge from hospital. The extracted data set was composed of a sample of 20,422 hospitalizations from 2015 to 2018, with similar numbers of annual hospitalizations: 5134, 5095, 5125, and 5068, respectively.

Medicines prescribed before hospital admission were not considered in the analysis due to a lack of data accuracy and validity. Indeed, information on medication at hospital admission is often collected from patients themselves, who may not accurately report their prescriptions, particularly in cases of unplanned hospitalization.

Data Set Extraction and Importing

The hospital data set was extracted from a public teaching hospital's registry, delivered to the investigators in a CSV (.csv) format file via an encrypted email and saved on a secure server. Finally, the data set was imported into the *R* statistical package for cleaning, data transformation, and synthesis [20]. Routinely collected data included information derived from patients' medical and clinical statuses (patient-reported data, clinical examination, medical diagnoses, or medicines prescribed). The data set had to be cleaned up and synthesized to be suitable for analyzing descriptive, predictive, and survival statistics.

Data Cleaning and Transformation

Clinical coding was carried out directly by health care professionals during routine daily care, using a pre-established drop-down menu. Official clinical coding of established medical (10th revision of the International Statistical Classification of Diseases and Related Health Problems [ICD-10]) and surgical diagnostics (CHOP) is mandatory under Swiss Federal Office of Public Health regulations. The variables represented by free text in the original database were excluded.

The distributions of each variable in the data set were explored, according to type (categorical and continuous variables), in order to identify any extreme values and obtain a better view of missing values and associations. Our data cleaning and transformation were guided by a literature review on cleaning-up large data sets, the quantity of information available to us, and the study aim [21]. One major challenge was to find a way to select or summarize a significant volume of information so that further descriptive and predictive statistical analyses could be performed (ie, summarize as many variables as possible, while losing the least amount of information). The large number of variables describing an inpatient's somatic and cognitive status and medical diagnoses represents a significant challenge: we must find a balance between the variability of data and the essential, detailed information they provide without losing the ability to perform descriptive, predictive, and survival analyses [22].

Presentation of the Data Set

Description of the Sociodemographic and Hospitalization Data Set

The sociodemographic data set—almost exclusively composed of ordinal variables—included just 2 categorical variables (sex and place of discharge) and 1 continuous variable (age). There were no missing sociodemographic variables except among the place-of-discharge data.

The hospitalization data set included 2 continuous variables (date of entry and discharge) and 1 categorical variable (the personal identification data number [PID]). These 3 variables enabled us to compute the length of stay (LOS) and the frequency of hospitalization and rehospitalization, respectively. Rehospitalization rates were important health status indicators in relation to drug prescriptions. Many polymedicated, home-dwelling older adults were hospitalized more than once during the 4-year study period. Almost one-third (n=3678) of older inpatients were rehospitalized 3 times or more; a small fraction was hospitalized more than 9 times. We found 18 polymedicated, home-dwelling older adults who were rehospitalized 17 times and considered them as outliers. Besides computing the average age and hospital LOS, no other interventions were necessary to clean up this section of the data set. Our analyses found an almost equal distribution of men and women, with an average age close to 79 (SD 7.7). Most older inpatients were discharged home after an average LOS of about 10 days (Multimedia Appendix 1).

Description of the Somatic Data Set

Nurses routinely collect clinical data during hospitalization using a drop-down menu, and the data set was composed of 18 categorical variables: 16 measured as ordinal variables (mobility, changing position, falls in the last year, exhaustion, upper- and lower-body care, upper- and lower-body [un]dressing, eating, drinking, micturition and defecation-related movements, hearing, vision, verbal expression, and pain intensity) and 2 measured as nominal variables (altered gait and chronic pain). Missing values in the data set were resolved by recoding them as "not available" (NA; Multimedia Appendix 2).

Description of the Cognitive Data Set

Inpatients' cognitive status was measured at an ordinal level using 5 categorical variables. More than 72.60% (14,826/20,422) of adults showed no deterioration in their cognitive status (Multimedia Appendix 3).

Description of the Medical Diagnoses and Surgical Interventions Data Set

This data set of medical information was composed of patients' principal medical diagnosis and 4 secondary medical diagnoses (active or passive comorbidities), based on the WHO's ICD-10 adopted by Switzerland's health care system [23]. This was completed with the patient's principal surgical intervention and 4 additional surgical interventions, based on Switzerland's surgical classification system (named CHOP) [24]. This data set showed no missing values (Multimedia Appendix 4).

The data set has no specific coding for MRPs (the corresponding ICD-10 is "Poisoning by drugs, medicaments and biological substances") [25].

Description of the Prescribed Medicines Data Set

The hospital data set showed that discharged patients had been prescribed 2370 different medicines. This huge number of medicines and their heterogeneous therapeutic focus needed a structured classification built based on best practices (Multimedia Appendix 5). Based on expert opinion and a literature review on medicine classification systems, we chose the Anatomical Therapeutic Chemical (ATC) classification system's 14 top-level codes to structure the set of prescribed medicines [25,26] (Multimedia Appendix 6).

Synthesizing the Raw Data Set

Summarizing the data set was especially challenging because most of the variables documented different parts of inpatients' overall health status, with all the diverse dimensions of their somatic and cognitive conditions. Special attention was given to the large data set of prescribed medicinal treatments. In many fields, the most common means of coping with such difficulties is the use of statistical clustering, a technique which combines all the available information (all variables) to reveal one or several underlying dimensions or health concepts.

In addition, the data set's large number of variables and dimensions made it extremely complex to investigate the relationships and interactions between the different somatic and cognitive variables. The data set should allow the analysis of the risks of adverse health outcomes and their relationships with the medicines prescribed. For this reason, computing every variable in the same model may not be the optimal modeling choice if we consider the multidimensional aspect and dependency between those variables. This is especially true if these variables are significant ($P<.01$) for the discrimination and discovery of mechanisms leading to rehospitalization and a nonreturn home due to medical conditions and MRPs. In the absence of any scientific models, this study used an empirical approach.

Data Clustering

Overview

Little research to date has explored specific combinations or clusters of clinical data and health status. Our study's objective was to transform and synthesize valuable inpatient health information (health concepts such as mobility), rather than to reduce the dimensions of the data. It is, therefore, worth considering a larger number of principal components in the analysis to explain a larger part of the data variability. Almost all the studies which have examined specific comorbidities start from a specific disease rather than examining all the co-occurring clinical and medical conditions [27,28]. Nosology clusters groups of diseases, disorders, or syndromes with meaningful associations into a type of classification, so that diseases, for example, within a cluster, are very similar to one another, but are dissimilar to diseases in other clusters [29]. Among older inpatients, some associations are useful for identifying those at risk of in-hospital adverse clinical events and death in relation to those disease or health-syndrome clusters.

A large variety of clustering methods exist in the literature. However, the majority are focused on either continuous or nominal data alone. Only a limited number of techniques and strategies manage to incorporate both variable types into the same clusters [30].

Distance Measurement

This approach aims to create a measure of the distance between individuals or sequences that includes nominal and continuous variables. The Gower distance is the most widely used distance measure, and it can be used to calculate the distance between 2 entities whose shared attribute has a mixture of categorical and numerical values [31]. However, because it uses a range of continuous variables to determine the distance and assumes that nominal variables have a distance of either 0 or 1, the Gower distance may underestimate the impact of continuous variables because they are valued at 1 much less often than nominal variables are. Furthermore, weightings are selected arbitrarily. However, they define each data type's contribution to the overall distance. As with all distance measures, the Gower distance should be used as an input for clustering methods, such as k-means.

K-Means Method

The k-means algorithm is mainly used for continuous variables [32]. Several other applications, such as the *R* statistical package KAMILA [33], integrate different types of variables. In this case, it uses the probabilities of a multinomial distribution for the discrete variables. The continuous variable distribution is estimated using univariate kernel densities [34]. The probabilities resulting from both distribution types are added together to obtain a measure of how close an observation is to the center of each cluster.

K-Medoids Method

The k-medoids method is a more robust version of k-means [35]. The difference is that in k-medoids real data points are selected as cluster centers, whereas in k-means the centers are the computed averages. The PAM function in the *R* statistical package KAMILA is a popular application of this approach [33,34].

Multiple Correspondence Analysis

The standard method for clustering factor variables is multiple correspondence analysis [36]. This model is implemented in the FactoMineR and PCAmixdata *R* packages. It splits all factors into multiple binary variables and applies a type of principal component analysis. The principal components obtained are then usually clustered using a k-means algorithm.

Hierarchical Cluster Analysis

Our data analysis strategy applied a hierarchical cluster analysis, using the ClustOfVar R package [37,38]. As with any statistical analysis, results of a hierarchical cluster should not be accepted as they first appear, but should be taken as suggestions or questioned instead. When the final set of groups of variables was defined, a statistical model to cluster the individuals within each group was applied. This created one new variable for each group, indicating the type of characteristics the individual displayed in his/her health status assessment. For example, if we separate the individuals into 3 groups according to their cognitive status, we might obtain a variable indicating that a person belongs to a group with significant, minor, or no cognitive impairment. This type of aggregated variable was used in our final analysis of risk factors.

Our analysis explored several different clustering methods. However, the results displayed here most often used the following variable clustering procedure. First, a one-factor analysis model was typically used; second, the most important latent factors were selected. At this stage, it was essential to obtain accurate clustering rather than reduce the dimensionality, which takes place in the final cluster partition. Third, these factors were considered as variables and served as the input to a k-means clustering algorithm. Finally, the number of clusters was then selected using the Rousseeuw silhouette statistic, also with regard to the interpretability of the resulting partition [39].

Two-Step Clustering Framework

In this approach, *n* and *p* denote the numbers of the patients and health conditions (indicators), respectively. The data can thus be represented by an $n \times p$ matrix, where the observed value for the *i*th column and the *j*th row of the data matrix is 1 or 0, indicating the presence or absence of the *i*th health condition for the *j*th respondent ($i = 1,\ldots, p$; $j = 1,\ldots, n$).

In the 2-step clustering approach, step 1 involves clustering the *p* conditions into non-overlapping groups of clinical or health conditions. Based on individual patterns in these groups of clinical and medical conditions, step 2 involves clustering the *n* respondents into clusters which correspond to different patterns of clinical or health conditions.

To thoroughly analyze the data and identify the MRPs leading to adverse health outcomes—such as rehospitalization, nonreturn home, and early death [40,41]—among older adult inpatients, a literature review was conducted [27].

Treatment of Missing Data

As in every real-life data collection exercise, missing values are unavoidable, and it is important to define how these are integrated into the study. Four approaches were considered: ignoring all observations with 1 or more missing values; defining "NA" as a separate potential variable value; replacing every missing value by the mode of the corresponding variable; or performing multiple imputations on the data set. The first approach was obviously inappropriate, especially in cases where the number of missing data was significant ($P<.01$). Considering NA as a separate modality for each variable inflates the number of modalities, but it reduces the possibility of bias due to incorrect imputation methods. Nevertheless, for the sake of comparison, it was also tempting to consider the 2 latter approaches. Before choosing between simple replacement using the variable's mode value and multiple imputation, we had to test for the type of missing data. If data are missing completely at random, we can simply impute using the mode. However, if this possibility is rejected, multiple imputation is theoretically more appropriate. The Little test (1988) [42] examines the null hypothesis H0: the data are missing completely at random. This test was applied to all subclusters of variables and the null hypothesis was rejected for every data set. This indicated that multiple imputation could be performed as an optional solution for estimating missing values.

Finally, defining NA values became our primary choice for the treatment of missing values. By creating an NA variable (an empty variable that does not influence the cluster result), all observations with an NA variable were still taken into account in the cluster analyses. This is why each cluster analysis contains every hospitalization (N=20,422).

Ethical Considerations

The hospital data set was coded and its use was contractually limited by the participating hospital center. Furthermore, because the data sets included highly sensitive electronic patient records from a hospital registry, ethical approval was sought before any synthesis or analysis. Data were stored on a dedicated secure data server, which included a log registry. Each access flow to the secure data environment was documented, and each change required approval. Only users working on the project and requiring access to the data were allowed to use the selected multifactor authentication mechanism in the secure environment. The Human Research Ethics Committee of the Canton of Vaud (CER-VD) (2018–02196) approved the study on February 1, 2019.

Results

Transformation of the Data Set

The original data set required some adjustments before our plan of analysis could move forward. Four empty variables and 1 observation containing mostly 0 or unavailable values were removed from the data set. The labels for all variables were rewritten and clarified, and many medicine names in French had accents and unreadable symbols corrected.

Missing Data

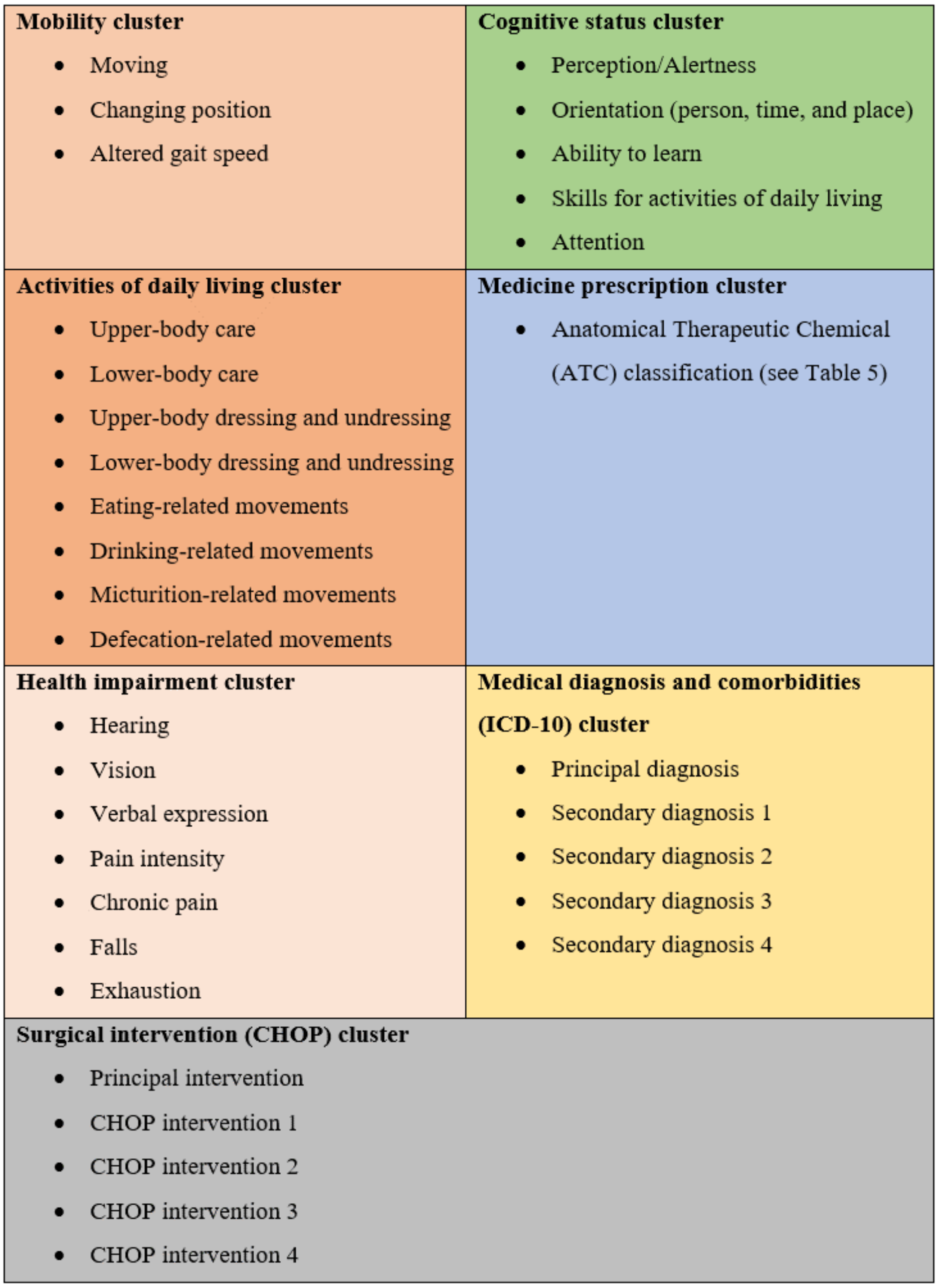
Tests made using both the BaylorEdPsych and RBtest *R* packages confirmed that the missing-completely-at-random hypothesis could be rejected [42]. Observations within each subcluster of the data set that only contained missing values were recoded as NA. Their presence might have been due to incorrect inputs, human or software error, or unavailable parts of some questionnaires. Missing data had very little impact on the sample size, appeared to be random, and concerned the first 4300 observations, especially. After recoding these observations, the cognitive status variables showed no more separate missing observations, and we had a complete data set.

Clustering of Clinical and Medical Data

Most of the hospital variables were partially independent and gathered into several groups according to the dimension of the patient's measured/assessed clinical and medical status. We used an empirical approach suggested by health care experts (FP, HV, and AvG) in an attempt to present homogenous groups within the set of variables. In cases involving clear and meaningful clustering, we relied on expert recommendations or opinions taken from a comprehensive literature review [27,33]. However, when evidence was scarce, we clustered variables using statistical methods. The results from statistical methods were compared against those from expert opinion, which served as a validation tool for addressing any possible subjectivity in those expert opinions [27,33].

Seven groups of clusters were developed: somatic/physical health conditions (3 orange groups in Figure 1), cognitive health conditions (green textbox in Figure 1), total number of prescribed medications based on the ATC classification, diagnoses based on the ICD-10 (yellow textbox in Figure 1), and the surgical interventions based on CHOP (gray textbox in Figure 1). Besides these more apparent distinctions between variables, other underlying subclusters may be present within these groups. This point is beyond the scope of this paper, however, and will be documented elsewhere with a complementary, within-group analysis (the presence of an interpretable clustering of variables within a group before clustering individuals). An examination of the *place of discharge* variable confirms this: of 20,422 hospitalizations, only 131 patients (<1%) were documented to have died during hospitalization. Bearing in mind that there was no explicit variable indicating this worst outcome, we developed indicators that were suggestive of imminent death or a highly and irreversibly deteriorated health condition. Based on a literature review of polymorbidity, 6 clinical indicators from the data set were associated with a functional deterioration leading to progressive decline and poor health status [43]: (1) restricted mobility, (2) incapacity to change position, (3) altered alertness, (4) altered orientation, (5) altered gait, and (6) reduced or absent cognitive skills necessary to carry out the activities of daily living. Each of these variables indicated a deteriorating health status. To ensure that only severely deteriorating health problems were captured, we only considered patients to be endangered if they had multiple problems. We therefore created a variable indicating the number of problems present, with values ranging from 0 to 6 (Multimedia Appendix 7). More than half of the sample presented with at least one deteriorated health condition. However, only a small fraction of the older adult patients had 4 or more deteriorated health conditions at discharge.

Figure 1. Structure and content of the data set clusters.



Mobility cluster

- Moving
- Changing position
- Altered gait speed

Cognitive status cluster

- Perception/Alertness
- Orientation (person, time, and place)
- Ability to learn
- Skills for activities of daily living
- Attention

Activities of daily living cluster

- Upper-body care
- Lower-body care
- Upper-body dressing and undressing
- Lower-body dressing and undressing
- Eating-related movements
- Drinking-related movements
- Micturition-related movements
- Defecation-related movements

Medicine prescription cluster

- Anatomical Therapeutic Chemical (ATC) classification (see Table 5)

Health impairment cluster

- Hearing
- Vision
- Verbal expression
- Pain intensity
- Chronic pain
- Falls
- Exhaustion

Medical diagnosis and comorbidities (ICD-10) cluster

- Principal diagnosis
- Secondary diagnosis 1
- Secondary diagnosis 2
- Secondary diagnosis 3
- Secondary diagnosis 4

Surgical intervention (CHOP) cluster

- Principal intervention
- CHOP intervention 1
- CHOP intervention 2
- CHOP intervention 3
- CHOP intervention 4

Cognitive Data Cluster

Overview

The cognitive data cluster (green textbox in Figure 1) was composed of 5 variables indicating cognitive status level (Table 1). As with many other variables in the total data set, cognitive data were considered nominal because they each had a small number of modalities. The first 400 observations in the data set were excluded from the cognitive status analysis because they contained only missing values and were excluded from other analyses for the same reason. These missing values were explained by the fact that new data variables were introduced into the hospital register during the first semester of 2015.

Cognitive Status Clustering

The *R* ClustOfVar package was used to perform a hierarchical clustering of the cognitive health variables to investigate any possible relationships and the presence of subclusters within these variables. The results did not suggest any clear interpretable structure within the variables included, as illustrated by the dendrogram (Figure 2). They indicated that only single-variable clusters (singletons) could be separated, one at a time, to form separate and not very distinct clusters. This information failed to provide any useful solution to our problem because it makes no sense to cluster individuals using a single variable. This result, combined with the small total number of 5 other data set clusters, led us to the conclusion that the 6 data set clusters illustrating different cognitive conditions should be considered together in the same clustering algorithm.

Multiple correspondence analysis was used to cluster individuals according to their cognitive status because all the variables were categorical. Even though the first 2 principal components do not explain much of the data (5310/20,422, 26.00%), we were able to discern the 4 most discriminant variables for clustering (and the importance of their categories). For further analysis, we selected numerous principal components (n=9) because of their relatively low explanatory power (65% of the variance). We found multiple different clustering partitions with respect to the number of clusters. Some groups and features were found systematically in all the partitions. This enabled us to make the following generalizations about the results, regardless of the number of clusters:

- The majority of observations indicated that cognitive status was not altered at the time of the assessment. We found a good solution and form in every cluster, including the largest cluster.
- When increasing the number of clusters, observations with average or poor cognitive status were split and nuanced.
- One group of individuals with mainly missing values was excluded from the analysis.

The optimal number of clusters was determined using the silhouette statistic (Figure 3). For each number of clusters, this statistic measures how similar each observation is to its own cluster in comparison to all other clusters, that is, the extent to which observations are grouped together. The results indicated that the 3-cluster solution would be the most appropriate in terms of within- and between-cluster distances. However, a partition using 2 clusters provided greater simplicity and also had a statistically sustainable silhouette value.

Figure 2. Dendrogram of cognitive status variables.

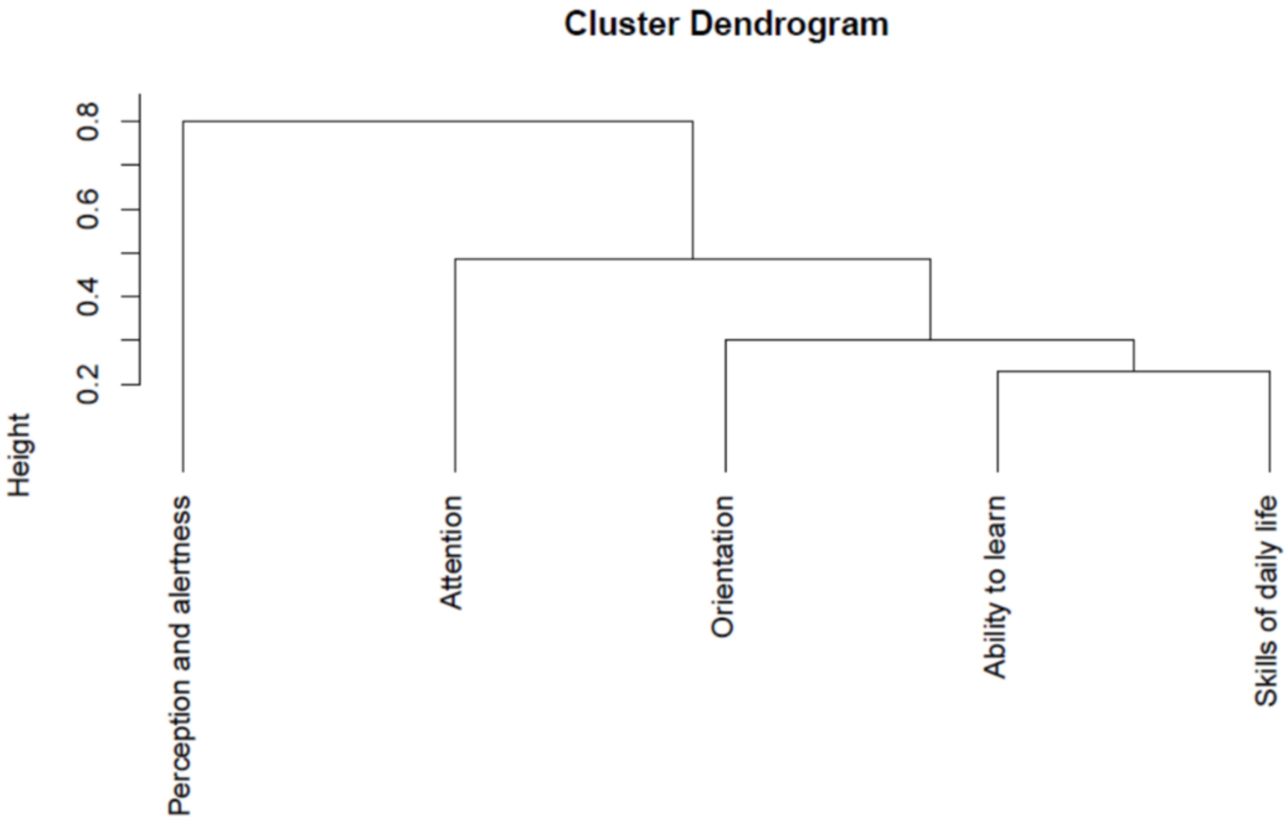
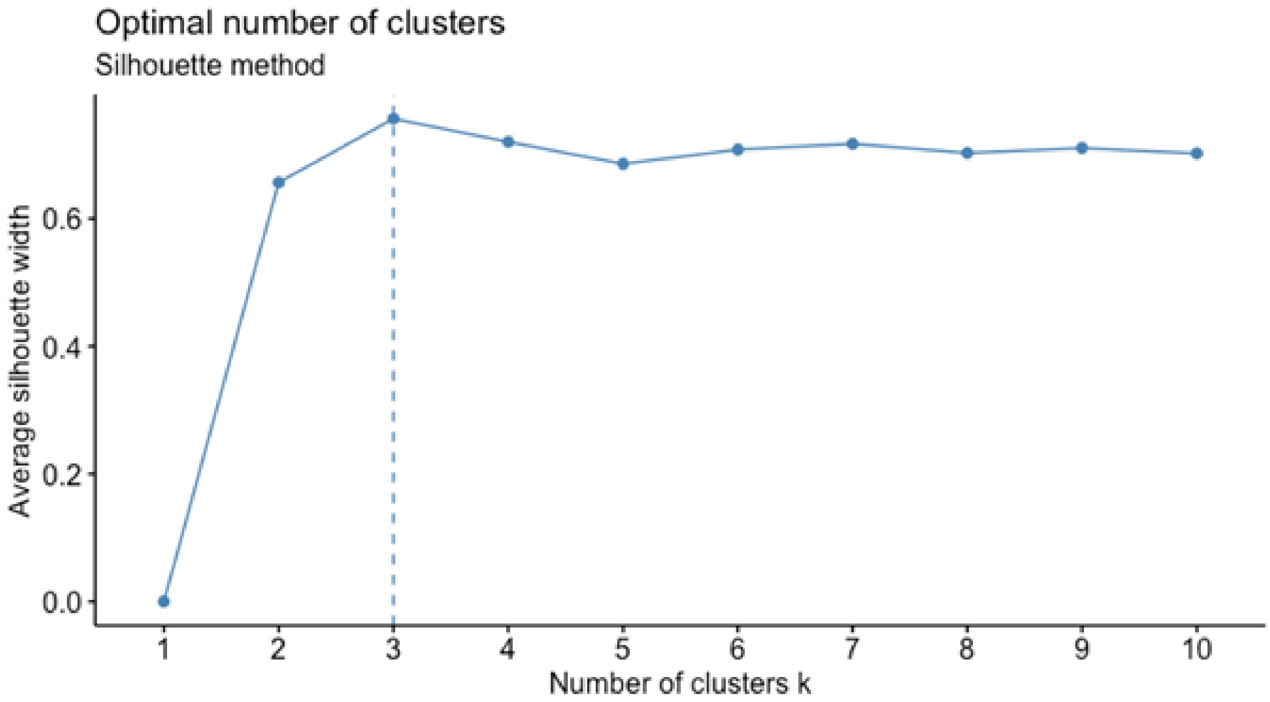


Figure 3. Silhouette statistics for choosing the optimal number of clusters: the two- or four-cluster solutions were suggested.



Two-Cluster Solution

Hierarchical clustering using 2 classes created a dominant group of 18,339/20,422 (89.80%) older inpatients with full cognitive ability and a smaller group of 2083/20,422 (10.20%) inpatients with cognitive impairment. The 2-cluster solution was differently distributed over the 5 variables and according to the type of diagnoses (ICD-10; Table 1), and it was highly significant (*P*<.001). Two other variables (number of medications prescribed and primary diagnosis) were added to the analysis for experimental purposes but were not included in the clustering model. A difference was observed in the average number of medications prescribed (9.63 vs 10.47; *P*<.001) between groups, and the primary diagnosis also appeared to be different (0.10 vs 0.08; *P*<.001; Table 1).

Table 1. Distribution of individuals in each group for all 5 cognitive status variables in the 2-cluster solution (N=20,422).

Cognitive status variables	Cognitive status	
	Full ability	Cognitive impairment
Perception/Alertness [a]		
Alert	1.00	0.85
Drowsy	0.00	0.13
Stupor	0.00	0.01
Coma	0.00	0.01
NA[b]	—	—
Distribution, n (%)	18,318 (89.70)	2083 (10.20)
Orientation[a]		
Full ability	0.91	0.11
3 abilities	0.08	0.24
1–2 abilities	0.01	0.40
Inability	0.00	0.20
NA	0.00	0.06
Distribution, n (%)	18,319 (89.70)	2083 (10.20)
Ability to learn[a]		
Full ability	0.81	0.02
Slightly reduced	0.18	0.10
Severely reduced	0.02	0.67
Inability	0.00	0.21
NA	—	—
Distribution, n (%)	18,319 (89.70)	2083 (10.20)
Activities of daily living[a]		
Full ability	0.83	0.03
Slightly reduced	0.15	0.16
Severely reduced	0.02	0.66
Inability	0.00	0.13
NA	0.00	0.01
Distribution, n (%)	18,319 (89.70)	2083 (10.20)
Attention		
Unaffected	0.98	0.36
Reduced	0.02	0.63
NA	0.00	0.01
Distribution, n (%)	18,319 (89.70)	2083 (10.20)
Number of medicines[a]		
Average number	9.63	10.47
ICD-10[c] **main diagnoses**[a]		
Systems	0.52	0.54
Mental	0.10	0.08
Cancers	0.01	0.01
Other	0.37	0.37

Cognitive status variables	Cognitive status	
	Full ability	Cognitive impairment
NA	—	—
Distribution, n (%)	18,339 (89.80)	2083 (10.20)

[a] Variables significantly different among clusters (χ^2 tests and *t* tests, *P*<.01). Each line represents 1 cluster and adds up to 1 (100%).

[b] NA: not available.

[c] ICD-10: 10th revision of the International Statistical Classification of Diseases and Related Health Problems.

Three-Cluster Solution

Hierarchical clustering using 3 classes created groups of 15,717/20,422 (76.96%) polymedicated older inpatients in full cognitive health, 4290/20,422 (21.01%) with mild cognitive impairment, and 415/20,422 (2.03%) with severe cognitive impairment. The 3-cluster solution’s results were similar to those of the 2-cluster solution (Table 2).

Table 2. Distribution of individuals in each group for all 5 cognitive status variables in the 3-cluster solution (N=20,422).

Cognitive status variables	Cognitive status		
	Full ability	Mild cognitive impairment	Severe cognitive impairment
Perception/Alertness[a]			
Alert	1.00	0.93	0.61
Drowsy	0.00	0.07	0.29
Stupor	0.00	0.07	0.06
Coma	0.00	0	0.04
NA[b]	—	—	—
Distribution, n (%)	17,855 (87.43)	2166 (10.61)	380 (1.86)
Orientation[a]			
Full ability	0.94	0.10	0.03
3 abilities	0.06	0.39	0.05
1–2 abilities	0.00	0.41	0.12
Inability	0.00	0.08	0.62
NA	0.00	0.02	0.18
Distribution, n (%)	17,856 (87.44)	2166 (10.61)	380 (1.86)
Ability to learn[a]			
Full ability	0.83	0.03	0.01
Slightly reduced	0.17	0.23	0.03
Severely reduced	0.01	0.70	0.09
Inability	0.00	0.05	0.87
NA			
Distribution, n (%)	17,856 (87.44)	2166 (10.61)	380 (1.86)
Activities of daily living[a]			
Full ability	0.85	0.06	0.01
Slightly reduced	0.13	0.29	0.02
Severely reduced	0.02	0.63	0.32
Inability	0.00	0.02	0.62
NA	0.00	0.00	0.03
Distribution, n (%)	17,856 (87.44)	2166 (10.61)	380 (1.86)
Attention[a]			
Unaffected	0.99	0.49	0.11
Reduced	0.01	0.51	0.84
NA	0.00	0.00	0.04
Distribution, n (%)	17,856 (87.44)	2166 (10.61)	380 (1.86)
Number of medicines[a]			
Average number	9.62	10.43	10.35
ICD-10[c] **main diagnoses**[a]			
Systems	0.52	0.54	0.57
Mental	0.10	0.07	0.09
Cancers	0.01	0.01	0.00

Cognitive status variables	Cognitive status		
	Full ability	Mild cognitive impairment	Severe cognitive impairment
Other	0.37	0.38	0.33
NA	—	—	—
Distribution, n (%)	17,876 (87.53)	2166 (10.61)	380 (1.86)

[a] Variables significantly different among clusters (χ^2 tests and *t* tests, *P*<.01). Each line represents 1 cluster and adds up to 1 (100%).

[b] NA: not available.

[c] ICD-10: 10th revision of the International Statistical Classification of Diseases and Related Health Problems.

Somatic Variables and Their Clustering Into Subclusters

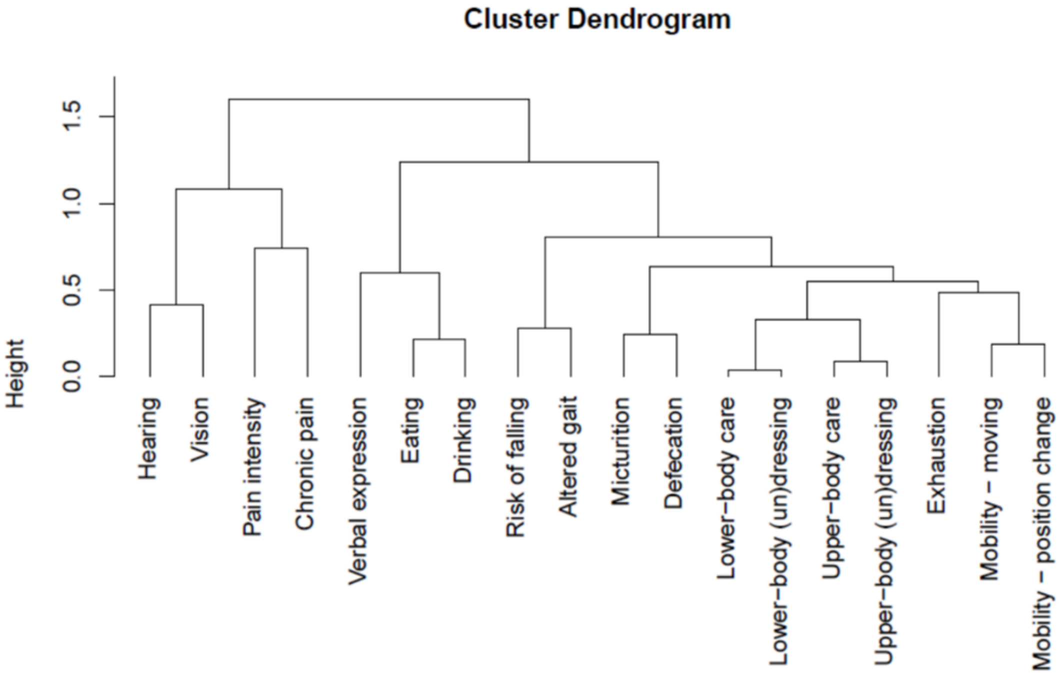
Multiple variables showed modalities that did not correspond exactly to those described in the list (Multimedia Appendices 1-6). The *risk of falling* variable in the list of somatic data (orange textbox, Figure 1) is continuous, and it was thus recoded into a 3-modality factor as no risk (0 falls), moderate risk (1-4 falls), and high risk (≥5 falls in the last year).

The number of somatic variables is large and heterogeneous, making the direct clustering of individuals challenging. We considered the hypothesis that there were probably dissimilarities in this whole set of somatic variables, and starting from this assumption, we split the variables into subclusters.

In the absence of any validated techniques, tools, or evidenced-based literature, we developed an empirical subcluster clustering strategy. The initial separation of the variables was guided by information retrieved from a literature review of communicable somatic diseases completed with the authors' experiences and expertise in patterns of somatic illness [27,28]. Four subclusters of somatic variables were constructed: mobility, health difficulties, capacities for the activities of daily living, and other health risks (orange textbox in Figure 1). The mobility subcluster was composed of the clinical variables of movement, changing position, altered gait, balance disorders, and past and recent falls. The general health status subcluster included exhaustion, hearing, vision, verbal expression, drowsiness, sleep rhythm, sleep impairment, pain intensity, and chronic pain. The capacities for the activities of daily living subcluster were composed of upper- and lower-body care, upper- and lower-body (un)dressing, eating, drinking, and micturition- and defecation-related movements. The other health risks subcluster was composed of clinical variables assessing the risks of sores, wounds, malnutrition, and falling during hospitalization. To reinforce the authors' opinions, a statistical validation model of the variable clustering was computed using the hierarchical clustering functions of the *R* ClustOfVar package (Figure 4).

Findings showed some differences between the authors' opinions and the statistical model. To optimize the composition of somatic health status variable subclusters, an adapted version was selected for further data analysis following discussion and a consensus agreement. Three subclusters of somatic variables were considered. The mobility subcluster was composed of the movement, changing position, and altered gait variables. The general health impairments subcluster included exhaustion, hearing, vision, verbal expression, risk of falling, chronic pain, and pain intensity. The capacities for the activities of daily living subcluster included upper- and lower-body care, upper- and lower-body (un)dressing, eating, drinking, and micturition- and defecation-related movements.

Figure 4. Dendrogram of the somatic health status variables.



Grouping Individuals Within the Somatic Health Status Subcluster

After separating the variables, the somatic health status subclusters of mobility, health impairments, and capacities for the activities of daily living were themselves partitioned, with the aim of discovering any possible underlying groupings of inpatients.

Mobility Subcluster

Using the silhouette statistic failed to give a clear optimal number of subgroupings n (Figure 5).

Our analysis demonstrated similar and increasing average silhouette widths as n increased. Consequently, we chose a 2-cluster partition, deciding that this best separated the variables in terms of interpretability of results and a clear implicit difference between the groups: a grouping of persons with mostly full mobility (n=12,540) and a grouping with an impaired mobility status (n=7,880). Roughly two-thirds of individuals had few or no mobility problems (Table 3). The remaining individuals exhibited problems in at least one of the three variables. That number is large but not surprising when considering the sample population’s advanced age. The χ^2 tests confirmed a clear difference between the groups across all variables (Table 3). Our analysis highlighted that the group with full mobility status was prescribed significantly fewer medications (P<.01) than the group with impaired mobility (9.07 vs 10.74).

Figure 5. Average silhouette width for each number of sub-clusters in the mobility sub-cluster.

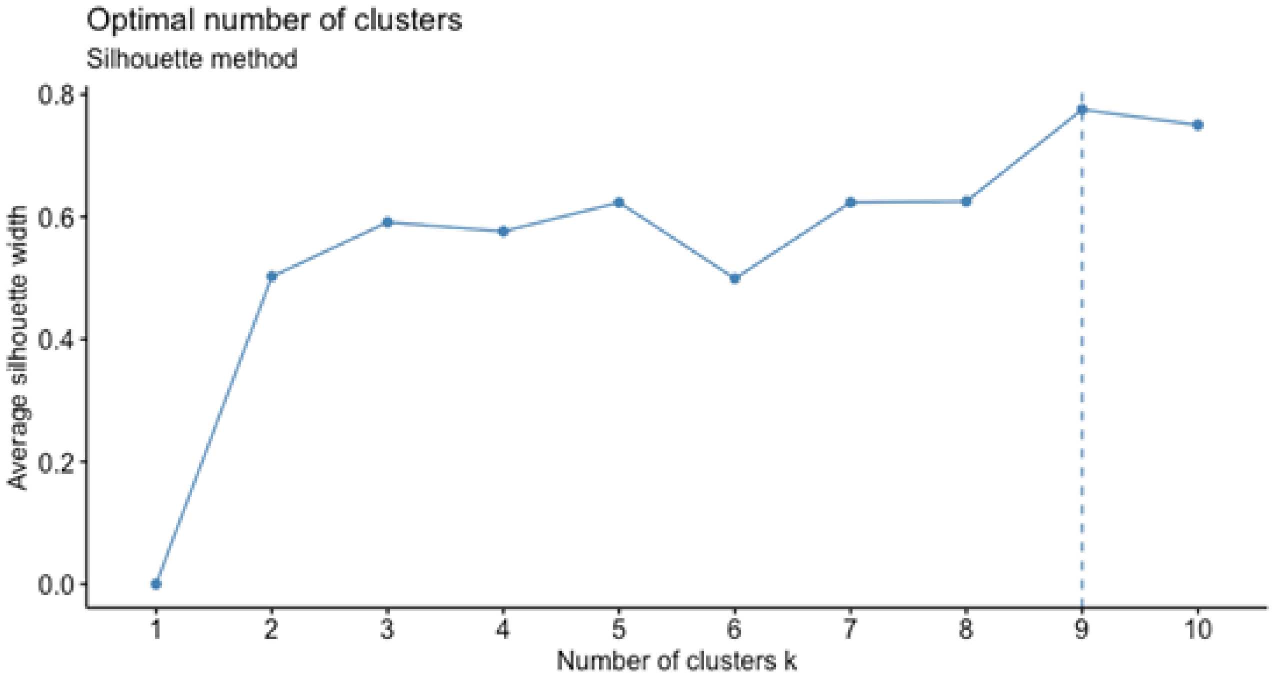


Table 3. Distribution of individuals in the 2-cluster solution for all mobility variables (N=20,422).

Mobility variables	Mobility status	
	Full mobility	Poor mobility
Movement[a]		
Full ability	0.90	0.01
Slightly reduced	0.09	0.61
Severely reduced	0.00	0.30
Inability	0.00	0.08
Distribution, n (%)	12,540 (61.40)	7878 (38.58)
Changing position[a]		
Full ability	0.99	0.25
Slightly reduced	0.01	0.51
Severely reduced	0.00	0.21
Inability	0.00	0.04
Distribution, n (%)	12,540 (61.40)	7878 (38.58)
Altered gait speed[a]		
No	0.85	0.13
Yes	0.15	0.82
Not available	0.00	0.06
Distribution, n (%)	12,540 (61.40)	7878 (38.58)
Number of medicines[a]		
Average number	9.07	10.74

[a] Variables significantly different among clusters (χ^2 tests and *t* tests, *P*<.01). Each line represents 1 cluster and adds up to 1 (100%).

Health Impairments Subclusters

Calculating the silhouette statistic suggested that the 4-cluster groupings solution was optimal, even though the results appear very surprising. However, we decided on the 2-grouping solution, mainly because it is easier to interpret (Figure 6 and Table 4).

Figure 6. Health impairments sub-cluster: silhouette statistics for choosing the number of groupings suggested the four-cluster grouping solution.

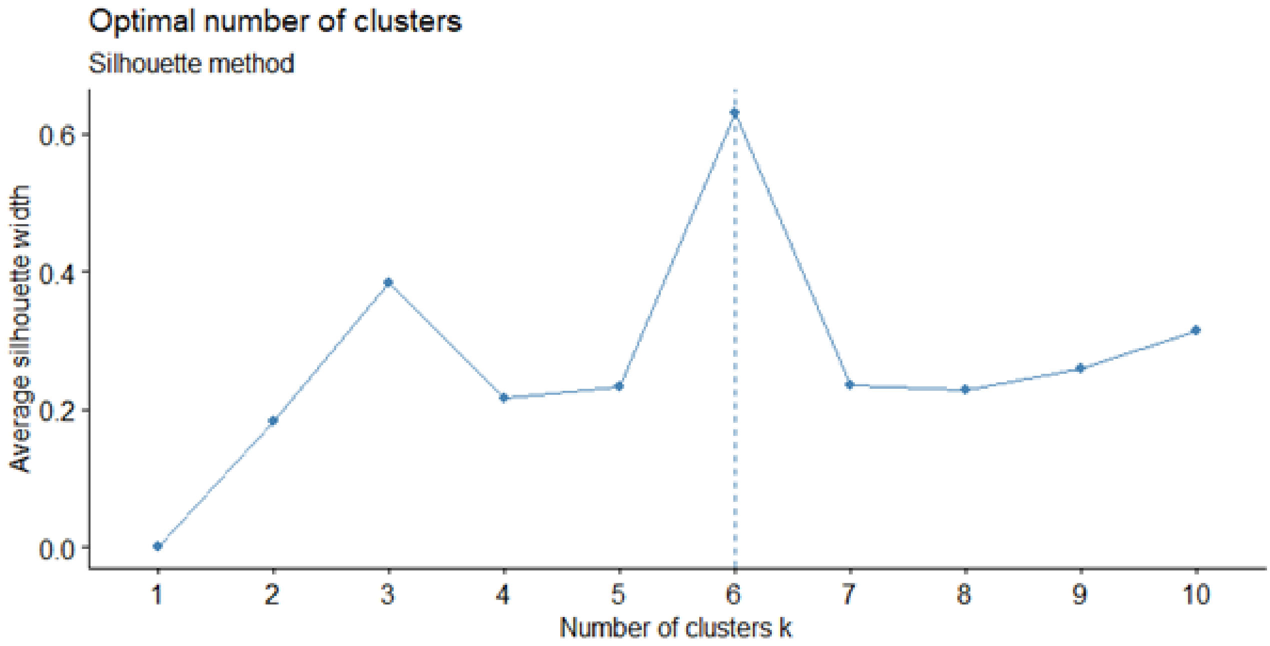


Table 4. Distribution of individuals in the 2-cluster solution for all health impairment variables (N=20,422).

Health impairment variables	Health status	
	Good health status	Impaired health status
Hearing[a]		
Full ability	0.88	0.77
Auditive problems	0.12	0.22
Deaf	0.00	0.10
Distribution, n (%)	17,897 (87.64)	2465 (12.07)
Vision[a]		
Full ability	0.92	0.73
View problems	0.08	0.27
Blind	0.00	0.01
Distribution, n (%)	17,897 (87.64)	2465 (12.07)
Verbal expression[a]		
Full ability	1.00	0.49
Limited capacity	0.00	0.47
Incapacity	0.00	0.04
Distribution, n (%)	17,898 (87.64)	2465 (12.07)
Risk of falling[a]		
No risk	0.37	0.05
Moderate risk	0.63	0.34
High risk	0.00	0.61
Distribution, n (%)	17,844 (87.38)	2464 (12.07)
Chronic pain[a]		
No pain	0.90	0.84
Pain	0.10	0.15
Not measurable	0.00	0.01
Distribution, n (%)	17,872 (87.51)	2462 (12.06)
Pain intensity[a]		
No pain	0.08	0.13
Improbable	0.26	0.29
Low	0.01	0.01
Moderate	0.00	0.01
Intense	0.00	0.01
Pain index	0.65	0.55
Distribution, n (%)	17,880 (87.55)	2462 (12.06)

[a] Variables significantly different among clusters (χ^2 tests and *t* tests, *P*<.01). Each line represents 1 cluster and adds up to 1 (100%).

Capacities for the Activities of Daily Living Subcluster

The 2-cluster solution appeared appropriate and confirmed the silhouette statistic, which highlighted the 2, 8, and 10-cluster solutions (Figure 7). We distinguished 1 large cluster grouping of 17,836/20,422 (87.34%) individuals composed of mainly *autonomous* inpatients with almost full capacity to carry out the majority of the activities of daily living. The second cluster grouping of more *dependent* inpatients included 2573/20,422 (12.60%) individuals with at least one serious problem in handling their activities of daily living. Overall, the partitioning into 2 cluster groupings was relevant in light of our aim to demonstrate that the observations were significantly different

(*P*<.01) among the overall variables and in relation to the number of prescribed medications (Table 5).

Figure 7. Silhouette statistics for the sub-cluster of capacities for the activities of daily living.

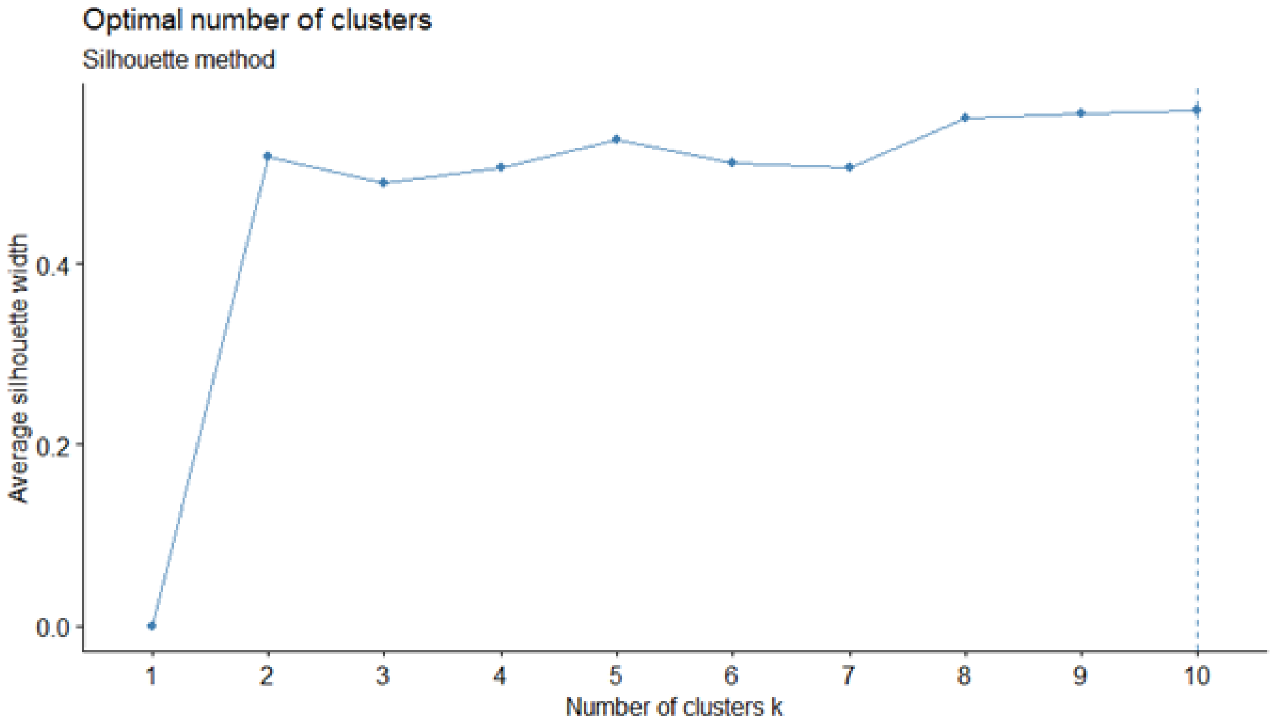


Table 5. Distribution of the capacities for the activities of daily living subcluster (N=20,422).

Activities of daily living	Subclusters	
	Autonomous grouping	Dependent grouping
Upper-body care[a]		
Full capacity	0.77	0.03
Slightly reduced	0.21	0.24
Severely reduced	0.02	0.47
Incapacity	0.00	0.26
Distribution, n (%)	17,836 (87.34)	2573 (12.60)
Lower-body care[a]		
Full capacity	0.61	0.00
Slightly reduced	0.25	0.01
Severely reduced	0.12	0.18
Incapacity	0.01	0.81
Distribution, n (%)	17,836 (87.34)	2573 (12.60)
Upper-body (un)dressing[a]		
Full capacity	0.80	0.01
Slightly reduced	0.18	0.16
Severely reduced	0.02	0.44
Incapacity	0.00	0.39
Distribution, n (%)	17,836 (87.34)	2573 (12.60)
Lower-body (un)dressing[a]		
Full capacity	0.64	0.00
Slightly reduced	0.22	0.01
Severely reduced	0.12	0.17
Incapacity	0.02	0.82
Distribution, n (%)	17,836 (87.34)	2573 (12.60)
Eating-related movements[a]		
Full capacity	0.95	0.35
Slightly reduced	0.05	0.38
Severely reduced	0.00	0.15
Incapacity	0.00	0.12
Distribution, n (%)	17,836 (87.34)	2573 (12.60)
Drinking-related movements[a]		
Full capacity	0.97	0.56
Slightly reduced	0.02	0.25
Severely reduced	0.00	0.12
Incapacity	0.00	0.08
Distribution, n (%)	17,836 (87.34)	2573 (12.60)
Micturition-related movements[a]		
Full capacity	0.85	0.12
Slightly reduced	0.11	0.19
Severely reduced	0.01	0.27

Activities of daily living	Subclusters	
	Autonomous grouping	Dependent grouping
Incapacity	0.02	0.42
Distribution, n (%)	17,836 (87.34)	2573 (12.60)
Defecation-related movements[a]		
Full capacity	0.88	0.18
Slightly reduced	0.10	0.19
Severely reduced	0.02	0.33
Incapacity	0.01	0.3
Distribution, n (%)	17,836 (87.34)	2573 (12.60)
Number of medicines[a]		
Average number	9.48	11.39

[a] Variables significantly different among clusters (χ^2 tests and *t* tests, *P*<.01). Each line represents 1 cluster and adds up to 1 (100%).

Synthesizing ICD-10 and CHOP Diagnoses

Clustering the large data set with more than 2000 different ICD-10 and 800 different CHOP diagnoses into general clusters was not interpretable. To make it suitable for further analysis, the ICD-10 data set was recoded into 4 groups: physiological systems, mental illnesses, oncological diseases, and others. The CHOP diagnoses were also recoded into 4 groups: physiological systems, sensorial, other, and measurement instruments for diagnostics (Table 6).

Table 6. Distribution of the recoded data set using the ICD-10 and CHOP diagnoses (N=20,422).

Diagnosis data set	Recoded data set					
	First	Second	Third	Fourth	Fifth	Total
ICD-10[a] **diagnoses**						
Physiological systems	10,666	10,311	10,277	10,034	9,495	50,783
Mental illnesses	2041	1181	856	609	465	5152
Oncological diseases	221	770	974	1012	1075	4052
Others	7490	7829	7308	6609	5768	35,004
No diagnosis	—	331	1008	2158	3619	7116
Total	20,418	20,422	19,415	20,422	20,422	
CHOP diagnostics						
Physiological systems	5086	3656	2255	2049	1293	14,339
Sensorial	526	1448	1370	740	489	4573
Other	8535	4964	3222	1964	1503	20,188
Measurement instruments	—	23	22	1	—	46
Total	14,147	10,091	6869	4754	3285	
No diagnosis/surgery	6275	10,331	13,553	15,668	17,137	

[a] ICD-10: 10th revision of the International Statistical Classification of Diseases and Related Health Problems.

Summary of Synthesized Registry Data

The different clustering and recoding methods resulted in the data set presented in Table 7.

Table 7. Summary of the variables and clusters in the synthesized data set ready for further advanced statistical analysis.

Domain	Variables per cluster in the synthesized database	Recoding[a] cluster level[b]	Inpatients in each cluster, n (%)
Sociodemographic characteristics (N=20,422)	6	—	20,422 (100.00)
Cognitive status (green textbox in Figure 1; n= 20,401)	5	2[b]	18,318 (89.79) and 2083 (10.21)
Somatic status (orange textbox in Figure 1)			
Mobility subcluster (n=20,418)	3	2[b]	12,540 (61.42) and 7878 (38.58)
Health impairments subcluster (n=20,362)	5	2[b]	17,897 (87.89) and 2465 (12.11)
Activities of daily living subcluster (n=20,409)	5	2[b]	17,836 (87.39) and 2573 (12.61)
Medical condition ICD-10[c] and CHOP (gray and yellow textboxes in Figure 1; N=20,422)	2,800	4[a]	Not applicable
Medicines (blue textbox in Figure 1; N=20,422)	2,370	14[a]	Not applicable

[a] Coded data.

[b] Clustered data (ability/impairment).

[c] ICD-10: 10th revision of the International Statistical Classification of Diseases and Related Health Problems.

Discussion

Principal Findings

This paper describes the rationale and methods used to synthesize a large, routinely collected data set of clinical and medical information concerning polymedicated home-dwelling older adults during hospitalization. The electronic patient records from a hospital center provided a valuable data resource for researchers wishing to perform a variety of analyses to explore health risk determinants, medication prescribing, rehospitalization, and death rates. Prospectively collecting research data is often time-consuming and expensive, resulting in biased samples of highly selected individuals, who are often unrepresentative of real-life patients [21]. Data that are already available for use in anonymized electronic patient records provide a valuable opportunity for a variety of different research designs and are particularly useful in the design of registries for evaluating patient outcomes [44]. In some situations, using population-based registries is even preferable to collecting primary data because selection bias due to nonresponders is not a problem [21]. However, large patient registries are sometimes also inconvenient as they frequently present raw data sets and, for several different reasons, they may not be immediately suitable for performing advanced statistical analyses [22]. Those large data sets usually need to be transformed, cleaned-up, and synthesized to be usable for advanced descriptive and predictive statistical analyses.

Our 4-year population-based data set was composed of polymedicated home-dwelling older inpatients with multiple chronic conditions, hospitalized and perhaps rehospitalized in a hospital center in the French-speaking part of Switzerland. The data came from multiple data set sources and were not easily exploitable for advanced statistical analyses, forcing the research team to explore and develop a synthesizing strategy for a large set of variables so as to respond to our research question. Synthesizing a large number of heterogeneous variables in a finite set of specific medical, clinical, and medication data groups was carried out using the principles of cluster methodologies [30,32] and following Olsen's recommendations for best practices in the analysis of population-based registries [22]. Most of the variables documenting patients' health status fulfilled the criteria for clustering into different groups according to the dimensions of their health status. Despite the existence of a large number of clustering algorithms, we observed that clustering variables remains a challenge [37]. First, our data set covered a large number of different domains, and it is often the case that clustering algorithms must be applied to heterogeneous sets of variables, creating an acute need for robust, scalable clustering methods for mixed continuous and categorical-scale data [45]. Current clustering methods for mixed-type data are generally unable to equitably balance the contributions of continuous and categorical variables without strong parametric assumptions. Second, stable cluster analysis is strongly dependent on the data set, especially on how well separated and how homogeneous the clusters are. In the same clustering exercise, some clusters will be more or less stable than others [46]. To overcome this challenge, our study used a combined empirical and statistical approach. In the empirical approach, the variables in the clusters and subclusters were selected following expert opinion (FP, HV, and AvG), presenting the most homogeneous groups possible within the set of variables described in the literature [47]. In the statistical approach, we used the most appropriate clustering methods and compared the results with the experts' opinions, which served as a validation tool to address any possible subjectivity in those opinions. Both methods were implemented independently and compared. This approach was similar to that used in 2 recent studies exploring frailty and comorbidity patterns [27,28]. Although this study developed 6 clusters based on best practices and the previously mentioned empirical statistical approach, other underlying subclusters could also be present within them. This was also noted in the study by Newcomer et al [48] which used agglomerative hierarchical clustering methods to identify clinically relevant subclusters based on groupings of coexisting conditions in a large sample of hospitalized adults.

This study demonstrated that constructing subclusters should not rely solely on an explicit statement indicating the worst outcome, such as death. Clinical indicators documenting functional deterioration which led to a progressive decline and a poor health status were integrated into the 7 clustered data sets. A recent population-based registry study by Vuik et al [49] confirmed the utility of this kind of approach and concluded that health status could not only be based on sociodemographic characteristics and medical diagnoses such as age or morbidity, but should also consider specific assessments of clinical care and patient function.

The procedure used in this study can be summarized as a 7-step approach to transforming and synthesizing a raw, multidimensional, hospital patient registry data set into an exploitable database:

1. Write a protocol including a problem statement, research questions or hypotheses, and data extraction methods incorporating inclusion and exclusion criteria.
2. Explore the hospital register's data catalog (content of administrative, clinical, medical, and drug data; frequency of assessment; types of measurement—health scores, structured observations, free text—as well as the period of data available) in collaboration with the hospital's clinical data warehouse.
3. Request ethical approval from an ethics committee for the use/reuse of existing patient data.
4. Select the most appropriate data for responding to the research questions/hypotheses.
5. Prepare the data set for further analysis by extracting hospital register data into a CSV (.csv) or Excel (.xls) format, cleaning the data in that format's file and importing the data set into a statistical package such as R, SPSS, or STATA.
6. Analyze missing data and strategies to address missing values based on best practice.
7. Synthesize the data with regard to the research questions by recoding and clustering.

Strengths and Limitations

The strengths of our retrospective registry study lie in its huge sample, allowing us to explore the data's variability and homogeneity in depth. Clustering data risks reducing their variability and the information that can be extracted from them, and some clinical variables showed a significant number of missing values. This fact raises questions about the accuracy and quality of the clinical data assessed, which would require measures of interrater reliability among the health care professionals inputting data into the registry. However, because this was beyond the study's aims, we did not explore interrater scores of clinical assessments or health care professionals' scoring of routinely assessed clinical data.

Another limitation to our study was that the sample was restricted to inpatients aged 65 years or older. Because this retrospective, register-based study was part of a larger project [50] focused on medication management among polymedicated, home-dwelling older adults with multiple chronic conditions, we did not have the ethics committee's approval to extend our extraction of data from the hospital register to all hospitalized adults. Furthermore, our analysis did not consider medicines prescribed before hospital admissions due to a lack of data accuracy and validity.

Finally, and surprisingly, our hospital data set revealed a low mortality rate. Considering the incidence of death in the region, our database showed that it was limited in its representativeness of mortality. Older inpatients presenting with a severe functional decline or at the end of their life probably left the hospital early to die at home or in a nursing home/intermediate care clinic.

Research Perspectives

Transforming and synthesizing electronic health records is an intermediate stage in the process of subsequently investigating risk profiles and predictive and survival outcomes. Proceeding to these types of analyses requires that each patient has a personal identifier (PID) for computing survival, predictive risk factors, re-admission rates, unplanned institutionalization, and other clinical outcomes explored in cohort and case–control studies. In addition, survival analysis must be performed up to 18 months after discharge—beyond our data analysis cut-off point. Within the framework of a trajectory analysis of health care, all the longitudinal data on 1 patient should be on the same horizontal line in the spreadsheet used for calculations. To do this, each patient must have a unique code allowing data to be linked across multiple hospitalizations. Risk and predictive analyses could be organized using multiple linear logistic regression models (generalized estimating equation [GEE statistics]).

In this study, the data synthesized to date will enable our research to be completed with additional longitudinal survival analyses. The construction of sequences of hospitalizations and rehospitalizations will allow us to better understand the impact of certain events from a longitudinal perspective. The registry data have some limitations because observations are equally spaced in time and all start from the same point, in 2015. However, this study promises to provide valid and robust results, because, despite the sample period, the next hospitalization may in fact be the best measure of treatment impact. For instance, the consequences of treatment decisions taken during one hospitalization (such as medications prescribed or surgical interventions) might only be measurable when the older inpatient needs to be rehospitalized. Yet those unequal periods between hospitalizations may actually prove to be advantageous because they provide a period of effect—that is, a period selected naturally by the evolving health status specific to each older inpatient (eg, inappropriate treatments make inpatients return to hospital at the exact moment their health worsens). A survival analysis would need to be performed to measure the impact of each important intervention (medical act or medication prescription).

Conclusions

This retrospective registry analysis study delivered a method to transform and synthesize a large, raw data set, which included patients' health records with sociodemographic, clinical, medical, health status, and medication data. Data were cleaned-up and the most appropriate approach for managing missing values was applied. The multicomponent data synthesis strategy integrated recoding together with empirical and evidence-based statistical clustering methods. Seven clusters were constructed to present the health status of hospitalized older adult inpatients. Medical status, comorbidity, and medication data were recoded to summarize the large data set. Finally, our overall strategy delivered an exploitable, population-based database for the advanced analysis of descriptive, predictive, and survival statistics for older inpatients.

Acknowledgments

The authors thank the Valais Hospital directorate for making their hospital patient registries available. This work is supported by the Swiss National Science Foundation grant number 407440_183434/1.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Sociodemographic characteristics, frequencies and hospital lengths of stay of older adult inpatients (N=20,422) for the period 2015–2018.

[DOCX File , 33 KB - medinform_v9i5e24205_app1.docx]

Multimedia Appendix 2

Distributions of the somatic status of hospitalized older inpatients at hospital discharge (N=20,422).

[DOCX File , 37 KB - medinform_v9i5e24205_app2.docx]

Multimedia Appendix 3

Distributions of cognitive status data for hospitalized older inpatients (N=20,422).

[DOCX File , 34 KB - medinform_v9i5e24205_app3.docx]

Multimedia Appendix 4

Distributions of ICD-10 and CHOP data for hospitalized older inpatients (N=20,422).

[DOCX File , 33 KB - medinform_v9i5e24205_app4.docx]

Multimedia Appendix 5

Distribution of the number of medicines at hospital discharge (N=20,422).

[DOCX File , 32 KB - medinform_v9i5e24205_app5.docx]

Multimedia Appendix 6

Distributions of prescribed medicines for discharged older adult inpatients based on the first level of the ATC classification system (N=20,422).

[DOCX File , 32 KB - medinform_v9i5e24205_app6.docx]

Multimedia Appendix 7

Distribution of the number of deteriorated health conditions among the sample of hospitalized older inpatients (N=20,422).

[DOCX File , 32 KB - medinform_v9i5e24205_app7.docx]

References

1. Gliklich R, Dreyer N, Leavy M. Registries for Evaluating Patient Outcomes: Patient Registries (3rd ed). Rockville, MD: Agency for Healthcare Research and Quality; 2014.
2. Strasberg H, Tudiver F, Holbrook AM, Geiger G, Keshavjee KK, Troyan S. Moving towards an electronic patient record: a survey to assess the needs of community family physicians. Proc AMIA Symp 1998:230-234. [Medline: 9929216]
3. Brooke E. The current and future use of registers in health information systems. World Health Organization: Geneva 1974:43 [FREE Full text]

4. Walsh K, Marsolo KA, Davis C, Todd T, Martineau B, Arbaugh C, et al. Accuracy of the medication list in the electronic health record-implications for care, research, and improvement. J Am Med Inform Assoc 2018 Jul 01;25(7):909-912 [FREE Full text] [doi: 10.1093/jamia/ocy027] [Medline: 29771350]
5. Chipps E, Tucker S, Labardee R, Thomas B, Weber M, Gallagher-Ford L, et al. The Impact of the Electronic Health Record on Moving New Evidence-Based Nursing Practices Forward. Worldviews Evid Based Nurs 2020 Apr;17(2):136-143. [doi: 10.1111/wvn.12435] [Medline: 32233009]
6. Hoque DME, Kumari V, Hoque M, Ruseckaite R, Romero L, Evans SM. Impact of clinical registries on quality of patient care and clinical outcomes: A systematic review. PLoS One 2017 Sep 8;12(9):e0183667 [FREE Full text] [doi: 10.1371/journal.pone.0183667] [Medline: 28886607]
7. Rogan E, Ranson CA, Valle-Oseguera CS, Lee C, Gumberg A, Nagin BN, et al. Factors associated with medication-related problems in an ambulatory medicare population and the case for medication therapy management. Res Social Adm Pharm 2020 Jun;16(6):783-786. [doi: 10.1016/j.sapharm.2019.08.033] [Medline: 31447267]
8. Nicosia FM, Spar MJ, Stebbins M, Sudore RL, Ritchie CS, Lee KP, et al. What Is a Medication-Related Problem? A Qualitative Study of Older Adults and Primary Care Clinicians. J Gen Intern Med 2020 Mar 01;35(3):724-731 [FREE Full text] [doi: 10.1007/s11606-019-05463-z] [Medline: 31677102]
9. van der Hooft CS, Dieleman JP, Siemes C, Aarnoudse AL, Verhamme KM, Stricker BH, et al. Adverse drug reaction-related hospitalisations: a population-based cohort study. Pharmacoepidemiol Drug Saf 2008 Apr 27;17(4):365-371. [doi: 10.1002/pds.1565] [Medline: 18302300]
10. OFS. Statistique médicale des hôpitaux 2015. Actualités OFS cited. 2016. URL: https://www.bfs.admin.ch/bfs/fr/home/statistiques/sante/systeme-sante/hopitaux/patients-hospitalisations.html [accessed 2021-04-14]
11. Pereira F, von Gunten A, Rosselet Amoussou J, De Giorgi Salamun I, Martins MM, Verloo H. Polypharmacy Among Home-Dwelling Older Adults: The Urgent Need for an Evidence-Based Medication Management Model. Patient Prefer Adherence 2019;13:2137-2143 [FREE Full text] [doi: 10.2147/PPA.S232575] [Medline: 31908421]
12. Stevenson J, Davies JG, Martin FC. Medication-related harm: a geriatric syndrome. Age Ageing 2019 Dec 01;49(1):7-11. [doi: 10.1093/ageing/afz121] [Medline: 31665207]
13. Belliardo C, Bouillot E, Heurte D, Curti C, Castera-Ducros C, Vanelle P, et al. Médicaments à haut risque : état des lieux de leur utilisation dans des services d'hospitalisation conventionnelle adulte au CHU. Le Pharmacien Hospitalier et Clinicien 2018 Jul;53(3):223-230. [doi: 10.1016/j.phclin.2018.04.005]
14. Lea M, Mowe M, Mathiesen L, Kvernrød K, Skovlund E, Molden E. Prevalence and risk factors of drug-related hospitalizations in multimorbid patients admitted to an internal medicine ward. PLoS One 2019 Jul 22;14(7):e0220071 [FREE Full text] [doi: 10.1371/journal.pone.0220071] [Medline: 31329634]
15. Nickel CH, Ruedinger JM, Messmer AS, Maile S, Peng A, Bodmer M, et al. Drug-related emergency department visits by elderly patients presenting with non-specific complaints. Scand J Trauma Resusc Emerg Med 2013 Mar 05;21(1):15 [FREE Full text] [doi: 10.1186/1757-7241-21-15] [Medline: 23497667]
16. Budnitz DS, Pollock DA, Weidenbach KN, Mendelsohn AB, Schroeder TJ, Annest JL. National surveillance of emergency department visits for outpatient adverse drug events. JAMA 2006 Oct 18;296(15):1858-1866. [doi: 10.1001/jama.296.15.1858] [Medline: 17047216]
17. Shehab N, Lovegrove MC, Geller AI, Rose KO, Weidle NJ, Budnitz DS. US Emergency Department Visits for Outpatient Adverse Drug Events, 2013-2014. JAMA 2016 Nov 22;316(20):2115-2125 [FREE Full text] [doi: 10.1001/jama.2016.16201] [Medline: 27893129]
18. Šteinmiller J, Routasalo P, Suominen T. Older people in the emergency department: a literature review. Int J Older People Nurs 2015 Jul 17;10(4):284-305. [doi: 10.1111/opn.12090]
19. Linkens AEMJH, Milosevic V, van der Kuy PHM, Damen-Hendriks VH, Mestres Gonzalvo C, Hurkens KPGM. Medication-related hospital admissions and readmissions in older patients: an overview of literature. Int J Clin Pharm 2020 May 30;42(5):1243-1251. [doi: 10.1007/s11096-020-01040-1]
20. Hummel M, Edelmann D, Kopp-Schneider A. Clustering of samples and variables with mixed-type data. PLoS One 2017 Nov 28;12(11):e0188274 [FREE Full text] [doi: 10.1371/journal.pone.0188274] [Medline: 29182671]
21. Thygesen LC, Ersbøll AK. When the entire population is the sample: strengths and limitations in register-based epidemiology. Eur J Epidemiol 2014 Aug 10;29(8):551-558. [doi: 10.1007/s10654-013-9873-0] [Medline: 24407880]
22. Olsen J. Register-based research: some methodological considerations. Scand J Public Health 2011 May 22;39(3):225-229. [doi: 10.1177/1403494811402719] [Medline: 21427148]
23. World Health Organization. International Statistical Classification of Diseases and Related Health Problems (ICD). Classification of Diseases (ICD). 2021. URL: https://www.who.int/standards/classifications/classification-of-diseases [accessed 2021-04-22]
24. Federal Statistical Office. Swiss Classification of Surgical Interventions (CHOP). Classification Suisse des Interventions Chirurgicales (CHOP). 2016. URL: https://www.bfs.admin.ch/bfs/fr/home/statistiques/sante/nomenclatures/medkk/instruments-codage-medical.assetdetail.350129.html [accessed 2021-04-22]
25. World Health Organization. ATC classification system. Collaborating Centre for Drug Statistics Methodology. 2018 Feb 18. URL: https://www.whocc.no/atc/structure_and_principles/ [accessed 2021-04-22]

26. Bergman U, Popa C, Tomson Y, Wettermark B, Einarson TR, Aberg H, et al. Drug utilization 90%--a simple method for assessing the quality of drug prescribing. Eur J Clin Pharmacol 1998 Apr 30;54(2):113-118. [doi: 10.1007/s002280050431] [Medline: 9626914]
27. Tangianu F, Gnerre P, Colombo F, Frediani R, Pinna G, Berti F, et al. Could clustering of comorbidities be useful for better defining the internal medicine patients' complexity? Ital J Med 2018 Jun 20;12(2):137. [doi: 10.4081/itjm.2018.940]
28. Guisado-Clavero M, Roso-Llorach A, López-Jimenez T, Pons-Vigués M, Foguet-Boreu Q, Muñoz MA, et al. Multimorbidity patterns in the elderly: a prospective cohort study with cluster analysis. BMC Geriatr 2018 Jan 16;18(1):16 [FREE Full text] [doi: 10.1186/s12877-018-0705-7] [Medline: 29338690]
29. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological) 2018 Dec 05;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]
30. Omran MG, Engelbrecht AP, Salman A. An overview of clustering methods. IDA 2007 Nov 09;11(6):583-605. [doi: 10.3233/ida-2007-11602]
31. Gower JC. A General Coefficient of Similarity and Some of Its Properties. Biometrics 1971 Dec;27(4):857. [doi: 10.2307/2528823]
32. Ahmad A, Dey L. A k-mean clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering 2007 Nov;63(2):503-527. [doi: 10.1016/j.datak.2007.03.016]
33. Ng SK. A two-way clustering framework to identify disparities in multimorbidity patterns of mental and physical health conditions among Australians. Stat Med 2015 Nov 20;34(26):3444-3460. [doi: 10.1002/sim.6542] [Medline: 26032906]
34. Węglarczyk S. Kernel density estimation and its application. ITM Web Conf 2018 Nov 07;23:00037. [doi: 10.1051/itmconf/20182300037]
35. Park H, Jun C. A simple and fast algorithm for K-medoids clustering. Expert Systems with Applications 2009 Mar;36(2):3336-3341. [doi: 10.1016/j.eswa.2008.01.039]
36. Moschidis OE. A different approach to multiple correspondence analysis (MCA) than that of specific MCA. msh 2009 Oct 12(186):77-88. [doi: 10.4000/msh.11091]
37. Foss A, Markatou M, Ray B, Heching A. A semiparametric method for clustering mixed data. Mach Learn 2016 Jul 15;105(3):419-458. [doi: 10.1007/s10994-016-5575-7]
38. Buchin K, Buchin M, van Kreveld M, Löffler M, Luo J, Silveira RI. Clusters in Aggregated Health Data. In: Lecture Notes in Geoinformation and Cartography. Berlin, Germany: Springer; 2008:77-90.
39. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 1987 Nov;20:53-65. [doi: 10.1016/0377-0427(87)90125-7]
40. Spinks JM, Kalisch Ellett LM, Spurling G, Theodoros T, Williamson D, Wheeler AJ. Adaptation of potentially preventable medication-related hospitalisation indicators for Indigenous populations in Australia using a modified Delphi technique. BMJ Open 2019 Nov 19;9(11):e031369 [FREE Full text] [doi: 10.1136/bmjopen-2019-031369] [Medline: 31748302]
41. Brandão D, Ribeiro O, Teixeira L, Paúl C. Perceived risk of institutionalization, hospitalization, and death in oldest old primary care patients. Arch Gerontol Geriatr 2020 Mar;87:103974. [doi: 10.1016/j.archger.2019.103974] [Medline: 31786410]
42. Little RJA. A Test of Missing Completely at Random for Multivariate Data with Missing Values. Journal of the American Statistical Association 1988 Dec;83(404):1198-1202. [doi: 10.1080/01621459.1988.10478722]
43. Fabbietti P, Ruggiero C, Sganga F, Fusco S, Mammarella F, Barbini N, et al. Effects of hyperpolypharmacy and potentially inappropriate medications (PIMs) on functional decline in older patients discharged from acute care hospitals. Arch Gerontol Geriatr 2018;77:158-162 [FREE Full text] [doi: 10.1016/j.archger.2018.05.007] [Medline: 29778885]
44. Wastesson JW, Cedazo Minguez A, Fastbom J, Maioli S, Johnell K. The composition of polypharmacy: A register-based study of Swedes aged 75 years and older. PLoS One 2018 Mar 29;13(3):e0194892 [FREE Full text] [doi: 10.1371/journal.pone.0194892] [Medline: 29596512]
45. Foguet-Boreu Q, Violán C, Rodriguez-Blanco T, Roso-Llorach A, Pons-Vigués M, Pujol-Ribera E, et al. Multimorbidity Patterns in Elderly Primary Health Care Patients in a South Mediterranean European Region: A Cluster Analysis. PLoS One 2015 Nov 2;10(11):e0141155 [FREE Full text] [doi: 10.1371/journal.pone.0141155] [Medline: 26524599]
46. Hennig C. Cluster-wise assessment of cluster stability. Computational Statistics & Data Analysis 2007 Sep;52(1):258-271. [doi: 10.1016/j.csda.2006.11.025]
47. Khalili S, Phongtankuel V, LaNoue M. Exploring Patterns of Multimorbidity and In-Network Healthcare Utilization Among Older Adults Using Cluster Analysis. Journal of the American Geriatrics Society 2018;66:S137-S137.
48. Newcomer SR, Steiner JF, Bayliss EA. Identifying subgroups of complex patients with cluster analysis. Am J Manag Care 2011 Aug 01;17(8):e324-e332 [FREE Full text] [Medline: 21851140]
49. Vuik SI, Mayer E, Darzi A. A quantitative evidence base for population health: applying utilization-based cluster analysis to segment a patient population. Popul Health Metr 2016 Nov 25;14(1):44 [FREE Full text] [doi: 10.1186/s12963-016-0115-z] [Medline: 27906004]

50. Pereira F, Roux P, Santiago-Delefosse M, von Gunten A, Wernli B, Martins MM, et al. Optimising medication management for polymedicated home-dwelling older adults with multiple chronic conditions: a mixed-methods study protocol. BMJ Open 2019 Oct 28;9(10):e030030 [FREE Full text] [doi: 10.1136/bmjopen-2019-030030] [Medline: 31662367]

Abbreviations

ATC: Anatomical Therapeutic Chemical classification system
ICD-10: 10th revision of the International Statistical Classification of Diseases and Related Health Problems.
MRPs: medication-related problems
NA: not available

Edited by G Eysenbach; submitted 09.09.20; peer-reviewed by KM Kuo, M Green; comments to author 10.11.20; revised version received 11.02.21; accepted 04.04.21; published 11.05.21.

Please cite as:
Taushanov Z, Verloo H, Wernli B, Di Giovanni S, von Gunten A, Pereira F
Transforming a Patient Registry Into a Customized Data Set for the Advanced Statistical Analysis of Health Risk Factors and for Medication-Related Hospitalization Research: Retrospective Hospital Patient Registry Study
JMIR Med Inform 2021;9(5):e24205
URL: https://medinform.jmir.org/2021/5/e24205
doi:10.2196/24205
PMID:33973865

Original Paper

Predicting Intensive Care Unit Length of Stay and Mortality Using Patient Vital Signs: Machine Learning Model Development and Validation

Khalid Alghatani[1], PhD; Nariman Ammar[2], PhD; Abdelmounaam Rezgui[3], PhD; Arash Shaban-Nejad[2], PhD

[1]Department of Computer Science and Engineering, New Mexico Institute of Mining and Technology, Socorro, NM, United States

[2]Oak Ridge National Laboratory Center for Biomedical Informatics, Department of Pediatrics, College of Medicine, The University of Tennessee Health Science Center, Memphis, TN, United States

[3]School of Information Technology, Illinois State University, Normal, IL, United States

Corresponding Author:
Khalid Alghatani, PhD
Department of Computer Science and Engineering
New Mexico Institute of Mining and Technology
801 Leroy Pl
Socorro, NM, 87801
United States
Phone: 1 5057204644
Email: khalid.alghatani@student.nmt.edu

Abstract

Background: Patient monitoring is vital in all stages of care. In particular, intensive care unit (ICU) patient monitoring has the potential to reduce complications and morbidity, and to increase the quality of care by enabling hospitals to deliver higher-quality, cost-effective patient care, and improve the quality of medical services in the ICU.

Objective: We here report the development and validation of ICU length of stay and mortality prediction models. The models will be used in an intelligent ICU patient monitoring module of an Intelligent Remote Patient Monitoring (IRPM) framework that monitors the health status of patients, and generates timely alerts, maneuver guidance, or reports when adverse medical conditions are predicted.

Methods: We utilized the publicly available Medical Information Mart for Intensive Care (MIMIC) database to extract ICU stay data for adult patients to build two prediction models: one for mortality prediction and another for ICU length of stay. For the mortality model, we applied six commonly used machine learning (ML) binary classification algorithms for predicting the discharge status (survived or not). For the length of stay model, we applied the same six ML algorithms for binary classification using the median patient population ICU stay of 2.64 days. For the regression-based classification, we used two ML algorithms for predicting the number of days. We built two variations of each prediction model: one using 12 baseline demographic and vital sign features, and the other based on our proposed quantiles approach, in which we use 21 extra features engineered from the baseline vital sign features, including their modified means, standard deviations, and quantile percentages.

Results: We could perform predictive modeling with minimal features while maintaining reasonable performance using the quantiles approach. The best accuracy achieved in the mortality model was approximately 89% using the random forest algorithm. The highest accuracy achieved in the length of stay model, based on the population median ICU stay (2.64 days), was approximately 65% using the random forest algorithm.

Conclusions: The novelty in our approach is that we built models to predict ICU length of stay and mortality with reasonable accuracy based on a combination of ML and the quantiles approach that utilizes only vital signs available from the patient's profile without the need to use any external features. This approach is based on feature engineering of the vital signs by including their modified means, standard deviations, and quantile percentages of the original features, which provided a richer dataset to achieve better predictive power in our models.

(JMIR Med Inform 2021;9(5):e21347) doi:10.2196/21347

KEYWORDS

intensive care unit (ICU); ICU patient monitoring; machine learning; predictive model; vital signs measurements; clinical intelligence

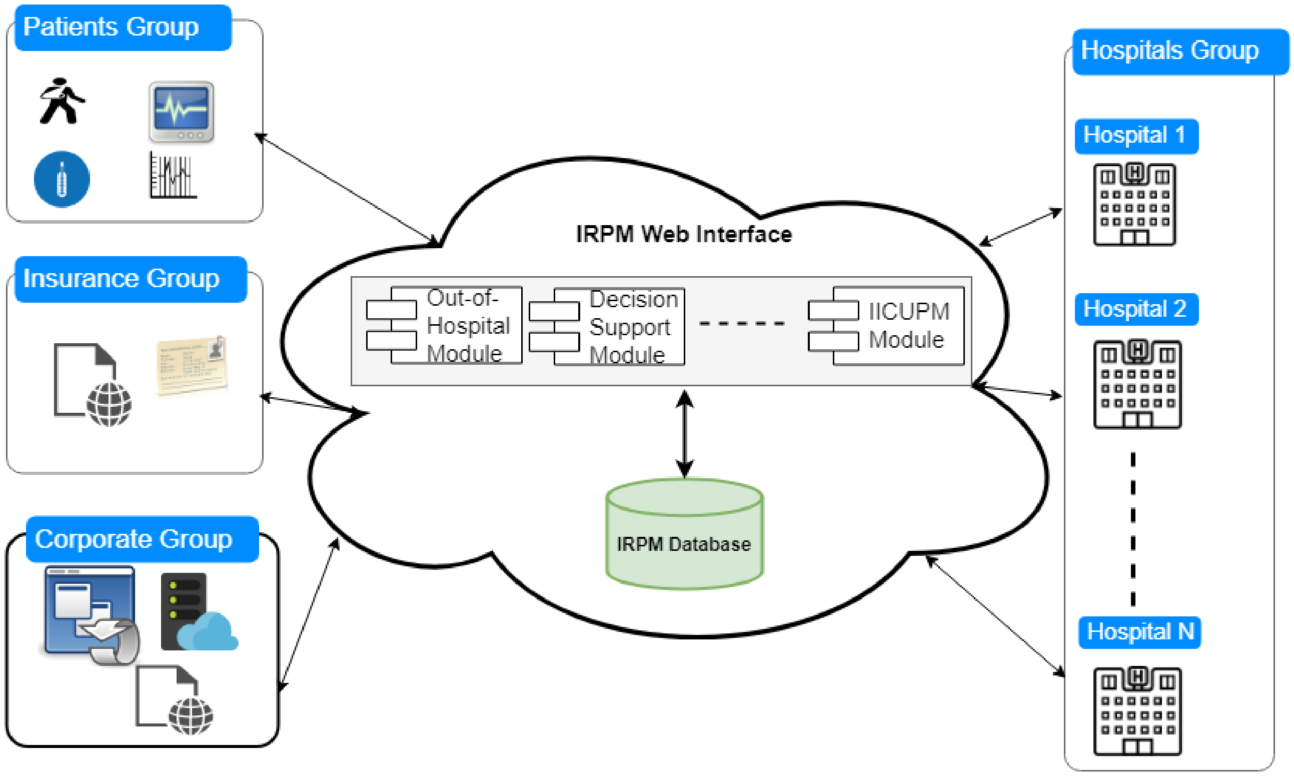
Introduction

Background

Precision observation and assessment are crucial tasks for "achieving an early diagnosis, informed planning, reflecting on the suitability of treatment options, information exchanging, and designing better health interventions" [1]. The use of artificial intelligence–based solutions to improve health care services is increasing [2] and patient monitoring is now an integral part of clinical intelligence [3]. The intensive care unit (ICU) is one of the most critical and resource-intensive units in hospitals, and ICU patient monitoring and continuous clinical surveillance have the potential to reduce morbidity and improve the quality of care. Therefore, hospitals often seek solutions that enable reducing waste and wait times, while increasing service efficiencies, accuracy, and productivity [2]. One of the issues in current monitoring approaches is that the data are collected via sensing devices and sent to remote diagnostic testing facilities for further, often manual or semiautomated, interpretation by a health care professional. Thus, there is a need for intelligent solutions for ICU patient monitoring that require minimal human intervention and that can monitor the health status of patients, and generate timely alerts, maneuver guidance, and reports whenever adverse medical conditions are anticipated.

In our previous work [4], we proposed an Intelligent Remote Patient Monitoring (IRPM) framework (Figure 1) that consists of three modules: (i) an out-of-hospital module that utilizes data collected via wearable devices (eg, Apple Watch and SleepO2); (ii) a decision support module that generates reports; and (iii) an intelligent ICU patient monitoring module, which utilizes data collected from ICUs. We here focus on the latter module.

Figure 1. Intelligent Remote Patient Monitoring (IRPM) framework. IICUPM: intelligent intensive care unit patient monitoring.



The IRPM framework is intended to serve as a global web service interface that exposes the different framework functionalities to hospitals, hospital managers, insurance companies, and other decision makers, including the host organizations that operate and maintain the IRPM system. The intelligent ICU patient monitoring functionality of the service performs analytics of the data exchanged between ICUs and the core IRPM system, and provides the different stakeholders with the analysis results in the form of timely and early warnings.

Three main factors impact the quality of prediction models: (1) the target patient population [5], (2) methods used for data fusion, and (3) algorithm type. Different populations lead to different prediction results. Moreover, different ways of combining information from physiological variables lead to various outcome measures. The IRPM framework is intended to be hosted in the cloud since the intelligent ICU patient monitoring module aims at applying machine learning (ML) within an architecture that allows any user (regardless of whether or not they are sick) as well as any hospital system to use the framework. Since most of the used physiological variables are often obtained inside and outside hospitals, the framework will enable performing continuous patient monitoring. Therefore, we built ML models by utilizing features that are easy to obtain outside the hospital setting, and we avoided features that are sophisticated and require high-level medical equipment.

Related Works

There has been some research effort toward developing ML models for predicting ICU-related outcomes [6-8]. McCarthy et al [9] performed a study on ICU mortality prediction in which they compared sliding-window predictors with recurrent predictors to classify patient state of health from ICU multivariate time-series data. They reported slightly improved performance for the recurrent neural network. Moreover, Zhu et al [10] proposed an ICU mortality prediction algorithm combining the bidirectional long short-term memory (LSTM) model with supervised learning. They trained and evaluated the LSTM model using 4000 ICU patients. They also performed a comparative analysis, which identified that their proposed method significantly outperformed several baseline methods.

A few studies have also focused on developing and validating ML models for predicting ICU-related outcomes using the Medical Information Mart for Intensive Care (MIMIC) database. Most of these works have used an exhaustive list of features to achieve higher accuracy in their models. Johnson and colleagues [11-13] developed models for predicting ICU mortality, achieving an area under the receiver operating characteristic (ROC) curve (AUROC) of 0.92 using a total of 148 features [12] and an AUROC of 0.86 using a range of features, including standard statistical descriptors [13]. Lehman et al [14] used basic physiological variables and applied the Simplified Acute Physiology Score (SAPS-I) algorithm to predict mortality, which achieved an AUROC of 0.72. Using the Cohen standardized mean and coefficient, Tyler et al [15] assessed the differences between ICU lab values, which were used to predict ICU length of stay (LOS) and mortality. Harutyunyan et al [16] selected 17 clinical variables to build a binary LOS model to predict whether a patient will stay in the ICU for a long (≥7 days) or short (<7 days) period with 84% accuracy. Gentimis et al [17] used several inputs from seven tables to build an LOS model to predict whether a patient will stay in the ICU for a long (>5 days) or short (≤5 days) period using neural networks, with around 80% accuracy; they removed patients who stayed in the ICU longer than 20 days. Bertsimas et al [18] used several static and dynamic variables (eg, general admission data, lab results, medical orders, pharmacy data, diagnosis codes, and notes) and different classification methods to predict different LOS with accuracy in the >80% range.

Some works have focused on developing ML models to be used in clinical information systems that assist in ICU discharge planning. Badawi and Breslow [19] developed and validated two models for predicting risks of death and readmission within 48 hours of ICU discharge. They used eICU Research Institute data from more than 400 ICUs and performed multivariate logistic regression (MLR) with 59 different features, including patient demographics, ICU admission diagnosis, admission severity of illness, intensive care interventions, complications occurring during the ICU stay, lab values, and physiological variables recorded within the last 24 hours of the ICU stay. They calibrated their models across deciles of risk, and their mortality model accurately discriminated between patients who would and would not experience a complication as early as 4 days before ICU discharge. However, to the best of our knowledge, predicting the LOS based on the population's median ICU patient stay using only vital signs and demographic attributes from MIMIC data has not been studied to date.

Objective

We here propose a new approach that focuses on the most critical observations in a patient's profile. The novelty of the approach lies in its ability to predict outcomes with reasonable accuracy by utilizing only vital signs that exist in the patient's profile without having prior knowledge about a patient's medical conditions or diagnoses. The approach enriches the original vital sign measures by adding extra features pertaining to their modified means, modified SDs, and quantile percentages. We evaluated the proposed approach (ie, the quantiles approach) in comparison to a baseline approach that uses the entire range of observations. We then applied both approaches to develop and validate two prediction models: (i) one focusing on classifying ICU mortality rate (survival or no survival), and (ii) another focusing on predicting the LOS in the ICU using public data from the MIMIC database.

Methods

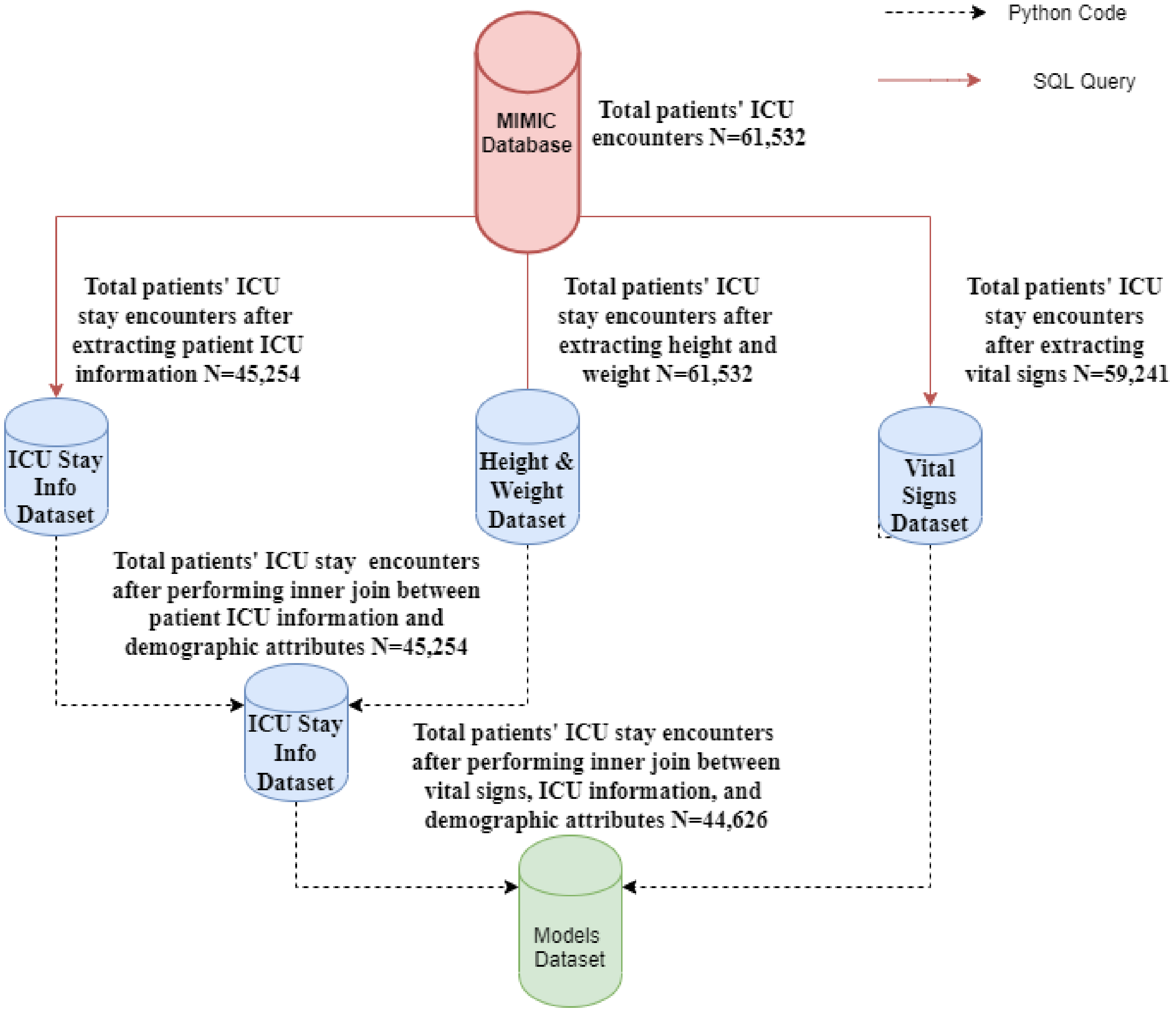
Study Population and Data Extraction

We used MIMIC-III (v1.4) [7], a publicly available ICU adult patient database that spans 11 years between 2001 and 2012. MIMIC-III has data for 53,423 distinct hospital admissions, including nearly 500 million rows in 26 tables. The database comprises features, including patient demographics, laboratory test results, medical reports, and results from imaging studies. To meet Health Insurance Portability and Accountability Act requirements, approximate ages for patients who are more than 89 years are reported by shifting their date of birth.

Figure 2 illustrates the data extraction pipeline of ICU stays data from the MIMIC database. We started with 61,532 total ICU stay encounters. In each hospital admission, a patient could have stayed in the ICU more than once. We performed this study based on unique ICU stays rather than unique patient identifiers since our goal was to predict mortality and LOS without having prior knowledge about patients' medical conditions or diagnoses.

For patients who stayed in the ICU for at least 1 day, we considered their data for only the first day. The population's median ICU LOS was 2.64 days, and therefore we discarded data from patients who stayed in the ICU for less than 1 day, which resulted in a total of 45,254 unique ICU stays. For each ICU stay, we ran separate SQL queries to extract the patients' vital sign measurements, and height and weight features from the total 61,532 encounters. We focused on six vital sign features (body temperature, heart rate, respiration rate, systolic blood pressure, diastolic blood pressure, and oxygen saturation [SpO_2]) along with glucose level. The total number of ICU stays for which vital sign features were available was 59,241. We extracted four demographic features (weight, height, age, and gender). We then performed consecutive inner joins between the results of the three queries; thus, the total ICU stays reduced to 44,626 unique ICU stays.

Figure 2. Data extraction pipeline from the Medical Information Mart for Intensive Care (MIMIC) database. ICU: intensive care unit.


Python Code
SQL Query
MIMIC Database
Total patients' ICU encounters N=61,532
Total patients' ICU stay encounters after extracting patient ICU information N=45,254
Total patients' ICU stay encounters after extracting height and weight N=61,532
Total patients' ICU stay encounters after extracting vital signs N=59,241
ICU Stay Info Dataset
Height & Weight Dataset
Vital Signs Dataset
Total patients' ICU stay encounters after performing inner join between patient ICU information and demographic attributes N=45,254
ICU Stay Info Dataset
Total patients' ICU stay encounters after performing inner join between vital signs, ICU information, and demographic attributes N=44,626
Models Dataset

Data Preprocessing

To enhance the accuracy of the predictive models, we eliminated extreme, trivial, and negative observations within each vital sign feature. The percentage of missing data was relatively low (less than 1% for heart rate, respiration rate, systolic blood pressure, diastolic blood pressure, SpO_2, and glucose level, and less than 2% for body temperature). Given the low percentage of missing values and the fact that vital signs are numerical values that are typically normally distributed [20], we filled missing values of vital sign observations using the mean.

Model and Variable Selection

We built two main prediction models: in-hospital mortality and LOS for each ICU admission. Table 1 defines the outcome variables in both models. The outcome variable for the mortality model was in-hospital mortality, which reduces to a binary classification problem with two classes: predicting a patient to survive or not. The dataset has a classification imbalance problem since the in-hospital mortality percentage was 11.897%, whereas the patient survival percentage was 88.103%. The outcome for the LOS model was the number of days a patient stayed in the ICU. Half of the population spent 2.64 days in the ICU, which led us to follow two approaches for classification. In the first approach, we followed a binary classification strategy by defining two classes with an equal number of observations and by considering 2.64 as a threshold. The first class predicts that a patient will stay in the ICU for 2.64 days or less, and the second class predicts that a patient will stay in the ICU for more than 2.64 days. In the second approach, we followed a regression-based classification strategy by considering the predicted outcome as a continuous variable.

We built two variations of each model: one using the baseline approach and another using the proposed quantiles approach. The models built with the baseline approach used the six vital sign features, glucose, and the five demographic features as predictor variables (Table 2). The models built with the quantiles approach used the same 12 baseline predictor variables, and augmented them with extra modified features. We discuss each model variation separately below.

Table 1. Descriptive statistics for outcome variables in the two models.

Model	Operationalization	Values
In-hospital mortality (binary classification)	0: survival; 1: nonsurvival	0: 11.897%; 1: 88.10 3%
Length of stay (LOS)		
Binary classification	0: LOS≤2.636 days; 1: LOS>2.636 days	0: 50%; 1: 50%
Regression-based classification	Number of days in intensive care unit	Mean 4.74959 (SD 6.49982)

Table 2. Descriptive statistics for baseline model predictors (N=44,626).

Input variables	Measurement	Value
HeartRate_mean	Heart rate (beats/minute), mean (SD)	85.99 (15.59)
sysbp_mean	Arterial systolic blood pressure (mmHg) mean (SD)	118.75 (16.90)
diasbp_mean	Arterial diastolic blood pressure (mmHg), mean (SD)	60.47 (10.89)
RespRate_mean	Respiratory rate (breaths/minute), mean (SD)	18.93 (4.05)
Tempc_mean	Body temperature (°C), mean (SD)	36.84 (0.62)
Spo_2_mean	Peripheral oxygen saturation (%), mean	97.27
Glucose_mean	Blood glucose (mg/dL), mean (SD)	138.74 (41.86)
Age	Age (years), mean (SD)	64.35 (16.87)
GenderM	Male population, n (%)	25,241 (56.56)
GenderF	Female population, n (%)	19,385 (43.44)
Height	Patient height (cm), mean (SD)	160.66 (11.76)
Weight	Patient weight (kg), mean (SD)	80.45 (23.47)

Baseline Approach

Table 2 shows the descriptive statistics for the predictor variables used in the baseline approach: the patients' vital signs for the first day and the demographic variables. The population had a slight majority of men with a mean age of 64.35 years.

Pearson correlation coefficients among the vital sign variables in the baseline approach (Table 3) showed weak correlations between the variables, except between systolic and diastolic blood pressure.

Table 3. Pearson correlation coefficients among vital signs of the baseline model.

Variable	Heart rate	Systolic BP[a]	Diastolic BP	Respiration rate	Body temperature	SpO_2[b]	Glucose
Heart rate	1	–0.104	0.211	0.326	0.268	–0.099	0.063
Systolic BP	–0.104	1	0.524	–0.032	0.065	0.045	0.063
Diastolic BP	0.211	0.524	1	0.0257	0.065	–0.0148	0.0142
Respiration rate	0.326	–0.032	0.0257	1	0.118	–0.259	0.069
Body temperature	0.268	0.065	0.0335	0.118	1	0.051	–0.022
SPO_2	–0.099	0.045	–0.0148	–0.259	0.051	1	–0.048
Glucose	0.063	0.078	0.0142	0.069	–0.022	–0.048	1

[a] BP: blood pressure.

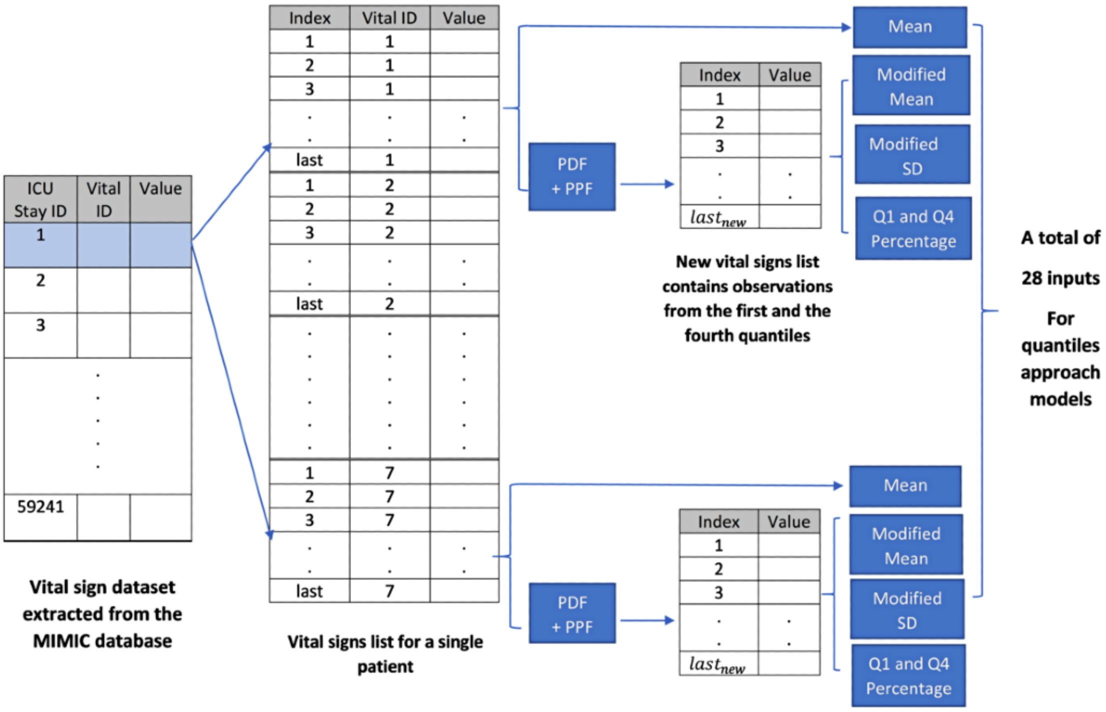
[b] SpO_2: oxygen saturation.

Quantiles Approach

When dealing with sequential data, observations that are far from the median are often ignored. We argue that a patient's deteriorating condition often comes with a high or low level of measurement. Thus, these observations are essential as they report the point at which the patient's health status changes dramatically. We propose the notion of the "quantiles approach," in which we perform feature engineering by emphasizing the high and low quantiles of a patient sample. Figure 3 demonstrates the steps performed in the feature engineering pipeline of the quantiles approach.

First, for each patient sample, we extracted values of the 7 vital sign features. Second, for each vital sign feature within that patient sample, we calculated the mean and SD. Third, we normalized the observations within each vital sign feature using the probability density function, and by passing the mean and SD calculated in step 2 as parameters to that function. The blue histograms in Figures 4 and 5 show the distribution of each vital sign feature before normalization, and the red curves show the distribution after normalization.

Figure 3. Feature engineering pipeline in the quantiles approach. MIMIC: Medical Information Mart for Intensive Care; ICU: intensive care unit; PDF: probability density function; PPF: percent point function.


ICU Stay ID
Vital ID
Value
1
2
3
59241
Vital sign dataset extracted from the MIMIC database
Index
Vital ID
Value
last
1
2
7
Vital signs list for a single patient
PDF + PPF
Index
Value
$last_{new}$
New vital signs list contains observations from the first and the fourth quantiles
Mean
Modified Mean
Modified SD
Q1 and Q4 Percentage
A total of 28 inputs For quantiles approach models

Fourth, we applied the percent point function (PPF) to each normalized vital sign feature to calculate two discrete values corresponding to the low and high values of that feature. The low values correspond to observations of the feature that are less than a given probability (the 25th percentile in our case) and the high values correspond to observations of the feature that are greater than or equal to a given probability (the 75th percentile in our case). Thus, for each vital sign feature, we calculated the values at which each percentage occurs.

Fifth, we used the calculated low and high values from step 4 to extract the observations of the vital sign features that occur in only the first and fourth quantiles (ie, we ignored the second and third quantiles). Sixth, we calculated the mean and SD of the extracted observations. In the remainder of the paper, we refer to these metrics as the modified mean and modified SD to distinguish from the original mean and SD calculated in step 2.

The final step is to calculate the quantile percentage for the vital sign feature by dividing the number of observations extracted in step 5 (ie, those that occur in the first and fourth quantiles) by the original number of observations (in all quantiles in the entire patient sample). Note that since we normalized the observations in the vital sign feature (step 3), the number of observations in the first and fourth quantiles will vary and will not always be 50% of the original observations.

Figure 4. Distribution of a sample patient observation before and after applying the quantiles approach.

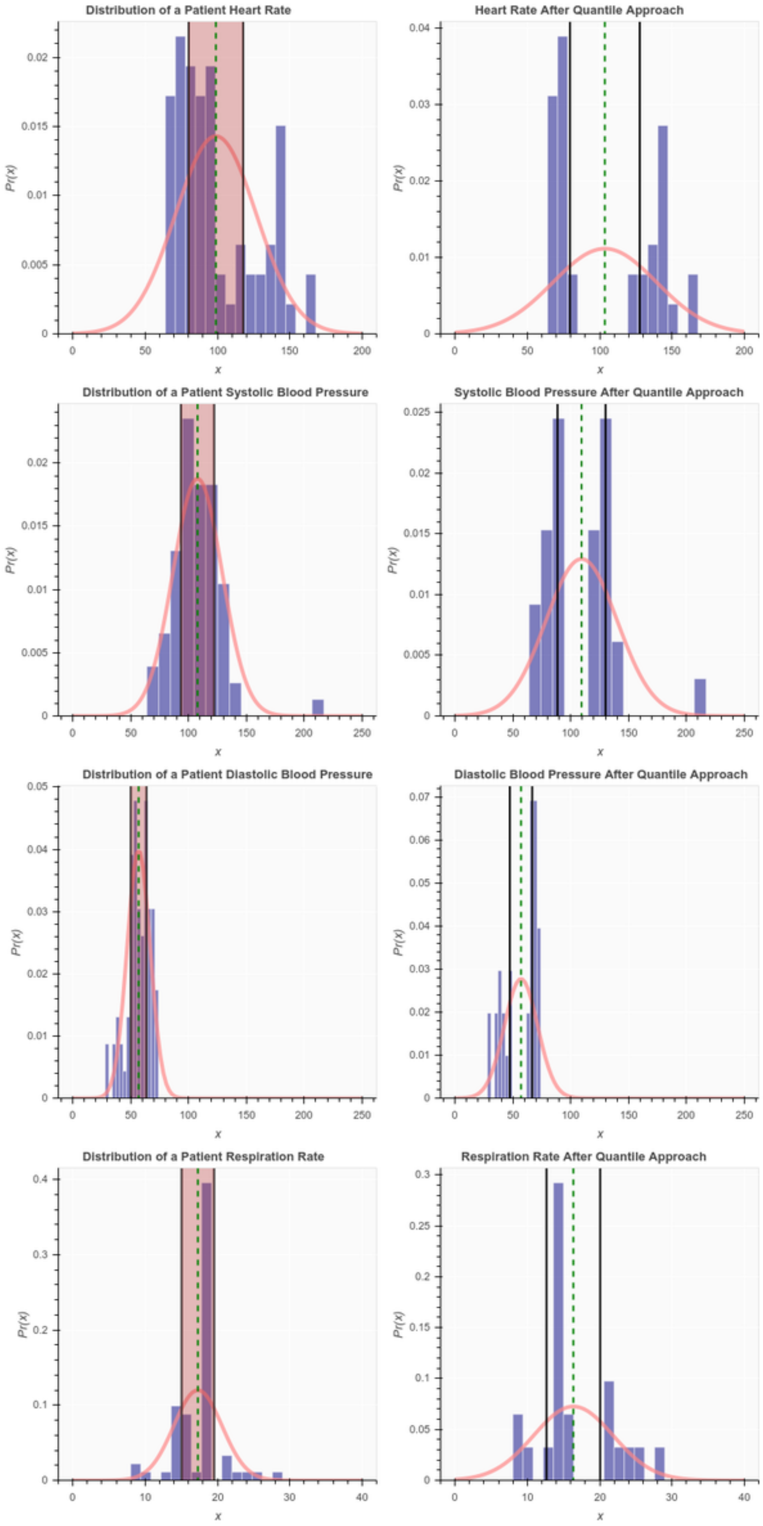
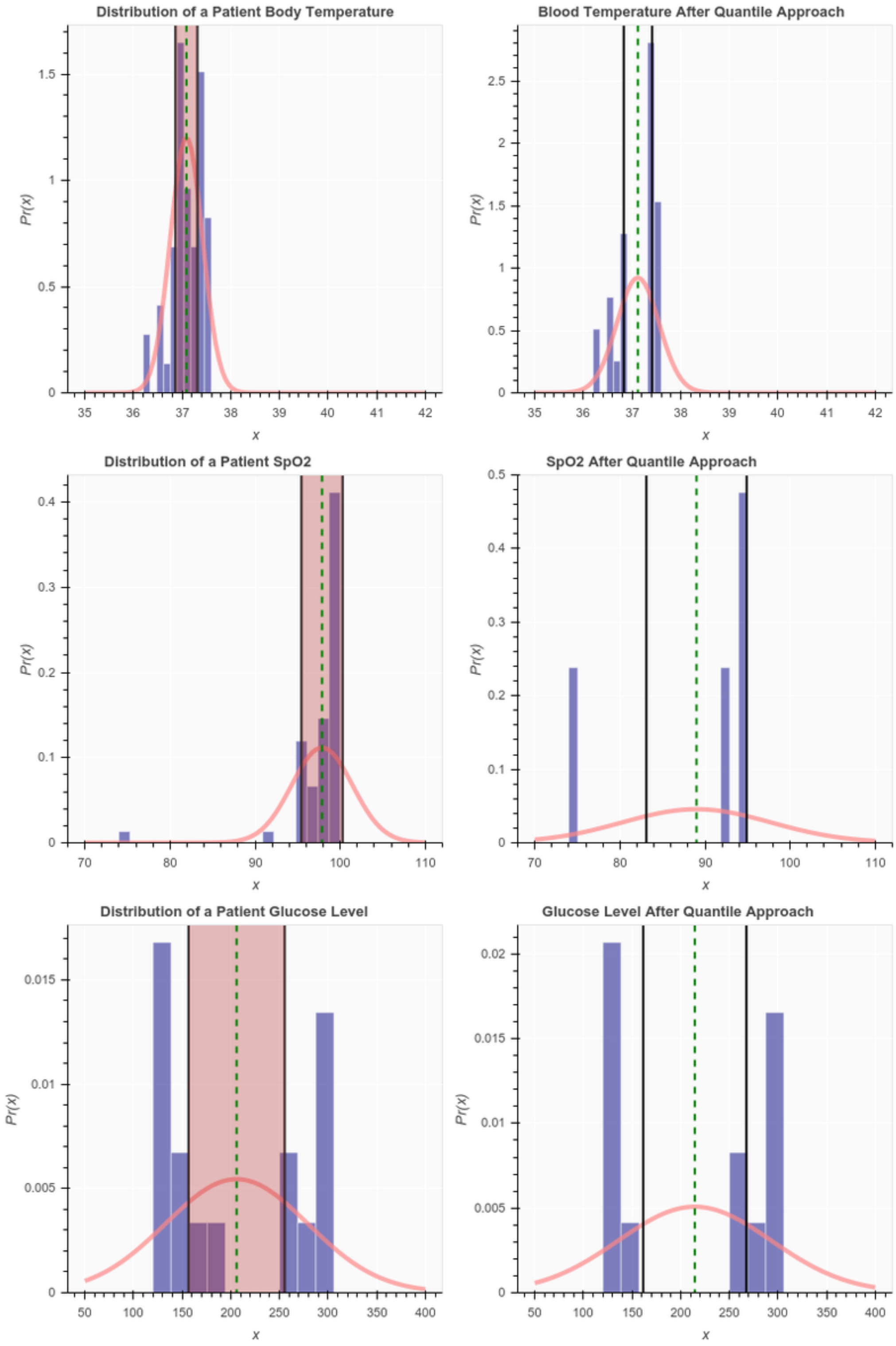


Figure 5. Distribution of a sample patient observation before and after applying the quantiles approach (continued from Figure 4).



Patient Use Case

To demonstrate the quantiles approach, we provide an example of a sample patient before and after applying the steps described above. Figures 4 and 5 show distributions of the 7 vital signs of the patient before (left) and (after) applying the quantiles approach. The shaded areas in Figures 4 and 5 show where the vital sign measurements are neglected. The right side of the figure shows the modified patient's observation after removing the values in the shaded area. After applying the change, the SD of the observation increased most of the time, whereas the mean (the green vertical line) did not change significantly. Table 4 shows an example of individual patient data before applying the quantiles approach. Table 5 demonstrates the features that were engineered from the original 7 vital sign measures for that patient sample.

Table 4. Sample data from an individual patient before applying the quantiles approach.

Feature	Operationalization	Mean (SD)
HeartRate_mean	Mean heart rate (beats/minute)	98.92 (27.89)
sysbp_mean	Mean systolic blood pressure (mmHg)	107.8 (21.26)
diasbp_mean	Mean diastolic blood pressure (mmHg)	56.88 (10.00)
resprate_mean	Mean respiration rate (breaths/minute)	17.29 (3.33)
tempc_mean	Mean body temperature (°C)	37.08 (0.33)
spo_2_mean	Mean oxygen saturation (%)	97.86 (3.57)
glucose_mean	Mean glucose level (mg/dL)	206.0 (73.26)

Table 5. Sample of patient data from after applying the quantiles approach.

Feature	Operationalization	Value
Modified mean		
HeartRate_mean_mod	Mean of modified heart rate (beats/minute)	103.59
sysbp_mean_mod	Mean of modified arterial diastolic blood pressure (mmHg)	109.34
diasbp_mean_mod	Mean of modified arterial systolic blood pressure (mmHg)	57.03
resprate_mean_mod	Mean of modified respiratory rate (breaths/minute)	16.36
tempc_mean_mod	Mean of modified body temperature (°C)	37.12
spo_2_mean_mod	Mean of modified peripheral oxygen saturation (%)	89.00
glucose_mean_mod	Mean of modified blood glucose level (mg/dL)	214.46
Modified SD		
heartRate_std_mod	SD of modified heart rate (beats/minute)	35.76
sysbp_std_mod	SD of modified arterial diastolic blood pressure (mmHg)	30.83
diasbp_std_mod	SD of modified arterial systolic blood pressure (mmHg)	14.36
resprate_std_mod	SD of modified respiratory rate (breaths/minute)	5.49
tempc_std_mod	SD of modified body temperature (°C)	0.43
spo_2_std_mod	SD of modified peripheral oxygen saturation (%)	8.74
glucose_std_mod	SD of modified blood glucose level (mg/dL)	78.60
Modified quantiles		
HeartRateQuantPer	First and fourth quantiles percent of heart Rate	0.5522
SystolicQuantPer	First and fourth quantiles percent of arterial diastolic blood pressure	0.4266
DiastolicQuantPer	First and fourth quantiles percent of arterial systolic blood pressure	0.4400
RespRateQuantPer	First and fourth quantiles percent of respiratory rate	0.3384
TempCQuantPer	First and fourth quantiles percent of body temperature	0.5384
SPO_2QuantPer	First and fourth quantiles percent of peripheral oxygen saturation	0.0689
GlucoseQuantPer	First and fourth quantiles percent of blood glucose level	0.8125

Table 6 lists additional features that were engineered from the original 7 vital sign measures using the quantiles approach for the entire patient population.

Pearson correlation analysis among the means of vital signs samples after applying the quantiles approach (Table 7) showed that there was no significant difference compared to the baseline model (Table 3). This implies that the quantiles approach does not considerably change the correlation between the variables.

Table 6. Vital sign data after applying the quantiles approach for the entire patient population.

Feature	Operationalization	Value
Modified mean, mean (SD)		
HeartRate_mean_mod	Mean of modified heart rate (beats/minute)	86.55 (15.8469)
sysbp_mean_mod	Mean of modified arterial diastolic blood pressure (mmHg)	119.06 (16.865)
diasbp_mean_mod	Mean of modified arterial systolic blood pressure (mmHg)	61.2201 (11.4944)
resprate_mean_mod	Mean of modified respiratory rate (breaths/minute)	19.22 (4.1363)
tempc_mean_mod	Mean of modified body temperature (°C)	36.82 (0.67382)
spo_2_mean_mod	Mean of modified peripheral oxygen saturation (%)	96.00 (5.28098)
glucose_mean_mod	Mean of modified blood glucose level (mg/dL)	144.50 (48.3843)
Modified SD, mean (SD)		
heartRate_std_mod	SD of modified heart rate (beats/minute)	11.33 (6.02761)
sysbp_std_mod	SD of modified arterial diastolic blood pressure (mmHg)	19.22 (7.64726)
diasbp_std_mod	SD of modified arterial systolic blood pressure (mmHg)	13.21 (6.06014)
resprate_std_mod	SD of modified respiratory rate (breaths/minute)	4.96 (2.05444)
tempc_std_mod	SD of modified body temperature (°C)	0.61 (0.35567)
spo_2_std_mod	SD of modified peripheral oxygen saturation (%)	2.53 (2.18251)
glucose_std_mod	SD of modified blood glucose level (mg/dL)	34.69 (32.2924)
Modified quantiles, quantile percentage		
HeartRateQuantPer	First and fourth quantiles percent of heart Rate	51.63
SystolicQuantPer	First and fourth quantiles percent of arterial diastolic blood pressure	50.49
DiastolicQuantPer	First and fourth quantiles percent of arterial systolic blood pressure	47.47
RespRateQuantPer	First and fourth quantiles percent of respiratory rate	49.02
TempCQuantPer	First and fourth quantiles percent of body temperature	56.57
SPO_2QuantPer	First and fourth quantiles percent of peripheral oxygen saturation	46.26
GlucoseQuantPer	First and fourth quantiles percent of blood glucose level	57.04

Table 7. Pearson correlation coefficients among the mean vital signs for a sample patient using the statistical model.

Variable	Heart rate	Systolic BP[a]	Diastolic BP	Respiration rate	Body temperature	SPO_2[b]	Glucose
Heart rate	1	–0.103	0.183	0.316	0.236	–0.065	0.053
Systolic BP	–0.103	1	0.504	–0.034	0.057	0.056	0.069
Diastolic BP	0.183	0.504	1	0.030	0.031	0.028	0.029
Respiration rate	0.316	–0.034	0.030	1	0.128	–0.095	0.064
Body temperature	0.236	0.057	0.031	0.128	1	0.016	–0.028
SPO_2	–0.065	0.056	0.028	–0.095	0.016	1	–0.028
Glucose	0.053	0.069	0.029	0.064	–0.028	–0.028	1

[a] BP: blood pressure.

[b] SpO_2: oxygen saturation.

Inputs to the Baseline Approach Versus the Quantiles Approach

The models built using the baseline approach used 12 predictor variables (ie, 5 demographic attributes and 7 vital signs) (Table 2). The feature engineering step performed in the quantiles approach augmented the original set of vital sign features with 21 extra features (ie, 7 variables corresponding to the mean of each patient observation, 7 variables corresponding to the SD of each patient observation, and 7 variables corresponding to the quantile percentages). Thus, in addition to the original 12 variables used in the baseline, the models built through the quantiles approach used the 21 engineered features.

Classification Methods

Models Applied

We used supervised learning techniques in both models for both variations because the model outputs were labeled accordingly. We split the dataset randomly into 75% as the training set (n=33,469 ICU stays) and 25% as the test set (n=11,157 ICU stays). To avoid overfitting, we used 10-fold cross-validation on the training set. We then trained both models using the training set and we validated the performance of both models using an unseen testing set.

We applied six commonly used ML algorithms for binary classification in both the mortality and LOS models: linear regression (LR), linear discriminant analysis, random forest (RF), k-nearest neighbors (kNN), support vector machine (SVM), and extreme gradient boosting (XGB). For the regression-based classification in the LOS model, we applied two ML algorithms to predict the number of days: MLR and support vector regression (SVR).

RF is an ensemble ML algorithm that generates bootstrapped samples from a dataset and uses the generated samples to construct several decision trees. Majority voting is then performed to decide the best classification of the generated samples. To avoid high correlation between the constructed trees, the algorithm uses a random subset of features to decide at each split point. This feature randomness increases the chances of having correct prediction results. Thus, one important parameter required by the algorithm is the number of features considered. In addition, choosing a high number of trees might increase the execution time with no considerable performance gain [21]. Therefore, another important parameter is the number of decision trees needed to compose the RF.

Parameter Tuning for Mortality Classifiers

For the RF algorithm, we set the maximum number of features to consider for finding a good split to 4, and we set the estimated number of trees in an RF to 500. For SVM, we used the radial basis function as a kernel type and we set the penalty parameter of error *C* to 1.60.

Parameter Tuning for LOS Classifiers

For the RF algorithm, we set the maximum number of features to consider in finding a good split to 4. We also set the estimated number of trees in the RF to 400. For SVM, we used the radial basis function as a kernel type and we set the penalty parameter of error *C* to 0.90.

Model Calibration

To assess the goodness of fit in our models, we compared the accuracy on the test set and the mean accuracy of the trained model. We also used five metrics (accuracy, sensitivity, specificity, negative predictive value, and positive predictive value, along with corresponding 95% CIs) to validate the classification models on an unseen test set from the same population. We examined the difference in AUROC values between the test and training sets. Finally, we examined calibration across deciles using the sigmoid test supported with a visual inspection of calibration curves.

Results

Mortality Prediction Model

Table 8 shows the performance of the mortality models on both the training and test sets using the baseline and the quantiles approach with the six different ML algorithms.

The RF algorithm achieved the highest accuracy (88.61%) in predicting mortality on the test set using the quantiles approach, followed by the XGB algorithm with an accuracy of 88.22%. All models showed high specificity and low sensitivity, indicating that our models performed very well at identifying patients who will survive but not the opposite. XGB showed the highest sensitivity rate (0.16), demonstrating the advantage of using the XGB algorithm to identify patients who will not survive.

We observed relatively low improvement in model accuracy from the baseline approach to the quantiles approach. This can be explained by the imbalanced classification problem in the mortality model (ie, a low mortality rate of 11.89%). Another possible reason is that the sample size was reduced after applying the quantiles approach, which might have misled the classifier. The original sample size (44,626 ICU stays considering only the first day in the ICU) dropped by almost by half since we included only the first and fourth quantiles for each patient observation. The algorithm uses the PPF function to return discrete values that are less than or equal to the given probability, and the best probabilities achieved in our case were at the 25th and 75th percentiles. We tried other probabilities, but due to the small sample size, varying the PPF percent did not have a significant improvement on the results. Figure 6 shows a visual comparison between the accuracy of the six ML algorithms in the mortality model using the quantiles approach. The box plots to the left show the model accuracy on the training set using 10-fold cross-validation and the graph on the right shows the one-time model accuracy on the testing set.

Table 8. Mortality model results for six algorithms using different performance metrics.

Method and algorithm	Training set accuracy, mean (SD)	Test set accuracy (95% CI)	Test set sensitivity (95% CI)	Test set specificity (95% CI)	Test set NPV[a] (95% CI)	Test set PPV[b] (95% CI)
Baseline approach						
LR[c]	0.8826 (0.0058)	0.8806 (0.874-0.890)	0.0331 (0.033-0.034)	0.9979 (0.991-1.009)	0.8817 (0.875-0.891)	0.6923 (0.688-0.700)
LDA[d]	0.8817 (0.0058)	0.8788 (0.873-0.888)	0.0523 (0.052-0.053)	0.9932 (0.986-1.004)	0.8833 (0.877-0.893)	0.5182 (0.515-0.524)
RF[e]	0.8846 (0.0061)	0.8854 (0.879-0.895)	0.1127 (0.112-0.114)	0.9923 (0.985-1.003)	0.8898 (0.884-0.899)	0.6710 (0.666-0.679)
kNN[f]	0.8765 (0.0054)	0.8760 (0.870-0.886)	0.0854 (0.085-0.087)	0.9855 (0.978-0.996)	0.8861 (0.880-0.896)	0.4496 (0.447-0.455)
SVM[g]	0.8837 (0.0058)	0.8808 (0.875-0.890)	0.0272 (0.027-0.028)	0.9989 (0.992-1.010)	0.8811 (0.875-0.891)	0.7872 (0.782-0.796)
XGB[h]	0.8842 (0.0061)	0.8815 (0.875-0.891)	0.1429 (0.142-0.145)	0.9837 (0.977-0.994)	0.8923 (0.886-0.902)	0.5495 (0.546-0.556)
Quantiles approach						
LR	0.8838 (0.0063)	0.8815 (0.875-0.891)	0.0545 (0.054-0.055)	0.9960 (0.989-1.007)	0.8838 (0.878-0.893)	0.6548 (0.650-0.662)
LDA	0.8821 (0.0067)	0.8814 (0.875-0.891)	0.0935 (0.093-0.095)	0.9905 (0.983-1.001)	0.8875 (0.881-0.897)	0.5772 (0.573-0.584)
RF	0.8859 (0.0064)	0.8861 (0.880-0.896)	0.0891 (0.089-0.090)	0.9964 (0.989-1.007)	0.8876 (0.881-0.897)	0.7756 (0.770-0.784)
KNN	0.8802 (0.0060)	0.8764 (0.870-0.886)	0.0589 (0.059-0.060)	0.9895 (0.982-1.000)	0.8836 (0.877-0.893)	0.4395 (0.437-0.445)
SVM	0.8851 (0.0058)	0.8820 (0.876-0.892)	0.0449 (0.045-0.046)	0.9816 (0.991-1.009)	0.8829 (0.877-0.893)	0.7439 (0.739-0.752)
XGB	0.8844 (0.0061)	0.8822 (0.875-0.891)	0.1643 (0.164-0.167)	0.9816 (0.975-0.992)	0.8945 (0.888-0.904)	0.5533 (0.550-0.560)

[a] NPV: negative predictive value.

[b] PPV: positive predictive value.

[c] LR: logistic regression.

[d] LDA: linear discriminant analysis.

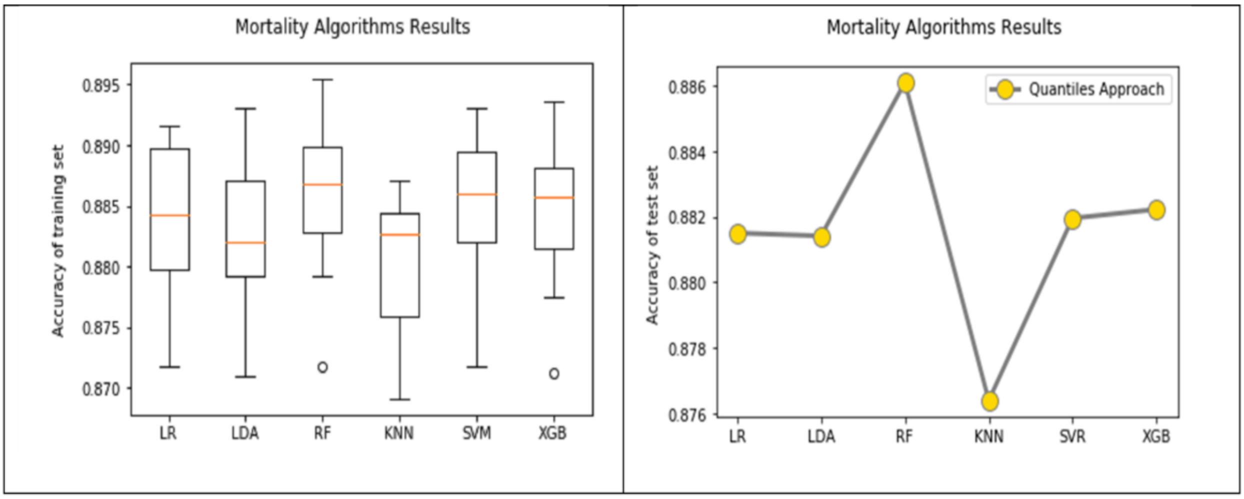
[e] RF: random forest.

[f] kNN: k-nearest neighbor.

[g] SVM: support vector machine.

[h] XGB: extreme gradient boosting.

Figure 6. Comparison of the mortality model results using the quantiles approach on the training set (left) and the test set (right). LR: logistic regression; LDA: linear discriminant analysis; RF: random forest; KNN: k-nearest neighbor; SVM: support vector machine; XGB: extreme gradient boosting.



The ROC curve is commonly used to evaluate the performance of an ML model by showing the relationship between the false-positive and true-positive rates. The AUROC metric can be used as a basis for comparison; higher values indicate that a model can identify classes using a specific ML algorithm better than another. In the case of the mortality model, the ROC curve shows the relationship between survival cases that scored as no survival and no survival cases that scored as no survival.

Table 9 shows the AUROC results of the mortality model on both the training and test sets using the baseline and quantile approaches for the different ML algorithms. Figure 7 shows the ROC curves for the six ML algorithms for both the baseline and the quantiles approach. XGB produced the highest AUROC (0.79) for predicting mortality on the test set using the quantiles approach (Table 9).

Table 9. Mortality model performance based on area under the receiver operating characteristic curve (AUROC).

Method and algorithm	Training set AUROC, mean (SD)	Test set AUROC
Baseline approach		
LR[a]	0.702047 (0.015652)	0.69313
LDA[b]	0.701731 (0.016077)	0.69247
RF[c]	0.764875 (0.009214)	0.76725
kNN[d]	0.629262 (0.008944)	0.63173
SVM[e]	0.653269 (0.011730)	0.66800
XGB[f]	0.771187 (0.012094)	0.76971
Quantiles approach		
LR	0.727331 (0.014217)	0.72810
LDA	0.725909 (0.014758)	0.72622
RF	0.783696 (0.010503)	0.78292
KNN	0.631649 (0.010416)	0.64087
SVM	0.719253 (0.008940)	0.72333
XGB	0.788908 (0.010665)	0.79036

[a] LR: logistic regression.

[b] LDA: linear discriminant analysis.

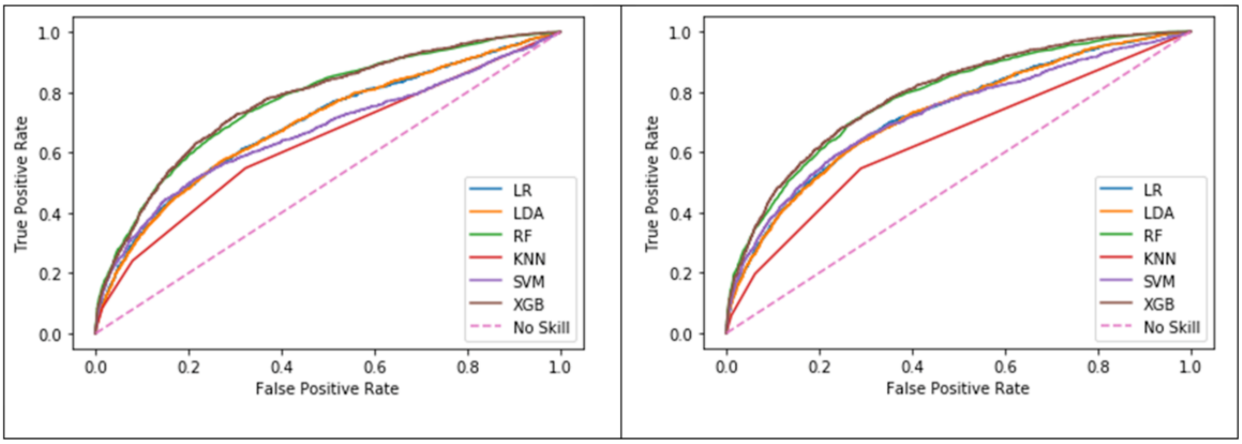
[c] RF: random forest.

[d] kNN: k-nearest neighbor.

[e] SVM: support vector machine.

[f] XGB: extreme gradient boosting.

Figure 7. Comparison of receiver operating characteristic curves in the mortality model using the baseline (left) and the quantiles approach (right). LR: logistic regression; LDA: linear discriminant analysis; RF: random forest; KNN: k-nearest neighbour; SVM: support vector machine; XGB: extreme gradient boosting.



LOS Prediction Model

Binary Classification Algorithms

Table 10 shows the performance of the binary classification models for the LOS model on both the training set and test set using the baseline and the quantiles approaches with the six different ML algorithms.

Table 10. Length of stay model results for six algorithms using different performance metrics.

Method and algorithm	Training set accuracy, mean (SD)	Test set accuracy (95% CI)	Test set sensitivity (95% CI)	Test set specificity (95% CI)	Test set NPV[a] (95% CI)	Test set PPV$_b$ (95% CI)
Baseline approach						
LR[c]	0.5787 (0.01)	0.5715 (0.57-0.58)	0.56 (0.554-0.564)	0.59 (0.58-0.59)	0.57 (0.563-0.573)	0.58 (0.57-0.58)
LDA[d]	0.5787 (0.01)	0.5710 (0.57-0.58)	0.56 (0.55-0.56)	0.59 (0.58-0.59)	0.57 (0.56-0.57)	0.58 (0.57-0.58)
RF[e]	0.6205 (0.01)	0.6193 (0.62-0.63)	0.61 (0.60-0.61)	0.63 (0.63-0.64)	0.61 (0.61-0.62)	0.63 (0.62-0.63)
kNN[f]	0.5639 (0.01)	0.5713 (0.57-0.58)	0.52 (0.520-0.529)	0.62 (0.616-0.627)	0.56 (0.559-0.569)	0.58 (0.58-0.59)
SVM[g]	0.6228 (0.01)	0.6141 (0.61-0.62)	0.56 (0.56-0.57)	0.67 (0.66-0.68)	0.60 (0.60-0.61)	0.63 (0.63-0.64)
XGB[h]	0.6303 (0.01)	0.6130 (0.61-0.62)	0.58 (0.58-0.59)	0.64 (0.64-0.65)	0.60 (0.60-0.61)	0.62 (0.62-0.63)
Quantiles approach						
LR	0.6126 (0.01)	0.6131 (0.61-0.62)	0.59 (0.59-0.60)	0.63 (0.629-0.640)	0.61 (0.60-0.61)	0.62 (0.62-0.63)
LDA	0.6131 (0.01)	0.6130 (0.61-0.62)	0.59 (0.59-0.60)	0.64 (0.63-0.64)	0.61 (0.60-0.61)	0.62 (0.62-0.63)
RF	0.6511 (0.01)	0.6461 (0.64-0.65)	0.64 (0.63-0.66)	0.66 (0.65-0.66)	0.64 (0.64-0.65)	0.65 (0.65-0.66)
kNN	0.5748 (0.01)	0.5768 (0.57-0.58)	0.4865 (0.483-0.49)	0.6681 (0.66-0.68)	0.56 (0.56-0.57)	0.60 (0.59-0.60)
SVM	0.6466 (0.01)	0.6386 (0.63-0.65)	0.5939 (0.59-0.60)	0.68 (0.68-0.69)	0.63 (0.62-0.63)	0.66 (0.65-0.66)
XGB	0.6496 (0.01)	0.6284 (0.62-0.64)	0.61 (0.60-0.62)	0.65 (0.64-0.66)	0.62 (0.62-0.63)	0.64 (0.63-0.64)

[a] NPV: negative predictive value.

[b] PPV: positive predictive value.

[c] LR: logistic regression.

[d] LDA: linear discriminant analysis.

[e] RF: random forest.

[f] kNN: k-nearest neighbor.

[g] SVM: support vector machine.

[h] XGB: extreme gradient boosting.

The best accuracy of predicting ICU LOS on the test set was 64.64% using the RF algorithm in the quantiles approach, followed by the SVM algorithm with an accuracy of 63.86%. The improvement in model accuracy from the baseline approach to the quantiles approach was better when compared with that found for the mortality model (Table 8). For example, the difference in accuracy between the baseline and the quantiles approach for the LOS model on the test set was 2.68% using RF and was 2.45% using SVM. The RF algorithm achieved the highest sensitivity (0.64), which indicates that the model using the RF algorithm can identify patients who will stay in the ICU for more than 2.64 days better than the other algorithms. SVM achieved the highest specificity (0.68), which indicates that the model using the SVM algorithm is better at identifying patients who will stay in the ICU for 2.64 days or less compared with the other algorithms. Figure 8 shows a visual comparison of the accuracy of the six algorithms in the LOS model results using the quantiles approach. The box plots on the left show the model accuracy on the training set using 10-fold cross-validation and the graph on the right shows the one-time model accuracy on the testing set.

Table 11 shows the AUROC results of the LOS model on both the training and test sets using the baseline and the quantiles approach with the six ML algorithms. Figure 9 shows the ROC curves for the algorithms in the baseline approach and the quantiles approach, respectively. The RF algorithm using the quantiles approach produced the highest AUROC (0.697) for predicting the LOS on the test set (Table 11).

Figure 8. Comparison of the length of stay model results using the quantiles approach on the training set (left) and the test set (right). LR: logistic regression; LDA: linear discriminant analysis; RF: random forest; KNN: k-nearest neighbor; SVM: support vector machine; XGB: extreme gradient boosting.

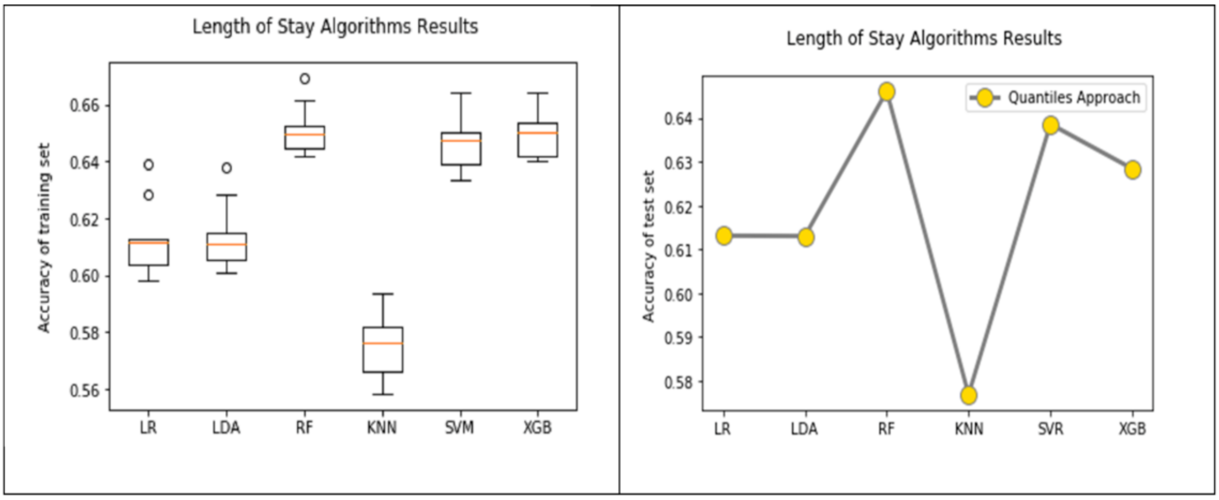


Table 11. Performance of the length of stay model results based on the area under the receiver operating characteristic curve (AUROC).

Method and algorithm	Training set AUROC, mean (SD)	Test set AUROC
Baseline approach		
LR[a]	0.612883 (0.006047)	0.60833
LDA[b]	0.612776 (0.006058)	0.60837
RF[c]	0.664959 (0.006147)	0.66325
kNN[d]	0.583710 (0.006401)	0.59110
SVM[e]	0.665992 (0.006041)	0.66118
XGB[f]	0.677454 (0.007311)	0.66586
Quantiles approach		
LR	0.654390 (0.012180)	0.65407
LDA	0.654178 (0.012102)	0.65384
RF	0.705115 (0.010004)	0.69782
kNN	0.598228 (0.007539)	0.60507
SVM	0.694473 (0.009834)	0.69272
XGB	0.704889 (0.011338)	0.69693

[a] LR: logistic regression.

[b] LDA: linear discriminant analysis.

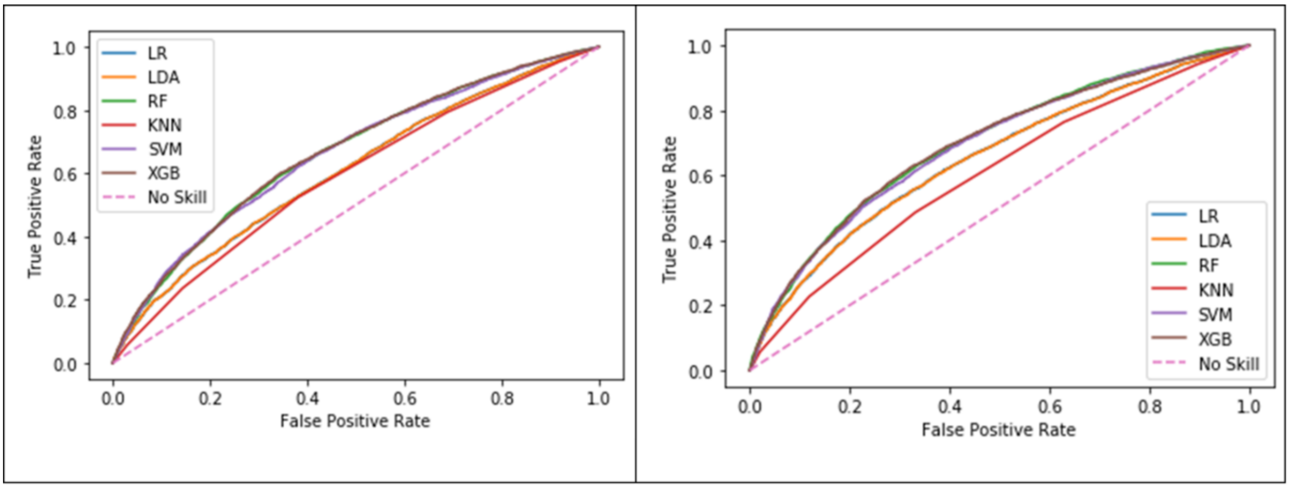
[c] RF: random forest.

[d] kNN: k-nearest neighbor.

[e] SVM: support vector machine.

[f] XGB: extreme gradient boosting.

Figure 9. Comparison of receiver operating characteristic curves in the length of stay model using the baseline (left) and quantiles (right) approaches. LR: logistic regression; LDA: linear discriminant analysis; RF: random forest; KNN: k-nearest neighbor; SVM: support vector machine; XGB: extreme gradient boosting.



Regression-Based Classifiers

As for the regression-based classifiers of the LOS model, we report the error between the predicted values and actual values in the test set using both the mean absolute error (MAE) and the root mean squared error metrics. The minimum, mean, and maximum LOS for the entire population was 1, 2.64, and 173.07 days, respectively. Table 12 shows the error value (per day) using both error metrics for the LOS model. The lowest error value obtained was 2.81 days using the MAE in the SVR algorithm with the quantiles approach.

Table 12. Regression error values of the length of stay model using the baseline and quantile approaches.

Method	MAE[a]	RMSE[b]
Baseline approach		
MLR[c]	3.509	6.029
SVR[d]	2.857	6.214
Quantiles approach		
MLR	3.446	5.961
SVR	2.810	6.137

[a] MAE: mean absolute error.

[b] RMSE: root mean square error.

[c] MLR: multivariate linear regression.

[d] SVR: support vector regression.

Discussion

Principal Results

Our findings indicate that we can build prediction models for ICU LOS and mortality with better accuracy using a combination of ML and the quantiles approach including only vital signs. Little improvement in the accuracy of the mortality model was achieved, but improvement of approximately 2.7% was achieved in the LOS model using the proposed quantiles approach. We examined model calibration across deciles for all six algorithms in both models. Figure 10 shows the probability calibration curves of the mortality model using the six algorithms. The six plots show good calibration of the models, especially in the case of the RF algorithm. Figure 11 shows the probability calibration curves of the LOS model using the six algorithms. The six plots show good calibration of the models except for the kNN algorithm.

Figure 10. Probability calibration curves of the mortality model for the six classification algorithms. LR: logistic regression; LDA: linear discriminant analysis; RF: random forest; KNN: k-nearest neighbor; SVM: support vector machine; XGB: extreme gradient boosting.

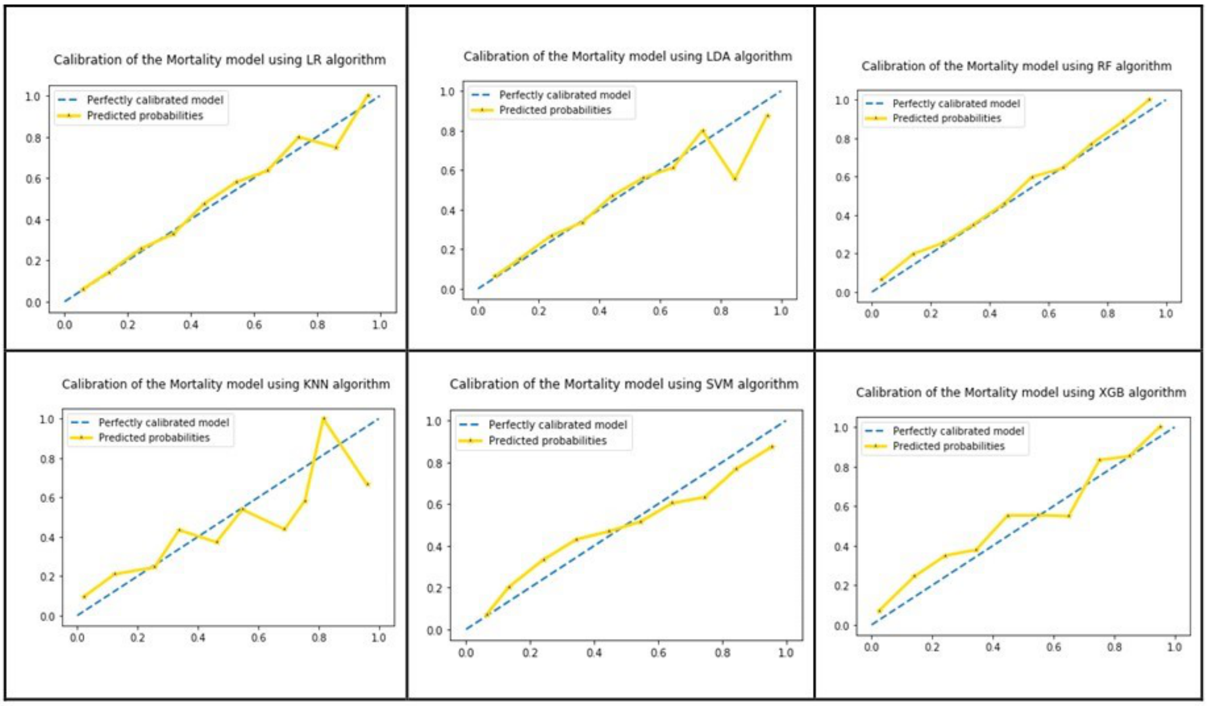
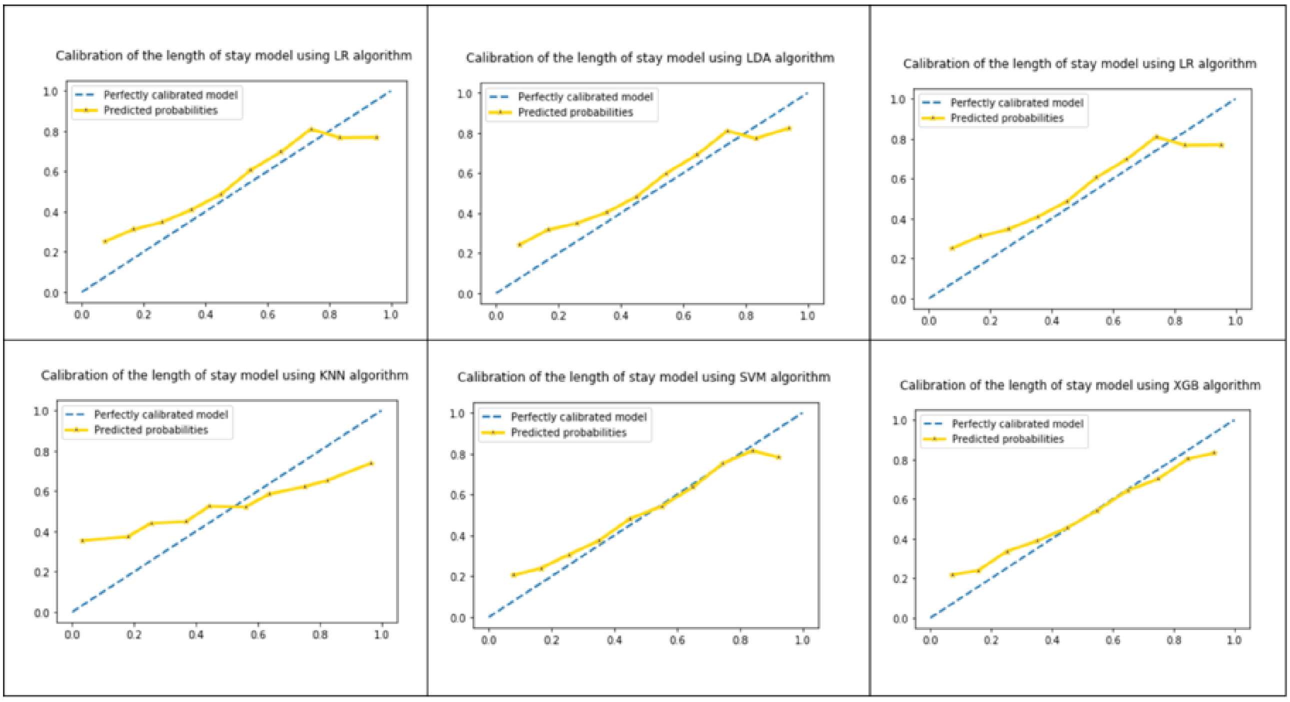


Figure 11. Probability calibration curves of the length of stay model for the six classification algorithms. LR: logistic regression; LDA: linear discriminant analysis; RF: random forest; KNN: k-nearest neighbor; SVM: support vector machine; XGB: extreme gradient boosting.



One might argue that we included only the mean and not the SD of the vital signs in the baseline approach when the comparison was to a model including both the mean and SD in the quantiles approach. Both the baseline and the quantile approaches include the means of vital signs. The quantiles approach includes an extra 21 features corresponding to modified means and modified SDs of the original values in addition to the quantile percentages. Had we chosen to include both the mean and SD of the original vital sign observations in the baseline approach, we would have also needed to include the SD of the original vital sign observations in the quantiles approach. In this case, we do not expect that there will be a significant impact.

Moreover, based on the method of population selection, the same patient could be in the training as well as in the test set but for different ICU admissions at different time points. For this study, we considered unique ICU admissions as opposed to unique patient identifiers. The rationale for focusing on unique admissions is that we sought to predict mortality and LOS without having prior knowledge about a patient's medical conditions or diagnoses.

Qualitative Comparison With Other Approaches

For the mortality model, we were able to achieve approximately 89% accuracy and an AUROC of 0.78 using only 7 vital sign features and 4 demographic attributes, along with 21 features engineered from the original features. Other researchers have used excessively more features to achieve similar or better accuracies. For instance, Johnson et al [12] used a total of 148 features to achieve an AUROC of 0.92. Lehman et al [14] applied the SAPS-I algorithm on commonly used physiological data to predict mortality and achieved an AUROC of 0.72. Johnson et al [13] used a range of features, including standard statistical descriptors, to achieve an AUROC of 0.86.

For LOS models, most researchers used an exhaustive list of features to achieve higher accuracy in their models, but they did not report on whether they had balanced classification problems. For example, Harutyunyan et al [16] achieved 84% accuracy using 17 clinical variables and by considering a target ICU LOS of 7 days. Gentimis et al [17] achieved 80% accuracy using several inputs from seven tables to build the LOS model with a target ICU stay of 5 days. Bertsimas et al [18] used several static and dynamic variables, and achieved accuracy in the >80% range. In our approach, we built balanced classification models (using the median LOS of the entire population) with minimal features. These two conditions made it harder to achieve high accuracy, which reached only 65% in the LOS model.

One contribution of our method is the unique combination of ML with the quantiles approach. Other researchers have used various techniques to assess a patient's deteriorating conditions. Tyler et al [15] found that the methods to normalize patients' abnormal values are not thoroughly correct and might affect the results negatively. Other researchers relied on scoring systems (eg, centile-based early warning score, National Early Warning Score, or SAPS) to estimate or recognize patients' deteriorating conditions. We avoided relying on existing early warning scoring systems since they vary from patient to patient, which may lead to uncertain results.

Sensitivity Analysis

Since we considered unique ICU stays rather than individual patients, the training set/testing set split was not performed at the patient level. This might raise the concern that the vital signs and LOS measured at different ICU visits for the same patient could be highly correlated. Thus, the mortality and the LOS models might risk overestimation in predictive performance. We mitigated this effect by performing a sensitivity analysis to compare the results of the models after excluding patient overlap to the results of the original model with the overlap included. In the original model, the population size was 44,626 (corresponding to ICU stays), the training set size was 33,469 ICU stays (75% of the population), and the test set size was 11,157 ICU stays (25% of the population).

The patient overlap between the training and test sets was 3886 ICU stays (34.83% of the test set). The number of ICU stays remaining in the test set after removing the patient overlap (ie, 3886) reduced to 7271 (65.17% of the original test set of size 11,157). Table 13 shows the results of the mortality model after removing the overlap and Table 14 shows the results of the LOS model after removing the overlap. There were no significant changes compared to the model results shown in Table 8 and Table 10, respectively.

Table 13. Mortality model results for six algorithms using different performance metrics.

Methods and algorithm	Training set accuracy, mean (SD)	Test set accuracy (95% CI)	Test set sensitivity (95% CI)	Test set specificity (95% CI)	Test set NPV[a] (95% CI)	Test set PPV[b] (95% CI)
Quantiles approach without overlap in the test set						
LR[c]	0.88263 (0.0058)	0.87636 (0.870-0.886)	0.0620 (0.062-0.063)	0.9963 (0.989-1.007)	0.8781 (0.872-0.888)	0.7160 (0.711-0.724)
LDA[d]	0.88171 (0.0058)	0.87663 (0.870-0.886)	0.0974 (0.097-0.099)	0.9914 (0.984-1.002)	0.8817 (0.875-0.891)	0.6275 (0.623-0.635)
RF[e]	0.88458 (0.0061)	0.88145 (0.875-0.891)	0.0952 (0.095-0.097)	0.9973 (0.990-1.008)	0.8820 (0.876-0.892)	0.8396 (0.834-0.849)
kNN[f]	0.87645 (0.0054)	0.87196 (0.866-0.881)	0.0620 (0.062-0.063)	0.9913 (0.984-1.002)	0.8776 (0.871-0.887)	0.5132 (0.510-0.519)
SVM[g]	0.88365 (0.0058)	0.87581 (0.870-0.885)	0.0428 (0.042-0.044)	0.9985 (0.991-1.009)	0.8762 (0.875-0.891)	0.8163 (0.811-0.825)
XGB[h]	0.88422 (0.0061)	0.87966 (0.873-0.889)	0.1670 (0.166-0.169)	0.9846 (0.978-0.995)	0.8891 (0.883-0.899)	0.6166 (0.612-0.624)
Quantiles approach						
LR	0.88380 (0.0063)	0.88150 (0.875-0.891)	0.0545 (0.054-0.055)	0.9960 (0.989-1.007)	0.8838 (0.878-0.893)	0.6548 (0.650-0.662)
LDA	0.88210 (0.0067)	0.88141 (0.875-0.891)	0.0935 (0.093-0.095)	0.9905 (0.983-1.001)	0.8875 (0.881-0.897)	0.5772 (0.573-0.584)
RF	0.88586 (0.0064)	0.88608 (0.880-0.896)	0.0891 (0.089-0.090)	0.9964 (0.989-1.007)	0.8876 (0.881-0.897)	0.7756 (0.770-0.784)
KNN	0.88018 (0.0060)	0.87640 (0.870-0.886)	0.0589 (0.059-0.060)	0.9895 (0.982-1.000)	0.8836 (0.877-0.893)	0.4395 (0.437-0.445)
SVM	0.88511 (0.0058)	0.88195 (0.876-0.892)	0.0449 (0.045-0.046)	0.9816 (0.991-1.009)	0.8829 (0.877-0.893)	0.7439 (0.739-0.752)
XGB	0.88443 (0.0061)	0.88222 (0.875-0.891)	0.1643 (0.164-0.167)	0.9816 (0.975-0.992)	0.8945 (0.888-0.904)	0.5533 (0.550-0.560)

[a] NPV: negative predictive value.

[b] PPV: positive predictive value.

[c] LR: logistic regression.

[d] LDA: linear discriminant analysis.

[e] RF: random forest.

[f] kNN: k-nearest neighbor.

[g] SVM: support vector machine.

[h] XGB: extreme gradient boosting.

Table 14. Length of stay model results for six algorithms using different performance metrics.

Method and algorithm	Training set accuracy, mean (SD)	Test set accuracy (95% CI)	Test set sensitivity (95% CI)	Test set specificity (95% CI)	Test set NPV[a] (95% CI)	Test set PPV[b] (95% CI)
Quantiles approach without overlap in the test set						
LR[c]	0.61262 (0.0117)	0.61312 (0.609-0.620)	0.5983 (0.594-0.605)	0.6267 (0.622-0.634)	0.6292 (0.625-0.636)	0.5957 (0.592-0.602)
LDA[d]	61.307 (0.0112)	0.61216 (0.608-0.619)	0.5951 (0.591-0.602)	0.6277 (0.624-0.635)	0.6277 (0.624-0.635)	0.5951 (0.591-0.602)
RF[e]	0.65108 (0.0081)	0.64778 (0.643-0.655)	0.6400 (0.636-0.647)	0.6550 (0.651-0.662)	0.6643 (0.660-0.672)	0.6304 (0.626-0.637)
kNN[f]	0.57483 (0.0104)	0.58740 (0.583-0.594)	0.4941 (0.491-0.500)	0.6731 (0.669-0.681)	0.5913 (0.587-0.598)	0.5816 (0.578-0.588)
SVM[g]	0.64659 (0.0088)	0.64379 (0.639-0.651)	0.5946 (0.591-0.601)	0.6890 (0.684-0.697)	0.6489 (0.645-0.656)	0.6374 (0.633-0.645)
XGB[h]	0.64961 (0.0076)	0.63540 (0.631-0.642)	0.6132 (0.609-0.620)	0.6557 (0.651-0.663)	0.6483 (0.644-0.656)	0.6209 (0.617-0.628)
Quantiles approach						
LR	0.61262 (0.0117)	0.61307 (0.609-0.620)	0.5930 (0.589-0.600)	0.6332 (0.629-0.640)	0.6058 (0.602-0.613)	0.6208 (0.617-0.628)
LDA	0.61307 (0.0112)	0.61298 (0.609-0.620)	0.5909 (0.587-0.598)	0.6352 (0.631-0.642)	0.6053 (0.601-0.612)	0.6212 (0.617-0.628)
RF	0.65108 (0.0081)	0.64614 (0.642-0.653)	0.6374 (0.633-0.645)	0.6549 (0.650-0.662)	0.6408 (0.636-0.648)	0.6516 (0.647-0.659)
KNN	0.57483 (0.0104)	0.57677 (0.573-0.583)	0.4865 (0.483-0.492)	0.6681 (0.664-0.676)	0.5624 (0.559-0.569)	0.5974 (0.593-0.604)
SVM	0.64659 (0.0088)	0.63861 (0.634-0.646)	0.5939 (0.590-0.601)	0.6838 (0.679-0.691)	0.6245 (0.620-0.632)	0.6553 (0.651-0.663)
XGB	0.64961 (0.0076)	0.62839 (0.624-0.635)	0.6085 (0.604-0.615)	0.6484 (0.644-0.656)	0.6206 (0.617-0.628)	0.6367 (0.632-0.644)

[a] NPV: negative predictive value.

[b] PPV: positive predictive value.

[c] LR: logistic regression.

[d] LDA: linear discriminant analysis.

[e] RF: random forest.

[f] kNN: k-nearest neighbor.

[g] SVM: support vector machine.

[h] XGB: extreme gradient boosting.

The total number of ICU stays was 44,626 and the total number of patients was 33,466. We calculated the frequency of ICU stays for the entire patient population. We found that 80% of the population visited the ICU only once and 20% visited the ICU more than once. Moreover, the MIMIC database includes data for patients who might have stayed in different ICU types (eg, general, cardiac) and due to different health conditions. In addition, a patient might have visited one ICU more frequently than another, and the time period between consecutive visits within a single ICU might be several years. The sensitivity analysis findings in our case might be due to the fact that our approach focused on the visits rather than the patients and ignored the details mentioned above.

Limitations

Admittedly, this study lacks quantitative comparisons with previous research on the same topic owing to substantial differences between the research questions tackled previously, and the associated data extraction pipelines and assumptions. We mitigated this limitation by providing a qualitative comparison between our models and previous models.

Previous research based on data from the MIMIC database likely demonstrated higher accuracy since excessively more features were used than applied in this study. We believe that it is difficult to achieve high model accuracy using a limited number of features.

Additionally, as in any ML-based method, our approach might have some limitations. In this study, we used the MIMC database, which represents a patient population from a single hospital in Boston, and does not generalize to other populations or hospital systems in other areas across the United States or the rest of the world. Future research will focus on applying our approach to other patient populations.

Moreover, we ran the models using only the vital signs to measure the impact of the demographic attributes. We found that the effects of demographic attributes on the results were low. For example, age did not have a considerable effect since we were only using adult patient data in the MIMIC database. The accuracy of the mortality model without the age feature using RF in the quantiles approach was 88.536%, which is very close to the model result obtained when including age. The mortality model achieved an AUROC of 0.77 without using age and 0.78 with age included. The accuracy of the LOS model without including the age feature using RF and the quantiles approach was 64.39%, which is very close to the result obtained with the age feature included. Table 15 also shows that the differences in AUROC and positive predictive value were not significant between the mortality and LOS models both including and excluding the age feature using the RF algorithm and the quantiles approach. This would be different in pediatrics and adolescent populations, for whom vital measurements are more age-sensitive. In addition, in the MIMIC database, the ages for patients older than 89 years are not accurate; we used 90 years as a dummy value for all of these patients. Another potential reason for the low impact of including the demographic attributes is the lack of variation in height due to missing values that had to be imputed using the population mean.

Table 15. Model results including and excluding the age feature.

Model	Accuracy	AUROC[a]	PPV[b] (95% CI)
Mortality			
Without age	88.536	0.76740	0.7468 (0.742-0.755)
With age	88.608	0.78292	0.7756 (0.770-0.784)
Length of stay			
Without age	64.390	0.69433	0.6487 (0.644-0.656)
With age	64.614	0.69782	0.6516 (0.647-0.659)

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

[b] PPV: positive predictive value.

Clinical Implications

Health professionals (ie, physicians, nurses, ICU specialists) can benefit from the advanced accurate predictive capabilities of the intelligent ICU patient monitoring module to help make better decisions regarding major challenges in health care, including bed management, patient flow, stock management, and effective provision of medical supplies. Poor bed management may result in the rejection of new patients, and a reduction in hospital revenue and overall quality of health services [22]. Patient flow involves making decisions regarding admissions, transfers, and referrals. Hospital administration needs solutions that enable reducing waste and wait times, and to increase service efficiency and productivity. Such tools need to consider the uncertainty of patients' recovery status. Poor stock management results in resource shortage or expiration, especially in the ICU where care should be delivered promptly. Thus, integrating the predictive functionalities of the intelligent ICU patient monitoring module within existing decision support platforms and clinical workflows may have several practical implications for improving the quality of care and reducing costs.

Conclusions

In this article, we proposed a novel approach for predictive modeling with reasonable performance based on a combination of ML algorithms and the quantiles approach that utilizes only vital signs available in the patient's profile without having to use external features. Using this quantiles approach, we engineered additional features by calculating the modified means, SDs, and quantile percentages from the baseline vital sign measures, which provided us with a richer dataset to achieve better predictive power in our models. We applied our approach to build two prediction models: one for mortality prediction and another for ICU LOS. Although the accuracy of the mortality model showed minimal improvement, we achieved better results in the LOS model by around 2.7%.

Intelligent ICU patient monitoring is a promising solution that will improve clinical workflows and enable hospitals to deliver higher-quality, cost-effective patient care, and to improve the overall quality of medical services in the ICU. The solution will support ICUs to put steps ahead and "nudge" health care providers to prepare for unexpected general health conditions of patients and better manage ICU facilities [23]. By relying on a minimal set of features that can be continuously collected from both inside and outside hospital systems and without requiring sophisticated medical devices, our predictive models can be used in cloud-based IRPM systems (see Exhibit X [24], a short video demonstrating the tool in action).

Relying on fewer features will be more feasible for realizing ML algorithms in real-world settings. Future directions of this research will involve adding more predictive modeling capabilities to the intelligent ICU patient monitoring module, including ICU readmission, severity level, and next-day patient vital sign measurements. We are currently working on applying this approach to a wider range of hospital systems within different geographic locations. Integrating intelligent ICU patient monitoring within existing clinical workflows and decision support platforms can support many hospitals in improving the quality of care and reducing costs.

Acknowledgments

This research was partially supported by King Fahad Medical City, Saudi Arabia. The authors would like to thank Drs. Robert L Davis, and Fatma Gunturkun from the Center for Biomedical Informatics, Department of Pediatrics, University of Tennessee Health Science Center, Memphis, Tennessee, for their feedbacks and insights.

Conflicts of Interest

None declared.

References

1. Shaban-Nejad A, Michalowski M, Peek N, Brownstein JS, Buckeridge DL. Seven pillars of precision digital health and medicine. Artif Intell Med 2020 Mar;103:101793. [doi: 10.1016/j.artmed.2020.101793] [Medline: 32143798]
2. Shaban-Nejad A, Michalowski M, Buckeridge DL. Health intelligence: how artificial intelligence transforms population and personalized health. NPJ Digit Med 2018;1:53. [doi: 10.1038/s41746-018-0058-9] [Medline: 31304332]
3. Riazanov A, Klein A, Shaban-Nejad A, Rose GW, Forster AJ, Buckeridge DL, et al. Semantic querying of relational data for clinical intelligence: a semantic web services-based approach. J Biomed Semantics 2013 Mar 13;4(1):9 [FREE Full text] [doi: 10.1186/2041-1480-4-9] [Medline: 23497556]
4. Alghatani K, Rezgui A. A cloud-based intelligent remote patient monitoring architecture. : CSREA Press; 2019 Presented at: International Conference on Health Informatics & Medical Systems; July 2019; Las Vegas, NV p. 29-21.
5. Cuthbertson BH, Boroujerdi M, McKie L, Aucott L, Prescott G. Can physiological variables and early warning scoring systems allow early recognition of the deteriorating surgical patient? Crit Care Med 2007 Feb;35(2):402-409. [doi: 10.1097/01.CCM.0000254826.10520.87] [Medline: 17205002]
6. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000 Jun 13;101(23):E215-E220. [doi: 10.1161/01.cir.101.23.e215] [Medline: 10851218]
7. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035. [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
8. Johnson A, Pollard T, Mark R. The MIMIC-III Clinical Database 2016. PhysioNet. URL: https://physionet.org/content/mimiciii/1.4/ [accessed 2016-09-04]
9. McCarthy A, Williams C. Predicting patient state-of-health using sliding window and recurrent classifiers. arXiv. 2016 Dec 02. URL: https://arxiv.org/pdf/1612.00662.pdf [accessed 2021-04-22]
10. Zhu Y, Fan X, Wu J, Liu X, Shi J, Wang C. Predicting ICU mortality by supervised bidirectional LSTM networks. 2018 Jul Presented at: IJCAI 2018 Joint Workshop on Artificial Intelligence in Health (AIH 2018); July 13-14, 2018; Stockholm, Sweden p. 49-60.
11. Johnson A, Pollard T, Mark R. Reproducibility in critical care: a mortality prediction case study. 2017 Nov 06 Presented at: Machine Learning for Healthcare Conference; August 18-19, 2017; Boston, MA p. 361-376.
12. Johnson AEW, Mark RG. Real-time mortality prediction in the Intensive Care Unit. AMIA Annu Symp Proc 2017;2017:994-1003 [FREE Full text] [Medline: 29854167]
13. Johnson A, Dunkley N, Mayaud L, Tsanas A, Kramer A, Clifford G. Patient specific predictions in the intensive care unit using a Bayesian ensemble. : IEEE; 2012 Sep 09 Presented at: 2012 Computing in Cardiology; September 9-12, 2012; Krakow, Poland p. 249-252.
14. Lehman L, Saeed M, Long W, Lee J, Mark R. Risk stratification of ICU patients using topic models inferred from unstructured progress notes. AMIA Annu Symp Proc 2012;2012:505-511 [FREE Full text] [Medline: 23304322]
15. Tyler PD, Du H, Feng M, Bai R, Xu Z, Horowitz GL, et al. Assessment of intensive care unit laboratory values that differ from reference ranges and association with patient mortality and length of stay. JAMA Netw Open 2018 Nov 02;1(7):e184521 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.4521] [Medline: 30646358]
16. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Sci Data 2019 Jun 17;6(1):96. [doi: 10.1038/s41597-019-0103-9] [Medline: 31209213]
17. Gentimis T, Ala' J, Durante A, Cook K, Steele R. Predicting hospital length of stay using neural networks on mimic iii data. 2017 Presented at: 2017 IEEE 15th International Conference on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech); . IEEE; November 6, 2017; Orlando, FL p. 1194-1201.
18. Bertsimas D, Pauphilet J, Stevens J, Tandon M. Predicting inpatient flow at a major hospital using interpretable analytics. medRxiv. 2020 Sep 16. URL: https://www.medrxiv.org/content/10.1101/2020.05.12.20098848v2 [accessed 2021-04-22]
19. Badawi O, Breslow MJ. Readmissions and death after ICU discharge: development and validation of two predictive models. PLoS One 2012 Nov 7;7(11):e48758 [FREE Full text] [doi: 10.1371/journal.pone.0048758] [Medline: 23144958]
20. Lo O, Fan L, Buchanan W, Thuemmler C. Technical evaluation of an e-health platform. 2012 Jul Presented at: IADIS E-Health; July 17-19, 2012; Lisbon, Portugal.

21. Ho TK. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Machine Intell 1998 Aug;20(8):832-844. [doi: 10.1109/34.709601]
22. Schmidt R, Geisler S, Spreckelsen C. Decision support for hospital bed management using adaptable individual length of stay estimations and shared resources. BMC Med Inform Decis Mak 2013 Jan 07;13:3 [FREE Full text] [doi: 10.1186/1472-6947-13-3] [Medline: 23289448]
23. Shaban-Nejad A, Mamiya H, Riazanov A, Forster AJ, Baker CJO, Tamblyn R, et al. From cues to nudge: a knowledge-based framework for surveillance of healthcare-associated infections. J Med Syst 2016 Jan;40(1):23. [doi: 10.1007/s10916-015-0364-6] [Medline: 26537131]
24. IRPM Prototype. URL: https://www.youtube.com/watch?v=9f25vDgM-qU [accessed 2020-11-07]

Abbreviations

AUROC: area under the receiver operating characteristic curve
ICU: intensive care unit
IRPM: Intelligent Remote Patient Monitoring
kNN: K-nearest neighbor
LOS: length of stay
LR: logistic regression
LSTM: long short-term memory
MAE: mean absolute error
MIMIC: Medical Information Mart for Intensive Care
ML: machine learning
MLR: multiple linear regression
PPF: percent point function
RF: random forest
ROC: receiver operating characteristic
SAPS: Simplified Acute Physiology Score
SpO$_2$: oxygen saturation
SVM: support vector machine
SVR: support vector regression
XGB: extreme gradient boosting

Edited by G Eysenbach; submitted 11.06.20; peer-reviewed by J op den Buijs, E Ahn; comments to author 29.07.20; revised version received 23.09.20; accepted 04.04.21; published 05.05.21.

Please cite as:
Alghatani K, Ammar N, Rezgui A, Shaban-Nejad A
Predicting Intensive Care Unit Length of Stay and Mortality Using Patient Vital Signs: Machine Learning Model Development and Validation
JMIR Med Inform 2021;9(5):e21347
URL: https://medinform.jmir.org/2021/5/e21347
doi:10.2196/21347
PMID:33949961

Original Paper

Automated Generation of Personalized Shock Wave Lithotripsy Protocols: Treatment Planning Using Deep Learning

Zhipeng Chen[1], PhD; Daniel D Zeng[2], PhD; Ryan G N Seltzer[3], PhD; Blake D Hamilton[4], MD

[1]Shenzhen Artificial Intelligence and Data Science Institute (Longhua), Longhua, Shenzhen, China
[2]The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3]Translational Analytics and Statistics, Tucson, AZ, United States
[4]School of Medicine, University of Utah, Salt Lake City, UT, United States

Corresponding Author:
Zhipeng Chen, PhD
Shenzhen Artificial Intelligence and Data Science Institute (Longhua)
Building 26, Technology Innovation Center
Hongshan 6979
Longhua, Shenzhen, 518110
China
Phone: 86 21071934
Email: zhipengchen@saidi.org.cn

Abstract

Background: Though shock wave lithotripsy (SWL) has developed to be one of the most common treatment approaches for nephrolithiasis in recent decades, its treatment planning is often a trial-and-error process based on physicians' subjective judgement. Physicians' inexperience with this modality can lead to low-quality treatment and unnecessary risks to patients.

Objective: To improve the quality and consistency of shock wave lithotripsy treatment, we aimed to develop a deep learning model for generating the next treatment step by previous steps and preoperative patient characteristics and to produce personalized SWL treatment plans in a step-by-step protocol based on the deep learning model.

Methods: We developed a deep learning model to generate the optimal power level, shock rate, and number of shocks in the next step, given previous treatment steps encoded by long short-term memory neural networks and preoperative patient characteristics. We constructed a next-step data set (N=8583) from top practices of renal SWL treatments recorded in the International Stone Registry. Then, we trained the deep learning model and baseline models (linear regression, logistic regression, random forest, and support vector machine) with 90% of the samples and validated them with the remaining samples.

Results: The deep learning models for generating the next treatment steps outperformed the baseline models (accuracy = 98.8%, F1 = 98.0% for power levels; accuracy = 98.1%, F1 = 96.0% for shock rates; root mean squared error = 207, mean absolute error = 121 for numbers of shocks). The hypothesis testing showed no significant difference between steps generated by our model and the top practices (*P*=.480 for power levels; *P*=.782 for shock rates; *P*=.727 for numbers of shocks).

Conclusions: The high performance of our deep learning approach shows its treatment planning capability on par with top physicians. To the best of our knowledge, our framework is the first effort to implement automated planning of SWL treatment via deep learning. It is a promising technique in assisting treatment planning and physician training at low cost.

(JMIR Med Inform 2021;9(5):e24721) doi:10.2196/24721

KEYWORDS

nephrolithiasis; extracorporeal shock wave therapy; lithotripsy; treatment planning; deep learning; artificial intelligence

Introduction

Shock wave lithotripsy (SWL, or extracorporeal shock wave lithotripsy) has been considered as a safe and effective noninvasive treatment option for nephrolithiasis since its introduction in early 1980s [1]. Reported SWL stone-free rates approach 74%-88% [2,3]; however, it is not without risk. Common contraindications to SWL include pregnancy, coagulopathy or use of platelet aggregation inhibitors, aortic aneurysms, severe untreated hypertension, and untreated urinary tract infections [4]. Failure of SWL treatment results in unnecessary exposure to various complications, such as loin pain, dysuria, analgesia, hematuria, and infection [3,5].

Given such risks, previous studies have identified proper patient selection, modifications in treatment technique, and employment of adjunctive measures as elements to improve SWL outcomes [6]. The treatment outcomes are strongly affected by a variety of preoperative patient characteristics (PPC), including BMI [7-9], stone location, overall stone burden [4], skin-to-stone distance [10,11], stone composition [12,13], stone density [14-17], and variation coefficients of stone density [18]. Various studies have also demonstrated that precise targeting [19,20] and tight coupling [21,22] increase fragmentation probability.

Appropriate control over shock wave delivery has a strong impact on treatment success and minimal complications. A treatment plan for shock wave delivery is a series of shock wave delivery steps with a specified power level, shock rate, and number of shocks; a successful sample SWL treatment plan is shown in Table 1. A plan precisely specifies step-by-step power levels, shock rates, and number of shocks. Each treatment step has a single power level, a constant shock rate, and shocks usually between 500-2500 [23-28]. Physicians are obliged to design plans that both deliver sufficient energy for breaking stones and minimize damage to body tissues. While the range of shock rates is typically 30-180 shocks/minute, a shock rate of 60-90 shocks/minute has been shown to improve efficacy [29-31] and decrease potential injury risks. The main reason is that the slower shock rate of 60-90 shocks/minute allows time for cavitation bubbles caused by the shock to disperse before the next shock arrives. Physicians can check stone fragmentation via x-ray. If the fragments of treated stones are ≤4 mm, they typically pass on their own without further treatment. An SWL treatment has to be stopped to reduce risks of tissue damage when the shock number reaches the maximum limit, even though the treated stone has not broken up.

Table 1. A sample shock wave lithotripsy (SWL) treatment plan.

Shock wave delivery steps	Power level	Shock rate (per minute)	Number of shocks
Step 1	1	120	100
Step 2	2	120	100
Step 3	3	120	100
Step 4	4	120	100
Step 5	5	120	100
Step 6	6	120	100
Step 7	7	120	100
Step 8	8	120	2300

Effective fragmentation leads to fewer shocks overall and therefore less damage to tissue [32,33]. In order to maximize treatment effect and control tissue damage, ramping protocols have been developed. The low-energy pretreatment allows for better pain management, thus preventing movement and subsequent decoupling of the shock head [34]. Clinical trials support that stepwise voltage ramping is associated with less tissue damage compared with a fixed maximal voltage protocol [23,25,26,35].

Although the strength, rates, and total number of shock waves are identified as the important factors of SWL treatment outcomes, there is no case-by-case guideline for physicians to optimize shock wave delivery protocols that take into account patient demographics and stone characteristics. The optimal energy delivery strategy remains controversial. In vitro and in vivo studies suggest that the strategy of ramping up shock wave energy is beneficial to improve fragmentation and stone clearance and limit renal damage, but clinical results are discordant [6,23]. In the current planning process, physicians adopt a trial-and-error approach to tune treatment plans. This approach involves nonintuitive iterations based on physicians' subjective decisions. Inexperienced physicians using this method may be more apt to produce inefficient or ineffective treatment plans. Such dependence on physicians' unique experience also leads to significant variability in the quality and consistency of treatment delivery. Moreover, different types of machines have different designs and different sources for generating shock waves. Therefore, an effective treatment plan for one machine may not transfer to a different machine.

As a result, SWL success rates are significantly different among physicians. Table 2 shows the percentiles of success rates of 171 physicians who recorded outcomes in the International Stone Registry, a database of accumulated treatment records for all patients treated within a national network of SWL services provided by Translational Analytics and Statistics, a lithotripsy service provider. Here, treatment success is defined as treated stone fragments ≤4 mm that typically pass on their own without further treatment. The top 20% of physicians have success rates higher than 94.3%, while the success rates of the bottom 20% of physicians are lower than 79.1%. Such variation indicates that the inexperience with and subjectivity of SWL treatment could lead to unnecessary damage to patients.

Table 2. Percentiles of treatment success rates.

Percentiles	Treatment success rates, %
Minimum	54.5
10th percentile	74.8
20th percentile	79.1
30th percentile	82.6
40th percentile	84.7
50th percentile	86.6
60th percentile	88.9
70th percentile	91.4
80th percentile	94.3
90th percentile	100
Maximum	100

Machine learning techniques have been applied in the planning process of high-quality personalized treatments, such as radiation therapies [36-39], chemotherapies [40,41] and diabetes treatments [42]. Most machine learning models only take independent vectors as inputs, so they are not suited to the sequential nature of SWL treatment plans. However, recurrent neural networks (RNNs) are naturally suited to temporal sequence inputs. Several variants like long short-term memory (LSTM) [43] and gated recurrent unit [44] have been developed for sequential features and applied to disease diagnosis [45,46]. Following these recent works, we aimed to validate the deep learning approach to generate next SWL treatment steps by learning the practices of top physicians and, based on the deep learning approach, develop a system to automatically produce personalized, unbiased, and consistent SWL treatment plans. The generated treatment plans can help physicians minimize the trial-and-error process and develop evidence-based personalized treatment based on PPC, including patient demographics and stone characteristics. An additional benefit is that this treatment planning framework can be generalized to different machine types, so physicians can easily adapt to new generations of SWL machines.

Methods

Data

To train and evaluate our models, we used a dataset of renal treatments with Storz SLX-T from the International Stone Registry provided by Translational Analytics and Statistics. Each treatment consisted of PPC and several treatment steps (ie, ternaries of a power level, a shock rate, and number of shocks). The power level ranged from 1 to 9. The options for shock rates were 60, 90, 120, and 180 shocks per minute. The maximum number of shocks was typically set at 3000 for renal stones. The PPC in our dataset included patient gender, age, stone location (one-hot encoding), stone size, mean arterial pressure before treatment, anticoagulant use, sedation use, whether multiple stones existed, and whether strapping was applied.

Our deep learning models were trained with the best treatment plans for obtaining the best planning capability. We selected 54 physicians in the top quartile of treatment success rates. These physicians had more than 91.4% treatment success rates. Then, we selected their successful treatment cases with no reported complications, in which they were stone free or had fragments ≤4 mm and typically passed on their own without further treatment. We identified 1216 cases in total and assumed these cases are the best practices in SWL treatment planning.

We then built the step dataset from the identified successful cases to train and evaluate the step generation model. We identified steps by power level change or shock rate change and limited the number of shocks to 1000 for each step, a natural step length in previous literature [25]. If more than 1000 shocks were delivered under the same power level and the same shock rate, we broke them into multiple steps with 1000 shocks maximum.

Then, we exhaustively decomposed each case into samples by step for the step generation task, where the ternary of each step was generated by its previous steps and PPC. An *n*-step treatment case was decomposed into $n - 1$ samples: we used the first *i* step(s) and PPC as the model inputs and the power level, shock rate, and number of shocks in the ($i + 1$)th step as the model outputs, where $0 < i < n$. For example, the SWL treatment case in Table 1 that consisted of 10 steps after the last 2300 shocks at power level 8 was split into 3 steps: (1) power level = 8, shock rate = 120, number of shocks = 1000; (2) power level = 8, shock rate = 120, number of shocks = 1000; and (3) power level = 8, shock rate = 120, number of shocks = 300. Then, we decomposed this case into 9 samples: (1) The input is the first step (power level = 1, shock rate = 120, number of shocks = 100) and PPC, and the output is the second step (power level = 2, shock rate = 120, number of shocks = 100); (2) the input is the first 2 steps and PPC, and the output is the third step (power level = 3, shock rate = 120, number of shocks = 100); …; and (9) the input is the first 9 steps and PPC, and the output is the last step (power level = 8, shock rate = 120, number of shocks = 300).

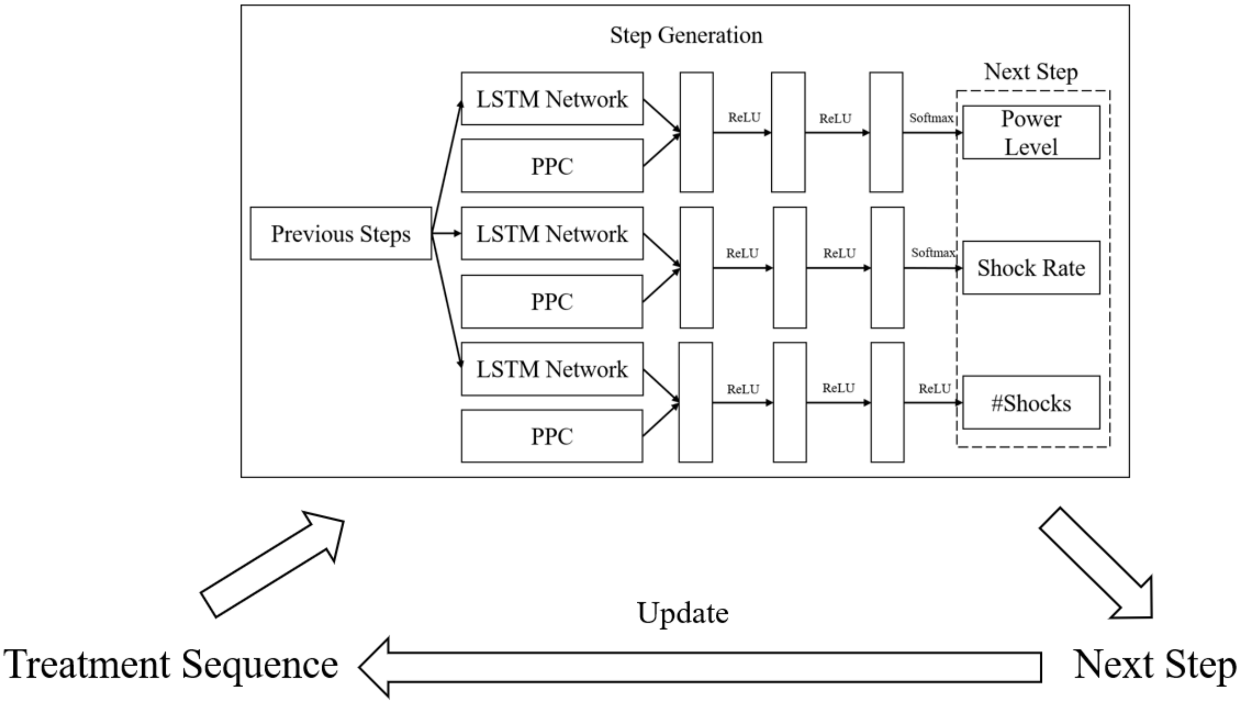
At last, we constructed 8583 samples for step generation. We randomly chose 90% of the samples for model training and used the remaining samples for validation. In the data split, we enforced that samples from the same treatment case were only contained within the same split.

Deep Learning for Step Generation

We first built deep neural networks to separately generate power levels, shock rates, and numbers of shocks for the next steps, given previous steps and PPC (Figure 1). Most off-the-shelf machine learning models only take inputs represented as independent vectors rather than a sequence of previous steps. However, RNNs are naturally suited to temporal sequence inputs, so we adopted an RNN variant, the LSTM model [43], which can keep track of arbitrary long-term dependencies in the input sequences, to encode the treatment sequences to vectors. More specifically, assume the *i*-th step is encoded as a vector x_i, then the LSTM model is defined iteratively as follows:

where the initial values c_0 and h_0 are zero vectors, ° denotes the element-wise product, σ is the sigmoid function, and h_1 is the representation of the first *i* treatment steps.

Figure 1. The framework for automated shock wave lithotripsy (SWL) treatment planning. LSTM: long short-term memory; PPC: preoperative patient characteristics; ReLU: rectifier linear unit.



Then, the encoded previous steps were concatenated to PPC vectors and fed to deep neural networks. In our implementation, we used 2 fully connected layers with a rectifier linear unit (ReLU) function as activation functions, because ReLU functions are nonsaturated and make the model less likely to overfit [47]. At last, we used different classifiers or regressors to generate power levels, shock rates and shock numbers. The formula are as follows:

where h_n is the *n* previous steps encoded by LSTM, and *p* denotes the PPC vector. The classifiers at the end of the networks were softmax functions for generating power level and shock rate because they are categorical, and we used categorical cross-entropy as the loss functions; for shock number generation in which the output is an integer, we used ReLU as the regressor and mean squared error (MSE) as the loss function. For all the deep neural networks, we chose the Adam SGD optimizer [48,49] in model training.

Statistical Analysis

We hypothesized that the deep learning approach is comparable to the treatment practices of top physicians and that it outperforms machine learning models which do not take treatment sequences as inputs. Thus, we compared the performance of the deep learning model and other up-to-date machine learning models.

Three classical machine learning approaches were selected as baselines for generating power level, shock rate, and number of shocks, respectively. We used logistic regression, random forest classifier (RFC), and support vector classifier (SVC) as the baseline models for power level generation and shock rate generation. We chose linear regression, random forest regression (RFR), and support vector regression (SVR) as the baseline models to generate the number of shocks. As these baseline models could not be fed with sequential data directly, the features for the baseline models were (1) the average power level, average shock rate, and average number of shocks in previous steps; (2) the power level, shock rate, and number of shocks in the last step; and (3) PPC.

We trained the deep learning models and baseline models with 90% of the samples. Then, we validated them with the remaining samples and calculated evaluation metrics. In the multiclass tasks of power level generation and shock rate generation, we used accuracy, macro-averaged precision, macro-averaged recall, and macro-averaged F1 score as the evaluation metrics [50,51]. Accuracy was defined as a ratio of correctly generated observations to the total observations. Suppose the number of categories is *n* and the confusion matrix of a classifier is a *n* x *n* matrix *C*, where $C_{i\,j}$ is the number of samples that is labeled as *i* but generated as *j*, then the accuracy is defined as

The precision and recall of category *k* are defined as

Macro-averaged precision and recall are the average of precisions and recalls for all categories:

The F1 score of category *k* is defined as the harmonic mean of precision and recall of category *k*

and macro-averaged F1 score is defined as the average of F1 scores for all categories:

Because the number of shocks is an integer, we used the root mean squared error (RMSE) and mean absolute error (MAE) as the metrics to evaluate the models generating the number of shocks and to measure the average magnitude of errors. At last, we conducted paired *t* test to detect the difference between treatment steps generated by machine learning models and treatment practices of top physicians.

Results

The deep learning models generated high-quality treatment steps and outperformed the baselines, as summarized in Tables 3-5. In power level generation (Table 3), the accuracy of the deep learning model was 0.988, and the precision, recall, and F1 scores were all 0.980. The best baseline was the SVC, for which the accuracy was 0.981, precision was 0.969, recall was 0.976, and F1 score was 0.972, lower than the performance of the deep learning model. For shock rate generation (Table 4), our model achieved an F1 score of 0.960 along with an accuracy of 0.981, precision of 0.963, and recall of 0.957. Among the baseline models, the logistic regression performed the best in accuracy and precision, at 0.978 and 0.932, respectively, while the RFC had the best recall and F1 score, at 0.986 and 0.956, respectively. Though the recall of the RFC and the logistic regression was better than that of the deep learning model, the accuracy, precision, and F1 score of our proposed model outperformed all the baseline models. The RMSE of the generation of the number of shocks (Table 5) by the deep learning model was 207, about 19% less than the best baseline model RFR. The MAE of the deep learning model was 121, about 23% less than the best baseline model.

Table 3. Model performance in power level generation.

Model	Accuracy	Precision	Recall	F1	*t* statistic	*P* value
Deep learning	0.988	0.980	0.980	0.980	0.707	.480
Logistic regression	0.974	0.964	0.964	0.964	1.257	.209
RFC[a]	0.708	0.823	0.859	0.803	4.976	<.001
SVC[b]	0.981	0.969	0.976	0.972	2.205	.028

[a]RFC: random forest classifier.

[b]SVC: support vector classifier.

Table 4. Model performance in shock rate generation.

Model	Accuracy	Precision	Recall	F1	*t* statistic	*P* value
Deep learning	0.981	0.963	0.957	0.960	0.277	.782
Logistic regression	0.978	0.932	0.960	0.945	2.331	.020
RFC[a]	0.952	0.930	0.986	0.956	2.064	.039
SVC[b]	0.976	0.926	0.956	0.939	2.510	.012

[a]RFC: random forest classifier.

[b]SVC: support vector classifier.

Table 5. Model performance in shock number generation.

Model	RMSE[a]	MAE[b]	*t* statistic	*P* value
Deep learning	207	121	0.350	.727
Linear regression	265	206	0.917	.359
RFR[c]	255	158	0.628	.530
SVR[d]	350	173	9.427	<.001

[a] RMSE: root mean squared error.

[b] MAE: mean absolute error.

[c] RFR: random forest regression.

[d] SVC: support vector regression.

The analysis also tested the difference between the generated step and the ground truth. In the paired *t* test result, there was no evidence indicating a difference between the generated steps of the deep learning model and treatment steps planned by top physicians, while the outputs of some baseline models significantly deviated from the ground truth. The power levels generated by the RFC and SVC, the shock rates generated by all the baseline models, and the numbers of shocks generated by the SVR were significantly different from the treatment steps in the successful SWL cases of top physicians.

Furthermore, we analyzed the performance of the deep learning models on samples of various treatment sequence lengths to gain a better understanding of how the treatment sequence information could aid decision making. We partitioned the validation dataset into 9 sets by the number of previous treatment steps and summarized the validation results in Tables 6-8. As shown in Table 6, the deep learning model was able to perfectly generate power levels when previous treatment steps were fewer than 6. As the number of previous treatment steps increases, the treatment becomes more complicated and leads to lower performance of power level generation by the deep learning model. The deep learning model reached the lowest accuracy (accuracy = 0.875) and lowest recall (recall = 0.500) in samples containing 9 previous treatment steps and the lowest precision (precision = 0.873) and lowest F1 score (F1 = 0.888) in samples containing 8 previous treatment steps. Similarly, the deep learning model generated highly accurate shock rates in samples with previous treatment steps fewer than 6; the model reached the lowest accuracy (accuracy = 0.889), lowest precision (precision = 0.857), and lowest recall (recall = 0.631) in samples containing 7 previous treatment steps and the lowest F1 score (F1 = 0.861) in samples containing 6 previous treatment steps (Table 7). For the performance of generating the number of shocks (Table 8), the RMSE and MAE generally increased as the number of previous treatment steps increased, and the maximum number of errors appeared in samples with 5 previous treatment steps (RMSE = 365; MAE = 310). The results show the excellent performance of deep learning models in step generation in the first 4-6 steps, where most successful cases end. It reflects the reliability of deep learning models in aiding treatment decision making. Longer treatment lengths typically indicate treatment difficulties; even our deep learning models cannot generate treatment steps with high accuracy in these rare cases.

Table 6. Power level generation performance in samples containing different numbers of previous treatment steps.

Number of previous treatment steps	Accuracy	Precision	Recall	F1
1	1.000	1.000	1.000	1.000
2	1.000	1.000	1.000	1.000
3	1.000	1.000	1.000	1.000
4	1.000	1.000	1.000	1.000
5	1.000	1.000	1.000	1.000
6	0.983	0.980	0.980	0.980
7	0.926	0.915	0.939	0.925
8	0.889	0.873	0.914	0.888
9	0.875	0.875	0.500	0.933

Table 7. Shock rate generation performance in samples containing different numbers of previous treatment steps.

Number of previous treatment steps	Accuracy	Precision	Recall	F1
1	1.000	1.000	1.000	1.000
2	0.992	0.997	0.972	0.984
3	1.000	1.000	1.000	1.000
4	1.000	1.000	1.000	1.000
5	0.992	0.997	0.974	0.985
6	0.975	0.976	0.802	0.861
7	0.889	0.857	0.631	0.888
8	0.917	0.864	0.642	0.902
9	1.000	1.000	1.000	1.000

Table 8. Performance of the generation of the number of shocks in samples containing different numbers of previous treatment steps.

Number of previous treatment steps	RMSE[a]	MAE[b]
1	31	25
2	32	17
3	34	24
4	139	60
5	365	310
6	317	233
7	273	190
8	275	242
9	99	76

[a] RMSE: root mean squared error.

[b] MAE: mean absolute error.

The validation showed that the capability of the deep learning model for step generation is on par with that of top physicians. Based on the high-quality step generation, we generated treatment plans by iteratively generating steps with the trained models (Figure 1). We started from an empty treatment sequence. We fed PPC and the current treatment sequence into the step generation model. The generated next step was then added to the current treatment sequence. We repeated such a process until the total number of shocks reached the upper limit. If a physician confirms stone fragmentation via x-ray before reaching the maximum limit of the number of shocks, they can stop immediately; if the number of shocks reaches the maximum limit, the physician has to stop for risk control. Thus, the generated treatment sequence is enough to guide practice. The specifications of individual lithotripters limit the maximum number of shocks per session to 2000-4500 [4], and for the majority of treatments of upper ureteral and renal stones, the range is 2000-3500 [52]. We used 3000 as the upper limit in our implementation, which is a typical shock limit in renal stone treatment practices and can be adjusted according to shock wave generating machines.

Discussion

Principal Findings

Previous literature has shown a series of work on standardizing SWL treatment [2,53]; however, energy delivery is still controversial and unclear [4,6], relying on physicians' subjective judgement. Manual treatment design is significantly affected by nonstandardizable radiographic appearance of stones, bias to a low power level for fear of complications, and preconceived expectations. Our study utilized deep learning to generate treatment steps and developed a framework for automated SWL treatment planning.

The analysis results revealed that deep learning models for treatment step generation effectively learn from SWL treatment plans and achieve the step generation capability of top physicians. The performance comparison indicated that utilization of a previous treatment sequence in deep learning improves the quality of generated steps. By iteratively generating treatment steps, our automated planning framework can avoid human biases and generate personalized, high-quality, and consistent SWL treatment plans based on PPC, including patient demographics and stone characteristics. With the help of these automatically generated treatment plans, physicians can minimize the trial-and-error process and implement evidence-based personalized treatment. This framework can be generalized to different machine types, so physicians can easily adapt to new generations of SWL machines.

Limitations

Our proposed model only learns and imitates the best practices, but cannot perform better than them. Even the best physician cannot plan successful SWL treatment plans for all cases, so successful difficult cases, including those requiring long treatment sequences, are rare for model training. Therefore, our model may be good at planning easier cases, but less adept in rare difficult cases, similar to physicians' actual practice. As the treatment cases, especially successful difficult cases, accumulate, our model is likely to gain an expert-level planning capability to handle difficult cases.

Due to data limitations, we were only able to consider a small set of patient demographics and stone characteristics. However, our framework can be easily extended to utilize a larger set of parameters than has previously been used. Moreover, the data are retrospective. Therefore, clinical studies are warranted to confirm the effectiveness and efficiency of this framework.

Conclusions

To the best of our knowledge, our framework is the first effort to implement automated planning of SWL treatment via deep learning. Its assistance for inexperienced urologists in designing SWL treatment plans is useful in both SWL treatment planning and physician training. While the applications of machine learning in diagnosis are becoming more mature, few studies exist in automated treatment plan generation. Our approach is a step forward in exerting the potential of machine learning in medical sciences.

Authors' Contributions

All authors designed the study. RGNS provided the data. ZC and DDZ developed the deep learning model for treatment plan generation. ZC implemented and evaluated the deep learning model and drafted the manuscript. All authors revised the manuscript.

Conflicts of Interest

RGNS was an employee of Translational Analytics and Statistics. BDH is a consultant for NextMed Management Services. The remaining authors declare no conflicts of interest.

References

1. Chaussy C, Brendel W, Schmiedt E. Extracorporeally induced destruction of kidney stones by shock waves. Lancet 1980 Dec 13;2(8207):1265-1268. [doi: 10.1016/s0140-6736(80)92335-1] [Medline: 6108446]
2. Assimos D, Krambeck A, Miller NL, Monga M, Murad MH, Nelson CP, et al. Surgical Management of Stones: American Urological Association/Endourological Society Guideline, PART I. J Urol 2016 Oct;196(4):1153-1160. [doi: 10.1016/j.juro.2016.05.090] [Medline: 27238616]
3. Al-Marhoon MS, Shareef O, Al-Habsi IS, Al Balushi AS, Mathew J, Venkiteswaran KP. Extracorporeal Shock-wave Lithotripsy Success Rate and Complications: Initial Experience at Sultan Qaboos University Hospital. Oman Med J 2013 Jul;28(4):255-259 [FREE Full text] [doi: 10.5001/omj.2013.72] [Medline: 23904918]
4. Reynolds LF, Kroczak T, Pace KT. Indications and contraindications for shock wave lithotripsy and how to improve outcomes. Asian J Urol 2018 Oct;5(4):256-263 [FREE Full text] [doi: 10.1016/j.ajur.2018.08.006] [Medline: 30364729]
5. Skolarikos A, Alivizatos G, de la Rosette J. Extracorporeal shock wave lithotripsy 25 years later: complications and their prevention. Eur Urol 2006 Nov;50(5):981-90; discussion 990. [doi: 10.1016/j.eururo.2006.01.045] [Medline: 16481097]
6. McClain PD, Lange JN, Assimos DG. Optimizing shock wave lithotripsy: a comprehensive review. Rev Urol 2013;15(2):49-60 [FREE Full text] [Medline: 24082843]
7. Pareek G, Armenakas NA, Panagopoulos G, Bruno JJ, Fracchia JA. Extracorporeal shock wave lithotripsy success based on body mass index and Hounsfield units. Urology 2005 Jan;65(1):33-36. [doi: 10.1016/j.urology.2004.08.004] [Medline: 15667858]
8. Perks AE, Schuler TD, Lee J, Ghiculete D, Chung D, D'A Honey RJ, et al. Stone attenuation and skin-to-stone distance on computed tomography predicts for stone fragmentation by shock wave lithotripsy. Urology 2008 Oct;72(4):765-769. [doi: 10.1016/j.urology.2008.05.046] [Medline: 18674803]
9. Hatiboglu G, Popeneciu V, Kurosch M, Huber J, Pahernik S, Pfitzenmaier J, et al. Prognostic variables for shockwave lithotripsy (SWL) treatment success: no impact of body mass index (BMI) using a third generation lithotripter. BJU Int 2011 Oct;108(7):1192-1197 [FREE Full text] [doi: 10.1111/j.1464-410X.2010.10007.x] [Medline: 21342413]
10. Wiesenthal JD, Ghiculete D, Ray AA, Honey RJD, Pace KT. A clinical nomogram to predict the successful shock wave lithotripsy of renal and ureteral calculi. J Urol 2011 Aug;186(2):556-562. [doi: 10.1016/j.juro.2011.03.109] [Medline: 21684557]
11. Patel T, Kozakowski K, Hruby G, Gupta M. Skin to stone distance is an independent predictor of stone-free status following shockwave lithotripsy. J Endourol 2009 Sep;23(9):1383-1385. [doi: 10.1089/end.2009.0394] [Medline: 19694526]
12. Dretler SP. Special article: calculus breakability--fragility and durility. J Endourol 1994 Feb;8(1):1-3. [doi: 10.1089/end.1994.8.1] [Medline: 8186775]

13. Ringdén I, Tiselius H. Composition and clinically determined hardness of urinary tract stones. Scand J Urol Nephrol 2007;41(4):316-323. [doi: 10.1080/00365590601154551] [Medline: 17763224]
14. Ouzaid I, Al-qahtani S, Dominique S, Hupertan V, Fernandez P, Hermieu J, et al. A 970 Hounsfield units (HU) threshold of kidney stone density on non-contrast computed tomography (NCCT) improves patients' selection for extracorporeal shockwave lithotripsy (ESWL): evidence from a prospective study. BJU Int 2012 Dec;110(11 Pt B):E438-E442. [doi: 10.1111/j.1464-410X.2012.10964.x] [Medline: 22372937]
15. El-Nahas AR, El-Assmy AM, Mansour O, Sheir KZ. A prospective multivariate analysis of factors predicting stone disintegration by extracorporeal shock wave lithotripsy: the value of high-resolution noncontrast computed tomography. Eur Urol 2007 Jun;51(6):1688-93; discussion 1693. [doi: 10.1016/j.eururo.2006.11.048] [Medline: 17161522]
16. Joseph P, Mandal A, Singh S, Mandal P, Sankhwar S, Sharma S. Computerized Tomography Attenuation Value of Renal Calculus: Can It Predict Successful Fragmentation of the Calculus by Extracorporeal Shock Wave Lithotripsy? A Preliminary Study. Journal of Urology 2002 May;167(5):1968-1971. [doi: 10.1016/s0022-5347(05)65064-1] [Medline: 11956419]
17. Abdelhamid M, Mosharafa AA, Ibrahim H, Selim HM, Hamed M, Elghoneimy MN, et al. A Prospective Evaluation of High-Resolution CT Parameters in Predicting Extracorporeal Shockwave Lithotripsy Success for Upper Urinary Tract Calculi. J Endourol 2016 Nov;30(11):1227-1232. [doi: 10.1089/end.2016.0364] [Medline: 27597174]
18. Yamashita S, Kohjimoto Y, Iguchi T, Nishizawa S, Iba A, Kikkawa K, et al. Variation Coefficient of Stone Density: A Novel Predictor of the Outcome of Extracorporeal Shockwave Lithotripsy. J Endourol 2017 Apr;31(4):384-390. [doi: 10.1089/end.2016.0719] [Medline: 28052698]
19. Bohris C, Stief CG, Strittmatter F. Improvement of SWL Efficacy: Reduction of the Respiration-Induced Kidney Motion by Using an Abdominal Compression Plate. J Endourol 2016 Apr;30(4):411-416. [doi: 10.1089/end.2015.0681] [Medline: 26558296]
20. Honey RJ, Healy M, Yeung M, Psihramis KE, Jewett MA. The Use of an Abdominal Compression Belt to Reduce Stone Movement During Extracorporeal Shock Wave Lithotripsy. Journal of Urology 1992 Sep;148(3 Part 2):1034-1035. [doi: 10.1016/s0022-5347(17)36808-8] [Medline: 1507324]
21. Pishchalnikov YA, Neucks JS, VonDerHaar RJ, Pishchalnikova IV, Williams JC, McAteer JA. Air pockets trapped during routine coupling in dry head lithotripsy can significantly decrease the delivery of shock wave energy. J Urol 2006 Dec;176(6 Pt 1):2706-2710 [FREE Full text] [doi: 10.1016/j.juro.2006.07.149] [Medline: 17085200]
22. Jain A, Shah TK. Effect of air bubbles in the coupling medium on efficacy of extracorporeal shock wave lithotripsy. Eur Urol 2007 Jun;51(6):1680-6; discussion 1686. [doi: 10.1016/j.eururo.2006.10.049] [Medline: 17112655]
23. Rabah DM, Mabrouki MS, Farhat KH, Seida MA, Arafa MA, Talic RF. Comparison of escalating, constant, and reduction energy output in ESWL for renal stones: multi-arm prospective randomized study. Urolithiasis 2017 Jun;45(3):311-316. [doi: 10.1007/s00240-016-0912-7] [Medline: 27687681]
24. Connors BA, Evan AP, Handa RK, Blomgren PM, Johnson CD, Liu Z, et al. Using 300 Pretreatment Shock Waves in a Voltage Ramping Protocol Can Significantly Reduce Tissue Injury During Extracorporeal Shock Wave Lithotripsy. J Endourol 2016 Sep;30(9):1004-1008 [FREE Full text] [doi: 10.1089/end.2016.0087] [Medline: 27307070]
25. Lambert EH, Walsh R, Moreno MW, Gupta M. Effect of escalating versus fixed voltage treatment on stone comminution and renal injury during extracorporeal shock wave lithotripsy: a prospective randomized trial. J Urol 2010 Feb;183(2):580-584. [doi: 10.1016/j.juro.2009.10.025] [Medline: 20018316]
26. Handa RK, McAteer JA, Connors BA, Liu Z, Lingeman JE, Evan AP. Optimising an escalating shockwave amplitude treatment strategy to protect the kidney from injury during shockwave lithotripsy. BJU Int 2012 Dec;110(11 Pt C):E1041-E1047 [FREE Full text] [doi: 10.1111/j.1464-410X.2012.11207.x] [Medline: 22612388]
27. McAteer JA, Evan AP, Williams JC, Lingeman JE. Treatment protocols to reduce renal injury during shock wave lithotripsy. Curr Opin Urol 2009 Mar;19(2):192-195 [FREE Full text] [doi: 10.1097/mou.0b013e32831e16e3] [Medline: 19195131]
28. Honey RJD, Ray AA, Ghiculete D, University of Toronto Lithotripsy Associates, Pace KT. Shock wave lithotripsy: a randomized, double-blind trial to compare immediate versus delayed voltage escalation. Urology 2010 Jan;75(1):38-43. [doi: 10.1016/j.urology.2008.12.070] [Medline: 19896176]
29. Pishchalnikov YA, McAteer JA, Williams JC, Pishchalnikova IV, Vonderhaar RJ. Why stones break better at slow shockwave rates than at fast rates: in vitro study with a research electrohydraulic lithotripter. J Endourol 2006 Aug;20(8):537-541 [FREE Full text] [doi: 10.1089/end.2006.20.537] [Medline: 16903810]
30. Pishchalnikov YA, McAteer JA, Williams JC. Effect of firing rate on the performance of shock wave lithotriptors. BJU Int 2008 Dec;102(11):1681-1686 [FREE Full text] [doi: 10.1111/j.1464-410X.2008.07896.x] [Medline: 18710450]
31. Kang DH, Cho KS, Ham WS, Lee H, Kwon JK, Choi YD, et al. Comparison of High, Intermediate, and Low Frequency Shock Wave Lithotripsy for Urinary Tract Stone Disease: Systematic Review and Network Meta-Analysis. PLoS One 2016;11(7):e0158661 [FREE Full text] [doi: 10.1371/journal.pone.0158661] [Medline: 27387279]
32. Delius M, Jordan M, Eizenhoefer H, Marlinghaus E, Heine G, Liebich HG, et al. Biological effects of shock waves: Kidney haemorrhage by shock waves in dogs—Administration rate dependence. Ultrasound in Medicine & Biology 1988 Jan;14(8):689-694. [doi: 10.1016/0301-5629(88)90025-7] [Medline: 3212839]

33. Willis LR, Evan AP, Connors BA, Shao Y, Blomgren PM, Pratt JH, et al. Shockwave lithotripsy: dose-related effects on renal structure, hemodynamics, and tubular function. J Endourol 2005;19(1):90-101. [doi: 10.1089/end.2005.19.90] [Medline: 15735392]
34. Mobley TB, Myers DA, Grine WB, Jenkins JM, Jordan WR. Low Energy Lithotripsy with the Lithostar: Treatment Results with 19,962 Renal and Ureteral Calculi. Journal of Urology 1993 Jun;149(6):1419-1424. [doi: 10.1016/s0022-5347(17)36404-2] [Medline: 8501779]
35. Skuginna V, Nguyen DP, Seiler R, Kiss B, Thalmann GN, Roth B. Does Stepwise Voltage Ramping Protect the Kidney from Injury During Extracorporeal Shockwave Lithotripsy? Results of a Prospective Randomized Trial. Eur Urol 2016 Feb;69(2):267-273. [doi: 10.1016/j.eururo.2015.06.017] [Medline: 26119561]
36. Fan J, Wang J, Chen Z, Hu C, Zhang Z, Hu W. Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique. Med Phys 2019 Jan;46(1):370-381. [doi: 10.1002/mp.13271] [Medline: 30383300]
37. Smith WP, Kim M, Holdsworth C, Liao J, Phillips MH. Personalized treatment planning with a model of radiation therapy outcomes for use in multiobjective optimization of IMRT plans for prostate cancer. Radiat Oncol 2016 Mar 11;11:38 [FREE Full text] [doi: 10.1186/s13014-016-0609-7] [Medline: 26968687]
38. Nicolae A, Morton G, Chung H, Loblaw A, Jain S, Mitchell D, et al. Evaluation of a Machine-Learning Algorithm for Treatment Planning in Prostate Low-Dose-Rate Brachytherapy. Int J Radiat Oncol Biol Phys 2017 Mar 15;97(4):822-829. [doi: 10.1016/j.ijrobp.2016.11.036] [Medline: 28244419]
39. Mak RH, Endres MG, Paik JH, Sergeev RA, Aerts H, Williams CL, et al. Use of Crowd Innovation to Develop an Artificial Intelligence-Based Solution for Radiation Therapy Targeting. JAMA Oncol 2019 May 01;5(5):654-661 [FREE Full text] [doi: 10.1001/jamaoncol.2019.0159] [Medline: 30998808]
40. Lee S, Celik S, Logsdon BA, Lundberg SM, Martins TJ, Oehler VG, et al. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. Nat Commun 2018 Jan 03;9(1):42 [FREE Full text] [doi: 10.1038/s41467-017-02465-5] [Medline: 29298978]
41. Lin H, Wei N, Chou T, Lin C, Lan Y, Chang S, et al. Building personalized treatment plans for early-stage colorectal cancer patients. Oncotarget 2017 Feb 21;8(8):13805-13817 [FREE Full text] [doi: 10.18632/oncotarget.14638] [Medline: 28099153]
42. Doubleday K, Zhou H, Fu H, Zhou J. An Algorithm for Generating Individualized Treatment Decision Trees and Random Forests. J Comput Graph Stat 2018;27(4):849-860 [FREE Full text] [doi: 10.1080/10618600.2018.1451337] [Medline: 32523325]
43. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
44. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Cornell University. 2014. URL: https://arxiv.org/abs/1406.1078 [accessed 2021-04-24]
45. 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]
46. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc 2017 Mar 01;24(2):361-370 [FREE Full text] [doi: 10.1093/jamia/ocw112] [Medline: 27521897]
47. Xu L, Choy CS, Li YW. Deep sparse rectifier neural networks for speech denoising. 2016 Presented at: IEEE International Workshop on Acoustic Signal Enhancement (IWAENC); September 13-16, 2016; Xi'an, China. [doi: 10.1109/iwaenc.2016.7602891]
48. Kingma DP, Ba J. Adam: A method for stochastic optimization. Cornell University. 2014. URL: https://arxiv.org/abs/1412.6980 [accessed 2021-04-24]
49. Reddi SJ, Kale S, Kumar S. On the convergence of Adam and beyond. Cornell University. 2019. URL: https://arxiv.org/abs/1904.09237 [accessed 2021-04-24]
50. Manning CD, Raghavan P, Schütze H. Natural language engineering. In: Introduction to information retrieval. Cambridge, MA: Cambridge University Press; 2010:100-103.
51. Narasimhan H, Pan W, Kar P, Protopapas P, Ramaswamy HG. Optimizing the multiclass F-measure via biconcave programming. 2017 Presented at: IEEE 16th International Conference on Data Mining (ICDM); December 12-15, 2016; Barcelona, Spain. [doi: 10.1109/icdm.2016.0143]
52. Rassweiler JJ, Knoll T, Köhrmann KU, McAteer JA, Lingeman JE, Cleveland RO, et al. Shock wave technology and application: an update. Eur Urol 2011 May;59(5):784-796 [FREE Full text] [doi: 10.1016/j.eururo.2011.02.033] [Medline: 21354696]
53. Türk C, Petřík A, Sarica K, Seitz C, Skolarikos A, Straub M, et al. EAU Guidelines on Interventional Treatment for Urolithiasis. Eur Urol 2016 Mar;69(3):475-482. [doi: 10.1016/j.eururo.2015.07.041] [Medline: 26344917]

Abbreviations

LSTM: long short-term memory

MAE: mean absolute error
MSE: mean squared error
PPC: preoperative patient characteristics
ReLU: rectifier linear unit
RFC: random forest classifier
RFR: random forest regression
RMSE: root mean squared error
RNN: recurrent neural network
SVC: support vector classifier
SVR: support vector regression
SWL: shock wave lithotripsy

Edited by G Eysenbach; submitted 02.10.20; peer-reviewed by JA Benítez-Andrades, G Lim; comments to author 06.11.20; revised version received 31.12.20; accepted 11.04.21; published 11.05.21.

Please cite as:
Chen Z, Zeng DD, Seltzer RGN, Hamilton BD
Automated Generation of Personalized Shock Wave Lithotripsy Protocols: Treatment Planning Using Deep Learning
JMIR Med Inform 2021;9(5):e24721
URL: https://medinform.jmir.org/2021/5/e24721
doi:10.2196/24721
PMID:33973862

Original Paper

Deciphering the Efficacy and Mechanisms of Chinese Herbal Medicine for Diabetic Kidney Disease by Integrating Web-Based Biochemical Databases and Real-World Clinical Data: Retrospective Cohort Study

Chien-Wei Wu[1*], MD; Hsing-Yu Chen[1,2,3*], MD; Ching-Wei Yang[1,2], MD; Yu-Chun Chen[4,5,6], MSc, MD

[1]Division of Chinese Internal and Pediatric Medicine, Center for Traditional Chinese Medicine, Chang Gung Memorial Hospital, Taoyuan, Taiwan
[2]School of Traditional Chinese Medicine, College of Medicine, Chang Gung University, Taoyuan, Taiwan
[3]Graduate Institute of Clinical Medical Sciences, College of Medicine, Chang Gung University, Taoyuan, Taiwan
[4]School of Medicine, Faculty of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
[5]Department of Family Medicine, Taipei Veterans General Hospital, Taipei, Taiwan
[6]Institute of Hospital and Health Care Administration, National Yang Ming Chiao Tung University, Taipei, Taiwan
[*]these authors contributed equally

Corresponding Author:
Yu-Chun Chen, MSc, MD
School of Medicine
Faculty of Medicine
National Yang Ming Chiao Tung University
No 155, Sec 2, Linong Street
Taipei, 112
Taiwan
Phone: 886 2 28712121 ext 7460
Fax: 886 2 28737901
Email: yuchn.chen@gmail.com

Abstract

Background: Diabetic kidney disease (DKD) is one of the most crucial causes of chronic kidney disease (CKD). However, the efficacy and biomedical mechanisms of Chinese herbal medicine (CHM) for DKD in clinical settings remain unclear.

Objective: This study aimed to analyze the outcomes of DKD patients with CHM-only management and the possible molecular pathways of CHM by integrating web-based biomedical databases and real-world clinical data.

Methods: A total of 152,357 patients with incident DKD from 2004 to 2012 were identified from the National Health Insurance Research Database (NHIRD) in Taiwan. The risk of mortality was estimated with the Kaplan-Meier method and Cox regression considering demographic covariates. The inverse probability of treatment weighting was used for confounding bias between CHM users and nonusers. Furthermore, to decipher the CHM used for DKD, we analyzed all CHM prescriptions using the Chinese Herbal Medicine Network (CMN), which combined association rule mining and social network analysis for all CHM prescriptions. Further, web-based biomedical databases, including STITCH, STRING, BindingDB, TCMSP, TCM@Taiwan, and DisGeNET, were integrated with the CMN and commonly used Western medicine (WM) to explore the differences in possible target proteins and molecular pathways between CHM and WM. An application programming interface was used to assess these online databases to obtain the latest biomedical information.

Results: About 13.7% (20,947/131,410) of patients were classified as CHM users among eligible DKD patients. The median follow-up duration of all patients was 2.49 years. The cumulative mortality rate in the CHM cohort was significantly lower than that in the WM cohort (28% vs 48%, P<.001). The risk of mortality was 0.41 in the CHM cohort with covariate adjustment (99% CI 0.38-0.43; P<.001). A total of 173,525 CHM prescriptions were used to construct the CMN with 11 CHM clusters. CHM covered more DKD-related proteins and pathways than WM; nevertheless, WM aimed at managing DKD more specifically. From the overrepresentation tests carried out by the online website Reactome, the molecular pathways covered by the CHM clusters in the CMN and WM seemed distinctive but complementary. Complementary effects were also found among DKD patients with concurrent WM and CHM use. The risk of mortality for CHM users under renin-angiotensin-aldosterone system (RAAS) inhibition therapy was lower than that for CHM nonusers among DKD patients with hypertension (adjusted hazard ratio [aHR] 0.47, 99% CI 0.45-0.51; *P*<.001), chronic heart failure (aHR 0.43, 99% CI 0.37-0.51; *P*<.001), and ischemic heart disease (aHR 0.46, 99% CI 0.41-0.51; *P*<.001).

Conclusions: CHM users among DKD patients seemed to have a lower risk of mortality, which may benefit from potentially synergistic renoprotection effects. The framework of integrating real-world clinical databases and web-based biomedical databases could help in exploring the roles of treatments for diseases.

(JMIR Med Inform 2021;9(5):e27614) doi:10.2196/27614

KEYWORDS

association rule mining; Chinese medicine network; social network analysis; survival

Introduction

Diabetic kidney disease (DKD) is one of the most crucial causes of chronic kidney disease (CKD) and end-stage renal disease (ESRD) at the final disease stage, especially when the prevalence of DKD keeps increasing yearly [1]. It has been reported that about one-third of DKD patients may experience ESRD during their lifetime [2]. Owing to the high prevalence and severe consequences, DKD has become a vital health care problem and causes tremendous financial burden [3-5]. The pathogenesis of diabetic nephropathy is complicated; however, the treatment modalities are still limited and need to be explored. Glomerular hyperfiltration, podocyte dysfunction, basement membrane thickening, mesangial cell proliferation, and collagen deposition with glomerular sclerosis are extensively reported [6-8]. Additionally, several precipitating factors have been identified, including hyperglycemia, advanced glycation end products, activation of the renin-angiotensin-aldosterone system (RAAS), decreased expression of nephrin and integrin, activation of cytokines, profibrotic elements, inflammation, oxidative stress, and vascular growth factors [9-12].

Although there are many Western medicine (WM) options for DKD, only blockade of the RAAS has been identified as an effective treatment, and the agents include angiotensin-converting enzyme inhibitors (ACEis), angiotensin receptor blockers (ARBs), and direct renin inhibitors (DRIs) [13-18]. Several notable novel agents have been recently reported to have benefits for reducing progression to DKD among diabetes mellitus (DM) patients, and these agents include sodium-glucose cotransporter 2 inhibitors (SGLT2is), glucagon-like peptide-1 (GLP-1) agonists, a selective endothelin-1 receptor antagonist, and a nonsteroidal mineralocorticoid receptor antagonist. However, the effectiveness of these agents among DM patients who are already diagnosed with DKD remains unclear, and some clinical trials are ongoing to address these issues [19-22]. Only GLP-1 agonists and SGLT2is have been found to be beneficial in DKD patients [23,24]. These novel agents inspire researchers to study other medications with similar effects on similar pathways and new therapeutic agents for CKD/DKD [16].

Complementary and alternative medicine may be another treatment option to relieve DKD in addition to WM. Several treatment modalities, including Chinese herbal medicine (CHM) and acupuncture, have been reported to have potential therapeutic benefits for DKD [25-28]. Moreover, some medications may be used to relieve proteinuria and ameliorate renal dysfunction, such as *Astragalus membranaceus* (Fisch.) and Liu-Wei-Di-Huang-Wan [29-31]. The potential mechanisms include anti-inflammation, antifibrosis, antioxidation, immunomodulation, and regulation of podocyte dysfunction [30-36]. Besides, some CHMs have been found to have effects on tubular cell cycle modulation [37]. However, only some of the abovementioned herbs/ingredients have been examined in terms of the clinical efficacy in treating DKD, and, on the other hand, only a small proportion of CHMs used in clinical trials have been examined in terms of the possible mechanisms in treating DKD owing to the high heterogeneity in used CHMs for DKD [31,38]. Additionally, the CHM prescriptions used for diseases are usually complicated in the clinical setting, and we previously found that the use of four to five kinds of CHMs in one prescription is not uncommon [39]. A comprehensive summary of the efficacy of CHM prescriptions becomes crucial to understand the effects of CHM for DKD [40,41].

Several methods have been proposed to extract valuable information from complicated CHM prescriptions, such as association rule mining, clustering, and decision tree [42]. In recent years, network pharmacology based on web-based biomedical resources has become one of the most critical tools to analyze CHM prescriptions [43-45]. However, the integration of these techniques with real-world clinical data has been lacking. For DKD, Zhang et al reported the potential effects of six representative compounds in the Gandi capsule (a mixture of several CHMs with fixed proportions) for 99 potential DKD-related target proteins [46]. Moreover, Shi et al tried to use the molecule-protein docking method to predict the possible mechanisms of Bushenhuoxue formula for treating CKD. They identified the potential of tanshinone IIA, rhein, curcumin, calycosin, and quercetin to act on CKD-related proteins, which may be related to the regulation of coagulation and fibrinolytic balance, aberrant extracellular matrix accumulation, and inflammation [47]. However, owing to the lack of clinical data, the effectiveness of these CHM formulae for DKD and the interactions between these CHMs and WMs remain unclear [48]. Besides, the interactions between CHMs and WMs are essential to understand the role of CHM in the modern health care system and the unexpected effects of CHM on DKD from the perspective of molecular medicine. For the above reasons, an integrative analysis on real-world data and web-based biomedical resources with the long-term effects of CHM and synergistic effects of CHM and WM is demanded and necessary for the management of DKD.

In our previous findings, DKD patients who received all kinds of Traditional Chinese medicine (TCM) treatments, including CHM, acupuncture, and moxibustion, had a better prognosis, which raised our interest in CHM use for DKD patients and the possible effective biomedical pathways [41]. In our previous successful integration of the most up-to-date web-based biomedical databases and real-world prescription databases, we identified the synergistic effects of CHM and WM for allergic rhinitis [40]. This study aimed to analyze the outcomes of DKD patients with CHM-only management and elucidate the roles of CHM and WM for DKD using an integrative platform with clinical and web-based biomedical databases.

Methods

Data Source and Study Design

The National Health Insurance Research Database (NHIRD), with high coverage (>99%) of all medical records in Taiwan, was used as a prescription data source for this study. The clinical data were preprocessed in our previous reports, including patient demographic features and prescriptions, and the protocol was approved by the Institutional Review Board of Chang Gung Memorial Foundation (number: 103-1259B) [41]. DKD patients were identified with a diagnosis of CKD after DM. From January 1, 2004, to December 31, 2012, DM patients were recognized by using International Classification of Diseases 9, Clinical Modification (ICD-9-CM) codes 250.0-250.9 and antidiabetic medications, including insulin and biguanides sulfonylurea, an alpha-glucosidase inhibitor, thiazolidinediones, and DDP-4 inhibitors. Furthermore, CKD was recognized by using ICD-9-CM codes 580.X-588.X, 250.4x, 274.1x, 283.11, 403.x1, 404.x2, 404.x3, 440.1, 442.1, 447.3, 572.4, 642.1x, and 646.2x. To recognize incident DKD patients, any subjects with previous CKD records or renal transplantation were excluded. Additionally, any visits with the use of acupuncture, massage, or other TCM modalities were excluded. In Taiwan, the diagnosis of DKD is consistent with the guidelines, and detection of diabetic nephropathy (DN) subjects by ICD-9-CM codes was consistent with previous studies [29,41]. CHM users were defined as DN patients who used CHM at least twice for DN from 2004 to 2012, and all CHM prescriptions were collected to build up the Chinese Herbal Medicine Network (CMN) with the integration of web-based biomedical databases.

Bias Assessment

This data set was unique and particularly suitable for CHM prescription analysis owing to its high coverage of Taiwan's general population and unbiased selection of CHM as a treatment option [39,49]. Possible selection bias and referral bias could be avoided as much as possible with a nationwide database than with a hospital-based database [50]. Additionally, the exclusion of acupuncture, moxibustion, or manual therapy is helpful to avoid confounding bias with possible influence on CHM prescriptions. Moreover, because there is no recommendation for initiation of CHM treatments, we found that the mean interval from diagnosis of DKD to initiation of CHM use was about 240 days among CHM users (data not shown), and immortal time bias may occur [51,52]. To overcome this problem, a 1-year landmark design was used to avoid the potential immortal time bias. Thus, the study index date was set as 1 year after DKD diagnosis for each patient, and patients who died within 1 year after DKD diagnosis were excluded as well. Moreover, to eliminate the possible baseline differences between CHM users and nonusers, inverse probability treatment weighting (IPTW) according to all assessable covariates described below was used [53].

Study Covariates and Outcome

Patient gender, age, comorbidities, medications, prior experience of CHM use, geolocation, and insured level were used as covariates in this study. The Charlson comorbidity index (CCI) and Diabetes Complications Severity Index (DCSI), with reduction of two factors (albuminuria and serum creatine) as a modification, were calculated as a summary of DKD-related comorbidities [54,55]. The identification of specific comorbidities was based on ICD-9-CM codes for diseases, including cerebrovascular disease (ICD-9-CM codes 430-432 and 433-435), heart failure (ICD-9-CM code 428), ischemic heart disease (IHD; ICD-9-CM codes 411, 413, and 414), hypertension (ICD‐9‐CM codes 401‐405), and hyperlipidemia (ICD‐9‐CM code 272). Only patients with at least two diagnosis codes in the outpatient service or one during the hospitalization 1 year before the DKD diagnosis were confirmed as having comorbidities. We also analyzed medications, including insulin; other antihyperglycemic agents; antihypertensive agents; antilipid agents; RAAS blockers, including ACEis, ARBs, and DRIs; aspirin; and nonsteroidal anti-inflammatory drugs (NSAIDs). Only medications with a cumulative duration of more than 30 days were included in the analysis. All-cause mortality was the outcome of this study, and it was recognized when patients permanently withdrew from the insurance program [56,57]. All enrolled DKD patients were followed up from the DKD starting point to the endpoint or the end of 2012.

CHM Prescriptions in the Database

Traditionally, the medicines used by TCM doctors include not only herbal plants, but also insects, animals, and minerals. In this study, we collectively referred to all medicines recorded in the database as CHM. There are two kinds of CHMs used in clinical practice, namely herbal formula (HF) and single herb (SH). SH is the extract or crude powder of a part of a herbal plant, insect, animal, or mineral and is made following ancient classics' process methods. On the other hand, HF is composed of more than one SH with the same proportion as recorded in TCM classics and is premixed in the pharmaceutical factory before marketing. More than 600 kinds of SHs and HFs are available for TCM doctors to choose freely, and all SHs and HFs are manufactured in a Good Manufacturing Practice pharmaceutical factory with strict regulation regarding the concentrations of heavy metals and pesticides.

Statistical Analysis: Outcome Evaluation and Online Pathway Analysis on the CMN

The first part of the statistical analysis was survival analysis. Descriptive statistics were used for CHM users' demographic characteristics, such as age, gender, comorbidities, insured level, living locations, previous medical use, and prescribing patterns.

After applying IPTW to balance the differences between CHM users and nonusers, survival analysis was carried out by performing a Kaplan-Meier estimation with the log-rank test. Additionally, Cox regression with adjustment of the abovementioned assessable covariates was used to estimate the adjusted hazard ratio (aHR) for CHM users. Furthermore, to ensure the analysis results, subgroup analysis was conducted based on age, gender, and comorbidities. Sensitivity tests were also performed with 1:1 matching on CHM users and nonusers, and different study populations.

Second, CHM prescription analysis with integration of the CMN and online biomedical databases was performed to reveal the potential molecular pathways of CHM for DKD. In this step, the application programming interface (API) was used to assess the biomedical databases to obtain the latest information about WM and CHM. As described in our previous studies about CHM prescription analysis, the CMN was constructed by applying association rule mining (ARM) and social network analysis (SNA) on CHM prescriptions for DKD [39,58,59]. Briefly, ARM on CHM prescriptions used for DKD could find out the CHM-CHM combinations commonly used for DKD. These CHM combinations could be connected to form the CMN for DKD, and SNA with these combinations could put CHMs used concurrently into the same cluster. CHM indications acquired from the Chinese Pharmacopoeia (2015 edition) were used to summarize the CHM clusters from TCM viewpoints [39]. On the other hand, four types of WMs used for DKD were proposed in this study, including ACEis, ARBs, GLP-1 agonists, and SGLT2is. Other possible molecular pathways could be obtained based on these CHM clusters and WMs by connecting WMs and CHM clusters to online biomedical databases [40]. Since biomedical databases contained only information about SHs, every HF in the CMN was disassembled to SHs according to the compositions provided by the Department of Chinese Medicine and Pharmacy of the Ministry of Health and Welfare, Taiwan [60]. Next, the ingredients of each SH were obtained from TCMSP [61], TCM-ID [62], and TCM@Taiwan [63], and the information was cross-validated with the Chinese Pharmacopoeia (2015 edition). Each ingredient's characteristics were also acquired from PubChem, such as oral bioavailability, XlogP, drug likeness, molecular weight, topological polar surface area (TPSA), and simplified molecular input line entry specification (SMILES). This information was crucial to realize the similarities between the ingredients of WMs and CHMs [48].

Moreover, to acquire each ingredient's possible target proteins for both WM and CHM, the Search Tool for Interacting Chemicals (STITCH) [64,65] was queried. STITCH is a well-developed database composed of known and predicted connections between a chemical compound and target proteins derived from genomic context predictions, high-throughput lab experiments, gene coexpression databases, text mining in journal databases, and previous knowledge from other databases [66]. Up to January 2021, STITCH contained 9,643,763 proteins, 2,031 organisms, and over 430,000 chemical compounds as ingredients in CHMs. Inferred chemical-target protein connections from experiments involving species other than humans and a scoring system to describe the confidence of the connections in this database could be used to explore the connections between chemical compounds and target proteins. The scoring system, ranging from 0 to 1, summarizes the probability of connection occurrence by combining the probability from individual data sources, such as experiments from mice and text mining from journal databases, in a native Bayesian fashion. A higher score symbolizes more substantial confidence in the connection between chemical compounds and target proteins. To select the most confident connections between drug ingredients and target proteins, a threshold of 0.950 was considered.

Furthermore, to assess the molecular pathways for CHMs in the CMN and WMs, the target proteins were sent to the Reactome pathway database via API, where overrepresentation tests were performed to disclose the potential acting pathways of CHMs and WMs [67-69]. Reactome is a freely accessible web resource to estimate, interpret, and visualize the molecular pathways of given groups of genes or proteins. A total of 15 species pathways were included in the Reactome pathway database, and there were 10,929 proteins, 13,534 reactions, and 2477 pathways for humans (last assessed date: January 2020). The overrepresentation analysis was carried out on the hypothesis that if a molecular pathway is relevant, the pathway's proteins should be more than randomly expected. The false discovery rate (FDR), calculated using the Benjamini-Hochberg approach, was used to demonstrate each pathway's statistical significance. Pathways with an FDR ≤0.05 were considered.

Figures 1 and 2 demonstrate the data processing flow of this study. The freeware KNIME (version 4.0) was used to deal with the clinical and web-based biomedical databases. NodeXL was used to build up the networks and perform SNA [70]. The commercial statistical software STATA (Release 16, StataCorp) was used to carry out survival analysis on core CHMs. A *P* value ≤.05 was considered significant.

Figure 1. Flow diagram. DKD: diabetic kidney disease.

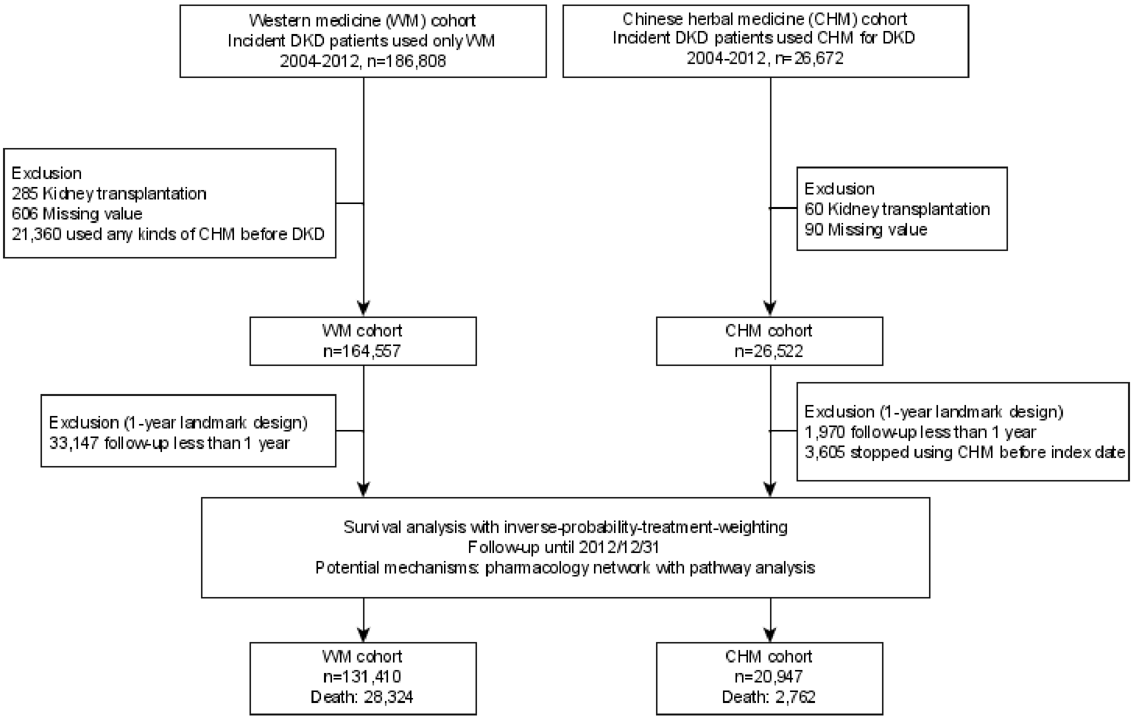
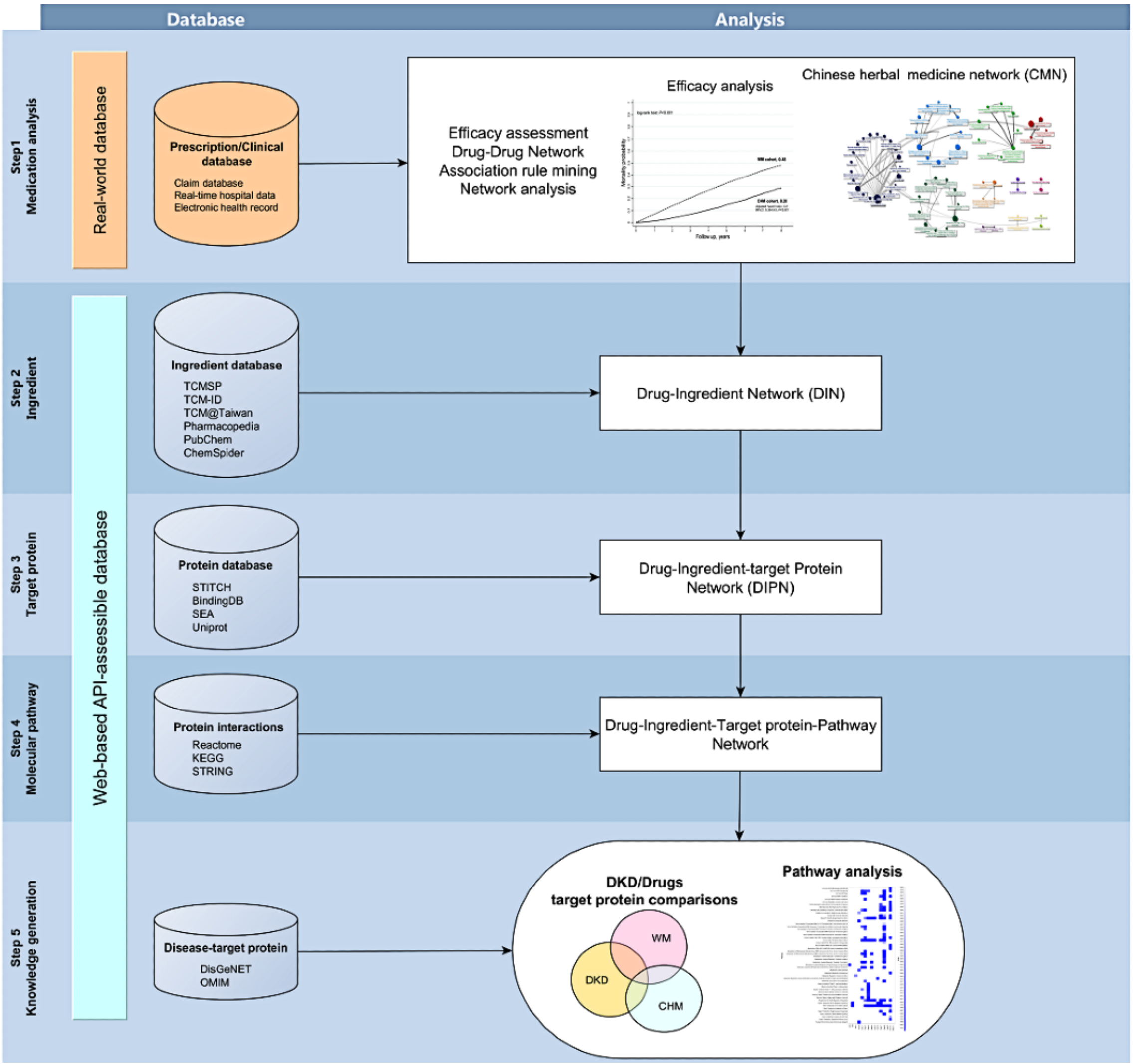

Western medicine (WM) cohort
Incident DKD patients used only WM
2004-2012, n=186,808
Chinese herbal medicine (CHM) cohort
Incident DKD patients used CHM for DKD
2004-2012, n=26,672
Exclusion
285 Kidney transplantation
606 Missing value
21,360 used any kinds of CHM before DKD
Exclusion
60 Kidney transplantation
90 Missing value
WM cohort
n=164,557
CHM cohort
n=26,522
Exclusion (1-year landmark design)
33,147 follow-up less than 1 year
Exclusion (1-year landmark design)
1,970 follow-up less than 1 year
3,605 stopped using CHM before index date
Survival analysis with inverse-probability-treatment-weighting
Follow-up until 2012/12/31
Potential mechanisms: pharmacology network with pathway analysis
WM cohort
n=131,410
Death: 28,324
CHM cohort
n=20,947
Death: 2,762

Figure 2. Data processing framework. API: application programming interface; DKD: diabetic kidney disease.



Results

Baseline Characteristics of CHM Users Among DKD Patients

Table 1 shows the demographic features of the CHM and WM cohorts. Except for the use of cyclooxygenase-2 inhibitors, these two cohorts were quite different. DKD patients who were female, aged 41 to 60 years, lived in urban areas, and had higher income were more commonly seen among CHM users. Regarding underlying diseases, less DKD-related comorbidities, such as hypertension, hyperlipidemia, heart failure, IHD, and cerebrovascular disease, and DM-related complications were found among CHM users. Regarding medications, except NSAIDs and acetaminophen, drugs to control hypertension, IHD, heart failure, hyperlipidemia, and DM were more commonly seen in the WM cohort. On applying the IPTW method, the baseline demographic features of the CHM and WM cohorts were well balanced (all standardized mean differences were within 10% among the CHM and WM cohorts; Multimedia Appendix 1).

Table 1. Characteristics of the Chinese herbal medicine and Western medicine cohorts among incident diabetic kidney disease patients from 2004 to 2012.

Characteristic	CHM[a] cohort (n=20,947), n (%) or mean (SD)	WM[b] cohort (n=131,410), n (%) or mean (SD)	*P* value
Gender			<.001
Female	9298 (44.4%)	55,546 (42.3%)	
Male	11,649 (55.6%)	75,864 (57.7%)	
Age (years)			<.001
≤40	991 (4.7%)	4437 (3.4%)	
41-60	9388 (44.8%)	41,605 (31.7%)	
≥61	10,568 (50.5%)	85,368 (65.0%)	
Comorbidities			
Hypertension	12,489 (59.6%)	89,219 (67.9%)	<.001
Hyperlipidemia	7748 (37.0%)	52,805 (40.2%)	<.001
Heart failure	1023 (4.9%)	10,053 (7.7%)	<.001
IHD[c]	3674 (17.5%)	24,820 (18.9%)	<.001
CVD[d]	1379 (6.6%)	12,765 (9.7%)	<.001
Hyperuricemia	1999 (9.5%)	14,477 (11.0%)	<.001
Modified DCSI[e] score, mean (SD)	3.3 (1.2)	3.4 (1.2)	<.001
Medications			
Diabetic drugs			
Insulin analogs	2674 (12.8%)	23,197 (17.7%)	<.001
OHAs[f]	14,487 (69.2%)	97,077 (73.9%)	<.001
Lipid-lowering agents			
Statin/fibrate	8139 (38.9%)	57,001 (43.4%)	<.001
Antihypertensives			
ACEi[g]/ARB[h]	10,820 (51.7%)	79,763 (60.7%)	<.001
Others	15,443 (73.7%)	99,808 (76.0%)	<.001
Analgesics/aspirin			
NSAIDs[i]	6783 (32.4%)	33,540 (25.5%)	<.001
COX-2[j] inhibitors	1576 (7.5%)	9931 (7.6%)	0.86
Acetaminophen	5767 (27.5%)	28,702 (21.8%)	<.001
Aspirin	6356 (30.3%)	43,116 (32.8%)	<.001
Insured level (NTD[k,l]/month)			<.001
0-20,000	15,802 (75.0%)	105,987 (81.0%)	
20,001-40,000	3441 (16.0%)	17,366 (13.0%)	
≥40,001	1704 (8.0%)	8057 (6.0%)	
Geolocation			<.001
Urban	15,437 (73.7%)	92,967 (70.7%)	
Rural	5510 (26.3%)	38,443 (29.3%)	
Previous TCM[m] users	11,161 (53.3%)	0 (0.0%)	<.001

[a]CHM: Chinese herbal medicine.

[b]WM: Western medicine.

[c]IHD: ischemic heart disease.

[d] CVD: cerebral vascular disease.

[e] DCSI: Diabetes Complications Severity Index.

[f] OHA: oral hypoglycemic agents.

[g] ACEi: angiotensin-converting enzyme inhibitor.

[h] ARB: angiotensin receptor blocker.

[i] NSAID: nonsteroidal anti-inflammatory drug.

[j] COX-2: cyclooxygenase-2.

[k] NTD: new Taiwan dollar.

[l] 1 NTD=0.033 USD.

[m] TCM: traditional Chinese medicine.

Risk of Mortality Among CHM Users

At the end of 2012, the median follow-up duration of all patients was 2.49 years, and the cumulative mortality among the CHM cohort was significantly lower than the WM cohort (28% vs 48%, *P*<.001; Figure 3). On adjusting age, gender, socioeconomic status, comorbidities, DM-associated complications, and medications, the risk of mortality was 0.41 among the CHM cohort (99% CI 0.38-0.43; *P*<.001). Furthermore, it seemed that prolonged use of CHM for DKD is safe. The risk of mortality reduced as the duration of CHM increased; DKD patients with CHM use less than 180 days had twice the risk of mortality than patients with CHM use more than 180 days (aHR 0.51 vs 0.25; both *P*<.001 compared to the WM cohort; Table 2). The sensitivity tests including propensity scores with 1:1 matching and the CHM cohort without late users demonstrated similar results (Multimedia Appendix 2). Moreover, when considering the influence of ESRD on DKD, patients in the CHM cohort had lower risk of mortality on either excluding ESRD patients (aHR 0.38, 99% CI 0.36-0.41; *P*<.001) or including ESRD patients (aHR 0.45, 99% CI 0.41-0.50; *P*<.001) (Multimedia Appendix 2). Moreover, the subgroup analysis on mortality risks using multivariate Cox regression showed reduced risks among CHM users stratified by age, gender, and comorbidities.

Figure 3. Overall survival benefit among patients using Chinese herbal medicine (CHM) for diabetic kidney disease. Kaplan-Meier curves by patient groups. WM: Western medicine.

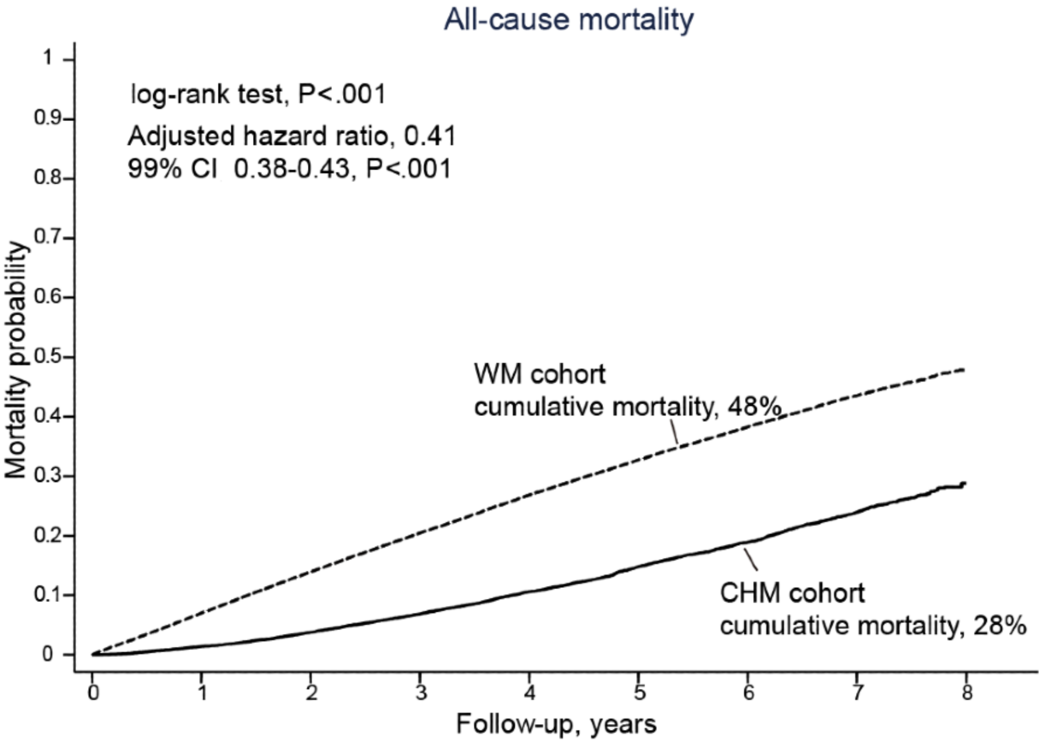


Table 2. Association of Chinese herbal medicine for diabetic kidney disease with lower mortality probability.

Variable	Subjects, n	Deaths, n	aHR[a]	99% CI	*P* value
WM cohort	131,410	28,324	1 (reference)	N/A[b]	N/A
CHM duration					
<180 days	13,670	2093	0.51	0.48-0.54	<.001
≥180 days	7277	669	0.25	0.22-0.27	<.001

[a] aHR: adjusted hazard ratio.

[b] N/A: not applicable.

CMN for DKD

A total of 173,525 CHM prescriptions were analyzed during the study period. There were 661 kinds of CHMs used (69.5% of all kinds of CHMs available in Taiwan), and about 5.7 CHMs were used in each prescription on average. Ji-Sheng-Shen-Qi-Wan was used most commonly (22.9%) (Multimedia Appendix 3). Figure 4 demonstrates the CMN for DKD, which was constructed by summarizing the CHM-CHM combinations selected by the ARM from all CHM prescriptions (the top 10 combinations listed in Multimedia Appendix 4). By using SNA to assemble the CHMs commonly used together, a total of 11 clusters could be defined, and the CHMs contained in each cluster are listed in Multimedia Appendix 5. A higher resolution of the CMN graph is provided online as well [71]. The within-cluster CHMs had closer relations than CHMs between clusters, which meant the CHMs in the same cluster were more commonly coprescribed. The network also revealed that other intracluster CHMs frequently connected some CHMs among clusters composed of more than two CHMs, such as Ji-Sheng-Shen-Qi-Wan in cluster 1, *Astragalus membranaceus* (Fisch.) Bge. in cluster 2, *Salvia miltiorrhiza* Bge. in cluster 3, *Epimedium sagittatum* (Sieb. et Zucc.) Maxim. in cluster 4, *Dipsacus asperoides* C. Y. Cheng at T. M. Ai in cluster 5, and *Aconitum carmichaelii* in cluster 6. In their clusters, other CHMs seemed to have to be used with these CHMs as adjuvants. Moreover, some between-cluster relations could be found as well (Figure 4), such as cluster 1-cluster 2, cluster 1-cluster 5, cluster 2-cluster 3, cluster 1-cluster 3, and cluster 1-cluster 11, which adequately represent the complexity of CHM prescriptions in the clinical setting. Taking these prescription patterns together, when dealing with DKD, TCM doctors may use CHM combinations in the same cluster and sometimes more than one cluster. The potential effects of each cluster were assessed, and the trends of reduced risks of mortality were similar among each cluster compared to CHM nonusers (Table 3).

Figure 4. Chinese Herbal Medicine Network (CMN) for diabetic kidney disease. Relations are indicated by grey lines connected to the center of clusters.

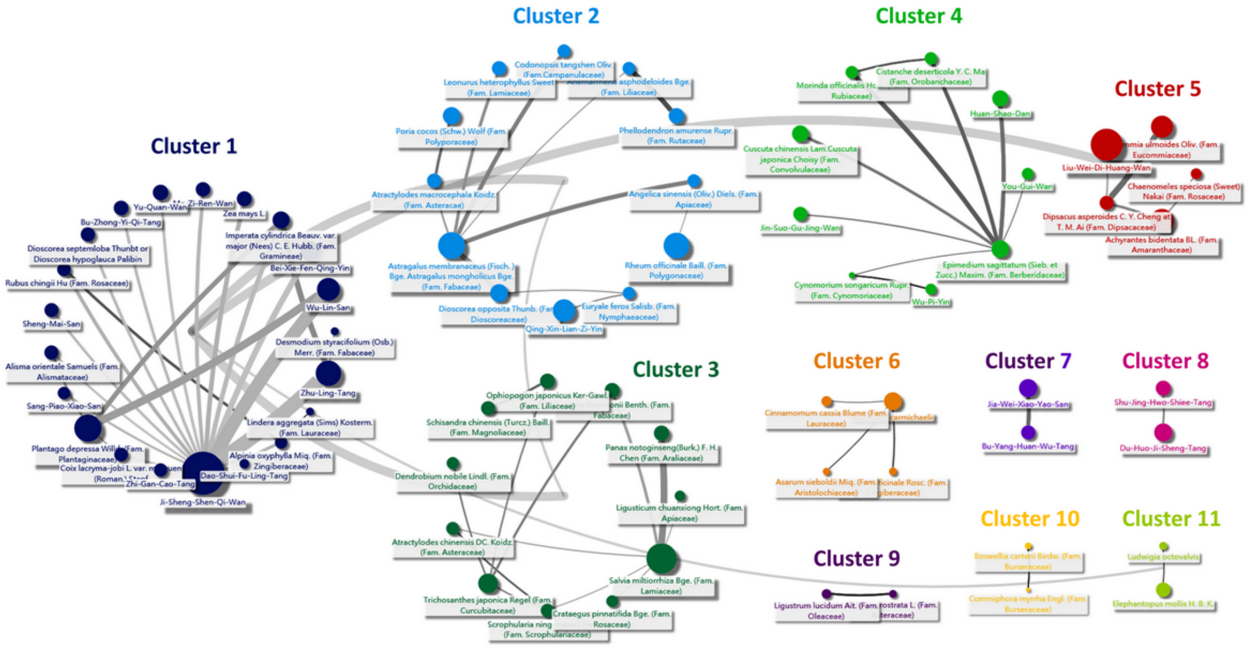

Cluster 1
Yu-Quan-Wan
Zi-Ren-Wan
Zea mays L.
Bu-Zhong-Yi-Qi-Tang
Imperata cylindrica Beauv. var. major (Nees) C. E. Hubb. (Fam. Gramineae)
Dioscorea septemloba Thunbt or Dioscorea hypoglauca Palibin
Bei-Xie-Fen-Qing-Yin
Rubus chingii Hu (Fam. Rosaceae)
Wu-Lin-San
Sheng-Mai-San
Desmodium styracifolium (Osb.) Merr. (Fam. Fabaceae)
Alisma orientale Samuels (Fam. Alismataceae)
Zhu-Ling-Tang
Sang-Piao-Xiao-San
Lindera aggregata (Sims) Kosterm. (Fam. Lauraceae)
Plantago depressa Willd. (Fam. Plantaginaceae)
Alpinia oxyphylla Miq. (Fam. Zingiberaceae)
Coix lacryma-jobi L. var. ma-yuen (Roman.) Stapf
Dao-Shui-Fu-Ling-Tang
Zhi-Gan-Cao-Tang
Ji-Sheng-Shen-Qi-Wan
Cluster 2
Codonopsis tangshen Oliv. (Fam.Campanulaceae)
Leonurus heterophyllus Sweet (Fam. Lamiaceae)
Anemarrhena asphodeloides Bge. (Fam. Liliaceae)
Poria cocos (Schw.) Wolf (Fam. Polyporaceae)
Phellodendron amurense Rupr. (Fam. Rutaceae)
Atractylodes macrocephala Koidz. (Fam. Asteracae)
Angelica sinensis (Oliv.) Diels. (Fam. Apiaceae)
Astragalus membranaceus (Fisch.) Bge. Astragalus mongholicus Bge. (Fam. Fabaceae)
Rheum officinale Baill. (Fam. Polygonaceae)
Dioscorea opposita Thunb. (Fam. Dioscoreaceae)
Euryale ferox Salisb. (Fam. Nymphaeaceae)
Qing-Xin-Lian-Zi-Yin
Cluster 3
Ophiopogon japonicus Ker-Gawl. (Fam. Liliaceae)
Sophora tonkinensis / ...onii Benth. (Fam. Fabaceae)
Schisandra chinensis (Turcz.) Baill. (Fam. Magnoliaceae)
Panax notoginseng(Burk.) F. H. Chen (Fam. Araliaceae)
Dendrobium nobile Lindl. (Fam. Orchidaceae)
Ligusticum chuanxiong Hort. (Fam. Apiaceae)
Atractylodes chinensis DC. Koidz. (Fam. Asteraceae)
Salvia miltiorrhiza Bge. (Fam. Lamiaceae)
Trichosanthes japonica Regel (Fam. Curcubitaceae)
Crataegus pinnatifida Bge. (Fam. Rosaceae)
Scrophularia ningpoensis Hemsl. (Fam. Scrophulariaceae)
Cluster 4
Cistanche deserticola Y. C. Ma (Fam. Orobanchaceae)
Morinda officinalis How (Fam. Rubiaceae)
Huan-Shao-Dan
Cuscuta chinensis Lam.Cuscuta japonica Choisy (Fam. Convolvulaceae)
You-Gui-Wan
Jin-Suo-Gu-Jing-Wan
Epimedium sagittatum (Sieb. et Zucc.) Maxim. (Fam. Berberidaceae)
Cynomorium songaricum Rupr. (Fam. Cynomoriaceae)
Wu-Pi-Yin
Cluster 5
Eucommia ulmoides Oliv. (Fam. Eucommiaceae)
Liu-Wei-Di-Huang-Wan
Chaenomeles speciosa (Sweet) Nakai (Fam. Rosaceae)
Dipsacus asperoides C. Y. Cheng at T. M. Ai (Fam. Dipsacaceae)
Achyrantes bidentata BL. (Fam. Amaranthaceae)
Cluster 6
Cinnamomum cassia Blume (Fam. Lauraceae)
Aconitum carmichaelii
Asarum sieboldii Miq. (Fam. Aristolochiaceae)
Zingiber officinale Rosc. (Fam. Zingiberaceae)
Cluster 7
Jia-Wei-Xiao-Yao-San
Bu-Yang-Huan-Wu-Tang
Cluster 8
Shu-Jing-Hwo-Shiee-Tang
Du-Huo-Ji-Sheng-Tang
Cluster 9
Ligustrum lucidum Ait. (Fam. Oleaceae)
Eclipta prostrata L. (Fam. Asteraceae)
Cluster 10
Boswellia carterii Birdw. (Fam. Burseraceae)
Commiphora myrrha Engl. (Fam. Burseraceae)
Cluster 11
Ludwigia octovalvis
Elephantopus mollis H. B. K.

Table 3. Cox regressions for mortality and 1-year landmark analysis among Chinese herbal medicine and Western medicine cohorts.

Cluster[a]	Unadjusted			Adjusted[b]		
	HR[c,d]	99% CI	*P* value	aHR[e,f]	99% CI	*P* value
Cluster 1 (n=5272)	0.39	0.35-0.43	<.001	0.40	0.36-0.44	<.001
Cluster 2 (n=2275)	0.37	0.32-0.43	<.001	0.37	0.32-0.44	<.001
Cluster 3 (n=2139)	0.38	0.33-0.45	<.001	0.39	0.34-0.46	<.001
Cluster 4 (n=905)	0.23	0.15-0.36	<.001	0.24	0.16-0.35	<.001
Cluster 5 (n=1144)	0.33	0.26-0.43	<.001	0.34	0.27-0.44	<.001
Cluster 6 (n=665)	0.38	0.28-0.50	<.001	0.38	0.29-0.51	<.001
Cluster 7 (n=173)	0.43	0.25-0.74	<.001	0.48	0.30-0.76	<.001
Cluster 8 (n=375)	0.22	0.10-0.49	<.001	0.23	0.11-0.47	<.001
Cluster 9 (n=243)	0.32	0.19-0.54	<.001	0.32	0.20-0.50	<.001
Cluster 10 (n=286)	0.33	0.20-0.53	<.001	0.32	0.21-0.48	<.001
Cluster 11 (n=196)	0.45	0.27-0.73	<.001	0.45	0.29-0.69	<.001

[a] Each cluster contained patients who took different groups of Chinese herbal medicines.

[b] Gender, age, geolocation, insured level, comorbidities, and medications were used as covariates in the adjusted regression models.

[c] HR: hazard ratio.

[d] The hazard ratio of each cluster was estimated after inverse probability treatment weighting in contrast to the Western medicine cohort.

[e] aHR: adjusted hazard ratio.

[f] The adjusted hazard ratio was calculated by a Cox regression model considering patient gender, age, comorbidities, medications, insured level, and geolocation. Inverse probability treatment weighting was estimated from the same covariates to relieve the accessible confounding bias between Chinese herbal medicine users and nonusers.

Web-Based Molecular Pathway Exploration Regarding CHM Clusters and WMs

Figure 5 shows the associations between DKD-related proteins and CHM clusters or WMs on searching potential target proteins for clinically commonly used CHMs and WMs in a web-based database as mentioned above. The examples of connections between CHMs, CHM ingredients, and target proteins are listed in Multimedia Appendix 6. There were 767 ingredients contained in CHMs in the CMN and 37 WMs in four types of WMs. The physiochemical characteristics of CHMs and WMs were quite different (Multimedia Appendix 7 and Multimedia Appendix 8). Figure 5 reveals CHM clusters often covering more DKD-related proteins than WMs commonly used for DKD; however, we also found that CHM clusters often covered much more DKD-unrelated target proteins than WMs (Figure 6).

Figure 5. The proportion of DKD-related proteins covered by WM and CHM. A higher proportion of DKD-related proteins is covered by CHM clusters. ACEi: angiotensin-converting enzyme inhibitor; ARB: angiotensin receptor blocker; CHM: Chinese herbal medicine; DKD: diabetic kidney disease; GLP-1: glucagon-like peptide-1; SGLT2i: sodium-glucose cotransporter 2 inhibitors; WM: Western medicine.

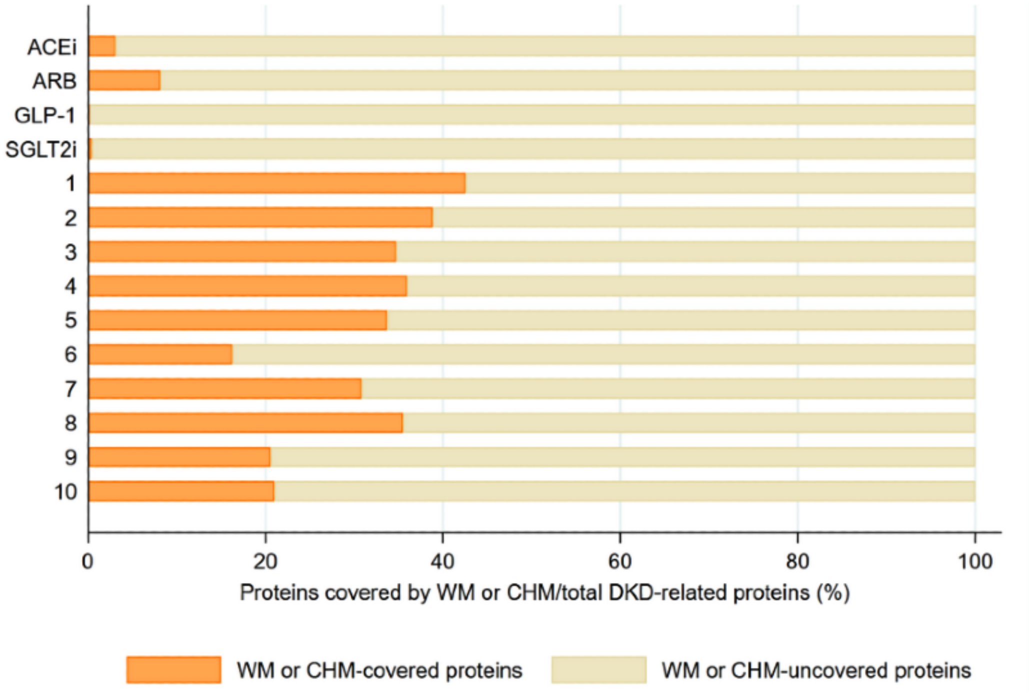
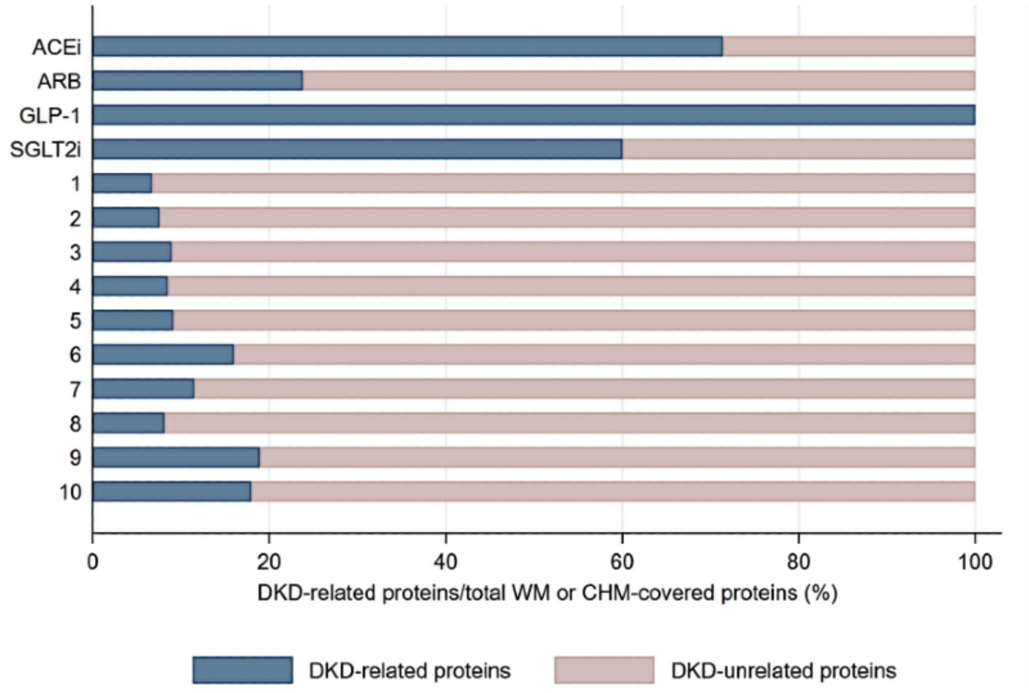


Figure 6. The proportion of proteins covered by CHM or WM specific to DKD. WM aimed more specifically at DKD-related proteins. ACEi: angiotensin-converting enzyme inhibitor; ARB: angiotensin receptor blocker; CHM: Chinese herbal medicine; DKD: diabetic kidney disease; GLP-1: glucagon-like peptide-1; SGLT2i: sodium-glucose cotransporter 2 inhibitors; WM: Western medicine.



Moreover, it was notable that Figure 7 shows the diverse molecular pathways covered by CHM clusters and WMs. CHM clusters potentially covered more pathways than WMs. The pathways in CHM clusters, which include GPCR ligand blinding and GPCR downstream signaling, overlapped with ARB pathways. On the contrary, pathways of ACEis, GLP-1 agonists, and SGLT2is had no intersection with CHM clusters. Moreover, many CHM cluster pathways were not covered by WMs, such as cell cycle, gene regulation, and metabolism pathways. Table 4 shows the possibly complementary effects of WMs and CHMs, since their molecular pathways seemed rather distinctive. DKD patients with hypertension, HF, and IHD who used RAAS blockers and CHMs had lower risks of mortality than those who used RAAS blockers alone (aHR 0.47, 99% CI 0.45-0.51; *P*<.001; aHR 0.43, 99% CI 0.37-0.51; *P*<.001; and aHR 0.46, 99% CI 0.41-0.51; *P*<.001, respectively).

Figure 7. Summary of biomedical pathways covered by clusters of Chinese herbal medicine (CHM) and Western medicine (WM). Pathway overrepresentation analysis and classification was performed by assessing the online Reactome database, and only pathways with a false discovery rate ≤0.05 were considered.

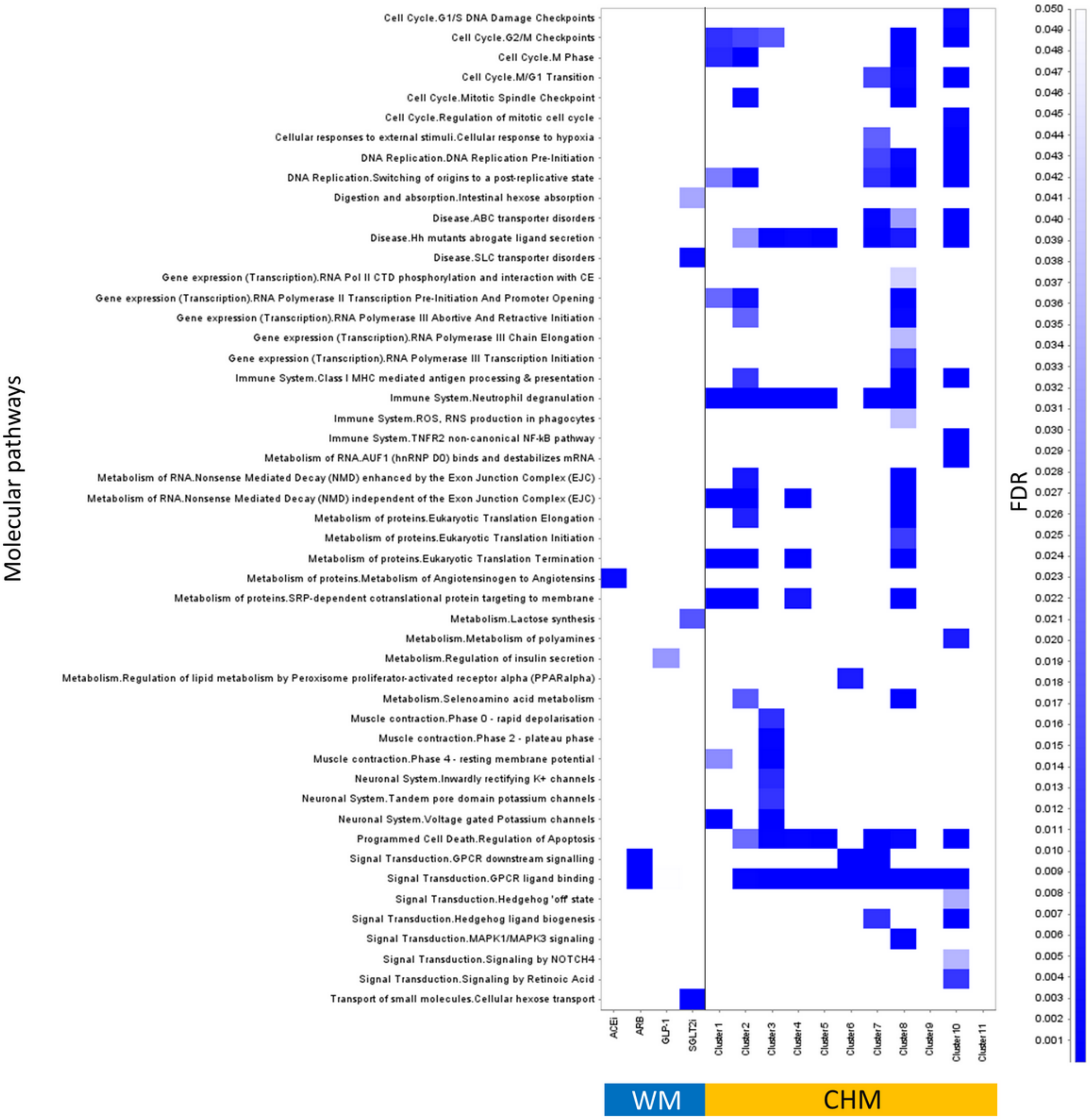


Table 4. Risks of mortality among Chinese herbal medicine users with chronic heart failure, hypertension, and ischemic heart disease under renin-angiotensin-aldosterone system inhibition therapy.

Disease in patients who received renin-angiotensin-aldosterone system inhibition blockers	CHM[a] nonusers		CHM users		aHR[b,c]	99% CI	*P* value
	Events, n	Subjects, n	Events, n	Subjects, n			
Hypertension	18,310	78,935	1813	10,814	0.47	0.45-0.51	<.001
Chronic heart failure	3876	9369	269	977	0.43	0.37-0.51	<.001
Ischemic heart disease	6408	21,359	600	2967	0.46	0.41-0.51	<.001
Hypertension, chronic heart failure, or ischemic heart disease	19,225	82,110	1919	11,240	0.48	0.45-0.51	<.001

[a] CHM: Chinese herbal medicine.

[b] aHR: adjusted hazard ratio.

[c] The aHR was calculated by a Cox regression model considering patient gender, age, comorbidities, medications, insured level, and geolocation. Inverse probability treatment weighting was estimated from the same covariates to relieve the accessible confounding bias between Chinese herbal medicine users and nonusers.

Discussion

Principal Findings

This is the first study to analyze CHM prescriptions for incident DKD patients. Potential survival benefits and molecular pathways present the potential complementary roles of CHMs in managing DKD. We previously proposed a framework to connect clinical databases to web-based biochemical and pathway databases to predict the efficacy of using CHM, and we reported the possible acting pathways of CHM for DKD with the same framework [40]. The different pathways among CHM and WM may have synergistic effects for DKD. By integrating web-based biomedical databases with the CMN, we found several clusters after analyzing the common use of CHM in the NHIRD, which reflected the TCM viewpoints and prescription patterns for DKD. The lower mortality risks among CHM clusters for DKD revealed the potential usefulness of CHM among DKD patients.

Most importantly, by using the web-based Reactome pathway database, the pathways of CHM could be comprehensively overviewed. There are many pathways covered by CHM clusters, but which are not seen in WM. Furthermore, the complementary effects could be validated by clinical data. The framework of cross-utilization of clinical and web-based biochemical databases showed the possibility to decipher CHM treatments for diseases.

Antihypertensive medicines, such as ACEis and ARBs, are recommended in patients with DKD. They are proven to reduce mortality rates and prevent cardiovascular morbidity. In addition, they can slow the degeneration of kidney function in patients with hyperalbuminuria and pre-ESRD [72-74]. Some studies have declared that simultaneous use of more than one drug is a good strategy for treating DKD patients because of different mechanisms [75-77]. However, whether CHM should be used with ACEis or ARBs remained unexplored, even though use of CHM with ACEis or ARBs simultaneously may improve blood pressure control among hypertension patients [78]. Our study reported the rationale of combining an ACEi or ARB with CHM for DKD patients by presenting the long-term benefits and the different coverage of DKD-related proteins and molecular pathways. Notably, CHM clusters used for DKD often cover more DKD-related proteins and pathways than WMs. Most pathways overrepresented by CHM are different from those related to ACEis or ARBs, such as cell cycle and gene regulation.

Cell cycle and gene regulation seemed to be the most different covered pathways between CHM and WM. A previous study found that specific CHMs are involved in DKD-related modulation of microRNA [79]. Besides, the importance of cell cycle arrest in treating CKD seems to be increasing in recent years [80-84]. Cell division involves the following four phases: G0-G1, S, G2, and M. To repair the injured tissue ultimately, DNA is replicated and divided in the process of the cell cycle. Checkpoints play an important role in the quality assurance process during cell division [85]. It is reported that proximal tubular cells arrested in the G2/M phase after an injury are responsible for the fibrotic response, which leads to CKD [83,84,86]. Hence, helping cells abrogate the G2/M arrest and preventing profibrotic growth factor release are new strategies for avoiding renal fibrosis [81,87]. According to our results, several pathways related to the cell cycle may be covered by CHM clusters, especially in pathways related to G2/M checkpoints. With probable regulation in the cell cycle, CHM may target specific sites and participate in cellular repair. Owing to multiple targets in CHM, the simultaneous use of CHM and WM provides a complementary treatment and another perspective in patients with DKD under ACEi or ARB therapy.

In addition to the pathways proposed by integrating biomedical databases with clinical databases, the CMN also revealed TCM viewpoints on managing DKD. TCM doctors categorize diseases into different specific patterns according to the patients' symptoms and clinical conditions. This characteristic approach to personalized diagnosis and treatment is named "bian-zheng-lun-zhi." In our study, 11 clusters according to the frequency and coprescription of CHMs were classified. These clusters with specific features explain that "bian-zheng-lun-zhi" has meticulous assessment and better outcomes. Ji-Sheng-Shen-Qi-Wan, which was used most commonly in our study, is the prescription to treat a patient with DKD diagnosed as having "Kidney-Yang Deficiency" in TCM. In TCM theory, “Kidney-Yang” indicates that the functions of the reproductive, endocrine, and urinary systems are normal, and “Kidney-Yang Deficiency” indicates hypofunctioning of these systems [88]. Several molecular pathways, such as cell cycle, gene regulation, signal transduction, immune system, and metabolism, were found to be involved in “Kidney-Yang” [89,90]. These pathways are similar to the pathways in which Ji-Sheng-Shen-Qi-Wan in cluster 1 is overrepresented. Therefore, the associations of DKD-related mechanisms and specific patterns of TCM are revealed by this framework.

Safety is one of the outcomes that cannot be ignored. Specific CHMs have been banned because they were proven to cause damage to the kidney, such as aristolochic acid–containing herbs [49,91]. We used the cohort from 2004 to exclude any potential adverse effects of aristolochic acid–containing herbs, and it helped us evaluate patient outcomes when using CHMs. Furthermore, we proved that using CHMs might be safe and beneficial for DKD patients in the long term. Our study indicates that more prolonged use of CHM among DKD patients reduces the mortality rate, especially in DKD patients who use CHM for more than 180 days.

Limitations

There are some limitations in this study. First, the severity and stage of DKD could not be analyzed owing to the lack of clinical information, such as glycemia, blood pressure, complete blood count, and biochemistry. The actual quality of control in DM and hypertension is crucial to patients with DKD. While this study aimed to compare mortality between WM and CHM, and explore the use of CHM, quite a few patients used CHM for nonfatal conditions that involved serious effects on quality of life, such as diabetic ophthalmopathy and limb necrosis. Future studies should focus on comparisons in such nonfatal conditions. Second, data about self-paid CHM and folk medicine were absent because only reimbursed CHM treatments were included. In Taiwan, most CHM treatments are fully reimbursed and convenient. Therefore, our study results would not be greatly affected by the use of self-paid CHM and folk medicine. Third, new oral hypoglycemic agents, such as SGLT2is, were not included in the analysis. These agents have the advantage of lowering mortality rates and cardiovascular morbidity among DKD patients [77]. Although they were not included in this study owing to approval in Taiwan in 2014, more studies about combined therapy involving SGLT2is are important in the future, since the pathways covered by SGLT2is were found to be quite different in our study.

Conclusion

By integrating clinical and biomedical databases, lower mortality rates among CHM users were found, and the complementary roles of CHM and WM may be the reason. Since CHM has complementary effects and proven safety, it may be beneficial to consider TCM treatment in DKD patients under WM therapy. Our study’s main advantage is the clarification of the summary of the mechanisms of CHM for DKD in the real world, which may broaden the horizon for DKD and facilitate the development of new drugs from active ingredients in CHM. Further studies, including those involving more detailed information about patients’ conditions and analysis of prescribed CHMs, are required.

Acknowledgments

Grant support was received from the Chang Gung Memorial Foundation (grant numbers: CMRPG1J0021, CMRPG1J0022, and CMRPG1J0023) and the Ministry of Science and Technology in Taiwan (grant numbers: 104-2320-B-182A-013 and 108-2320-B-182A-009-MY3).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Standardized mean differences (%) after inverse probability treatment weighting between Chinese herbal medicine clusters and the Western medicine cohort.
[DOCX File , 35 KB - medinform_v9i5e27614_app1.docx]

Multimedia Appendix 2
Sensitivity analyses on the risks of mortality among Chinese herbal medicine users.
[DOCX File , 27 KB - medinform_v9i5e27614_app2.docx]

Multimedia Appendix 3
The top 10 single Chinese herbal medicines for diabetic kidney disease among 173,525 prescriptions.
[DOCX File , 28 KB - medinform_v9i5e27614_app3.docx]

Multimedia Appendix 4
The top 10 two Chinese herbal medicine combinations for diabetic kidney disease.
[DOCX File , 28 KB - medinform_v9i5e27614_app4.docx]

Multimedia Appendix 5
Chinese herbal medicines in each cluster.
[DOCX File , 29 KB - medinform_v9i5e27614_app5.docx]

Multimedia Appendix 6
Examples of the Chinese herbal medicine-ingredient-target protein network.
[DOCX File , 28 KB - medinform_v9i5e27614_app6.docx]

Multimedia Appendix 7
Differences in the physiochemical characteristics of Chinese herbal medicine (767 ingredients) and Western medicine (37 ingredients) used for diabetic nephropathy.
[DOCX File , 28 KB - medinform_v9i5e27614_app7.docx]

Multimedia Appendix 8
Physiochemical characteristics of 767 ingredients contained in the Chinese Herbal Medicine Network (CMN) and 37 compounds of Western medicine for diabetic kidney disease. The comparisons between density functions were performed by the two-sample Kolmogorov-Smirnov test.
[PNG File , 208 KB - medinform_v9i5e27614_app8.png]

References

1. Yang WC, Hwang SJ, Chiang SS, Chen HF, Tsai ST. The impact of diabetes on economic costs in dialysis patients: experiences in Taiwan. Diabetes Research and Clinical Practice 2001 Nov;54:47-54. [doi: 10.1016/s0168-8227(01)00309-6] [Medline: 11580969]
2. Ritz E, Orth SR. Nephropathy in patients with type 2 diabetes mellitus. N Engl J Med 1999 Oct 07;341(15):1127-1133. [doi: 10.1056/NEJM199910073411506] [Medline: 10511612]
3. Kuo H, Tsai S, Tiao M, Yang C. Epidemiological features of CKD in Taiwan. Am J Kidney Dis 2007 Jan;49(1):46-55. [doi: 10.1053/j.ajkd.2006.10.007] [Medline: 17185145]
4. Tsai S, Tseng H, Tan H, Chien Y, Chang C. End-stage renal disease in Taiwan: a case-control study. J Epidemiol 2009;19(4):169-176 [FREE Full text] [doi: 10.2188/jea.je20080099] [Medline: 19542686]
5. Yang W, Hwang S, Taiwan Society of Nephrology. Incidence, prevalence and mortality trends of dialysis end-stage renal disease in Taiwan from 1990 to 2001: the impact of national health insurance. Nephrol Dial Transplant 2008 Dec;23(12):3977-3982. [doi: 10.1093/ndt/gfn406] [Medline: 18628366]
6. Harris RD, Steffes MW, Bilous RW, Sutherland DE, Mauer SM. Global glomerular sclerosis and glomerular arteriolar hyalinosis in insulin dependent diabetes. Kidney Int 1991 Jul;40(1):107-114 [FREE Full text] [doi: 10.1038/ki.1991.187] [Medline: 1921145]
7. Fioretto P, Mauer M, Brocco E, Velussi M, Frigato F, Muollo B, et al. Patterns of renal injury in NIDDM patients with microalbuminuria. Diabetologia 1996 Dec;39(12):1569-1576. [doi: 10.1007/s001250050616] [Medline: 8960844]
8. Fliser D, Wagner K, Loos A, Tsikas D, Haller H. Chronic angiotensin II receptor blockade reduces (intra)renal vascular resistance in patients with type 2 diabetes. J Am Soc Nephrol 2005 Apr;16(4):1135-1140 [FREE Full text] [doi: 10.1681/ASN.2004100852] [Medline: 15716329]
9. Wolf G, Ziyadeh FN. Molecular mechanisms of diabetic renal hypertrophy. Kidney Int 1999 Aug;56(2):393-405 [FREE Full text] [doi: 10.1046/j.1523-1755.1999.00590.x] [Medline: 10432377]
10. Hilgers KF, Veelken R. Type 2 diabetic nephropathy: never too early to treat? J Am Soc Nephrol 2005 Mar;16(3):574-575 [FREE Full text] [doi: 10.1681/ASN.2005010083] [Medline: 15703269]
11. Hohenstein B, Hausknecht B, Boehmer K, Riess R, Brekken RA, Hugo CPM. Local VEGF activity but not VEGF expression is tightly regulated during diabetic nephropathy in man. Kidney Int 2006 May;69(9):1654-1661 [FREE Full text] [doi: 10.1038/sj.ki.5000294] [Medline: 16541023]
12. Navarro-González JF, Mora-Fernández C. The role of inflammatory cytokines in diabetic nephropathy. J Am Soc Nephrol 2008 Mar;19(3):433-442 [FREE Full text] [doi: 10.1681/ASN.2007091048] [Medline: 18256353]
13. Lewis EJ, Hunsicker LG, Clarke WR, Berl T, Pohl MA, Lewis JB, Collaborative Study Group. Renoprotective effect of the angiotensin-receptor antagonist irbesartan in patients with nephropathy due to type 2 diabetes. N Engl J Med 2001 Sep 20;345(12):851-860. [doi: 10.1056/NEJMoa011303] [Medline: 11565517]
14. Parving H, Persson F, Lewis JB, Lewis EJ, Hollenberg NK, AVOID Study Investigators. Aliskiren combined with losartan in type 2 diabetes and nephropathy. N Engl J Med 2008 Jun 05;358(23):2433-2446. [doi: 10.1056/NEJMoa0708379] [Medline: 18525041]
15. Wang H, Deng JL, Yue J, Li J, Hou YB. Prostaglandin E1 for preventing the progression of diabetic kidney disease. Cochrane Database Syst Rev 2010 May 12(5):CD006872. [doi: 10.1002/14651858.CD006872.pub2] [Medline: 20464745]

16. Haller H, Ito S, Izzo JL, Januszewicz A, Katayama S, Menne J, ROADMAP Trial Investigators. Olmesartan for the delay or prevention of microalbuminuria in type 2 diabetes. N Engl J Med 2011 Mar 10;364(10):907-917. [doi: 10.1056/NEJMoa1007994] [Medline: 21388309]
17. Shan D, Wu HM, Yuan QY, Li J, Zhou RL, Liu GJ. Pentoxifylline for diabetic kidney disease. Cochrane Database Syst Rev 2012 Feb 15(2):CD006800. [doi: 10.1002/14651858.CD006800.pub2] [Medline: 22336824]
18. St Peter WL, Odum LE, Whaley-Connell AT. To RAS or not to RAS? The evidence for and cautions with renin-angiotensin system inhibition in patients with diabetic kidney disease. Pharmacotherapy 2013 May;33(5):496-514. [doi: 10.1002/phar.1232] [Medline: 23576066]
19. Alicic RZ, Johnson EJ, Tuttle KR. SGLT2 Inhibition for the Prevention and Treatment of Diabetic Kidney Disease: A Review. Am J Kidney Dis 2018 Aug;72(2):267-277. [doi: 10.1053/j.ajkd.2018.03.022] [Medline: 29866460]
20. Wiviott SD, Raz I, Bonaca MP, Mosenzon O, Kato ET, Cahn A, DECLARE–TIMI 58 Investigators. Dapagliflozin and Cardiovascular Outcomes in Type 2 Diabetes. N Engl J Med 2019 Jan 24;380(4):347-357. [doi: 10.1056/NEJMoa1812389] [Medline: 30415602]
21. Pecoits-Filho R, Perkovic V. Are SGLT2 Inhibitors Ready for Prime Time for CKD? Clin J Am Soc Nephrol 2018 Feb 07;13(2):318-320 [FREE Full text] [doi: 10.2215/CJN.07680717] [Medline: 28893920]
22. Cherney DZI, Bakris GL. Novel therapies for diabetic kidney disease. Kidney Int Suppl (2011) 2018 Jan;8(1):18-25 [FREE Full text] [doi: 10.1016/j.kisu.2017.10.005] [Medline: 30675435]
23. Perkovic V, Jardine MJ, Neal B, Bompoint S, Heerspink HJL, Charytan DM, CREDENCE Trial Investigators. Canagliflozin and Renal Outcomes in Type 2 Diabetes and Nephropathy. N Engl J Med 2019 Jun 13;380(24):2295-2306. [doi: 10.1056/NEJMoa1811744] [Medline: 30990260]
24. Wang M. GLP1 fragments protect the kidney. Nat Rev Nephrol 2018 Oct;14(10):599. [doi: 10.1038/s41581-018-0056-9] [Medline: 30166605]
25. Burrowes JD, Van Houten G. Use of alternative medicine by patients with stage 5 chronic kidney disease. Adv Chronic Kidney Dis 2005 Jul;12(3):312-325. [doi: 10.1016/j.ackd.2005.04.001] [Medline: 16010646]
26. Garcia GE, Ma S, Feng L. Acupuncture and kidney disease. Adv Chronic Kidney Dis 2005 Jul;12(3):282-291. [doi: 10.1016/j.ackd.2005.04.002] [Medline: 16010643]
27. Li X, Wang H. Chinese herbal medicine in the treatment of chronic kidney disease. Adv Chronic Kidney Dis 2005 Jul;12(3):276-281. [doi: 10.1016/j.ackd.2005.03.007] [Medline: 16010642]
28. Markell MS. Potential benefits of complementary medicine modalities in patients with chronic kidney disease. Adv Chronic Kidney Dis 2005 Jul;12(3):292-299. [doi: 10.1016/j.ackd.2005.03.004] [Medline: 16010644]
29. Hsu P, Tsai Y, Lai J, Wu C, Lin S, Huang C. Integrating traditional Chinese medicine healthcare into diabetes care by reducing the risk of developing kidney failure among type 2 diabetic patients: a population-based case control study. J Ethnopharmacol 2014 Oct 28;156:358-364. [doi: 10.1016/j.jep.2014.08.029] [Medline: 25178949]
30. Zhang J, Xie X, Li C, Fu P. Systematic review of the renal protective effect of Astragalus membranaceus (root) on diabetic nephropathy in animal models. J Ethnopharmacol 2009 Nov 12;126(2):189-196. [doi: 10.1016/j.jep.2009.08.046] [Medline: 19735713]
31. Li M, Wang W, Xue J, Gu Y, Lin S. Meta-analysis of the clinical value of Astragalus membranaceus in diabetic nephropathy. J Ethnopharmacol 2011 Jan 27;133(2):412-419. [doi: 10.1016/j.jep.2010.10.012] [Medline: 20951192]
32. Liu H, Tang X, Dai D, Dai Y. Ethanol extracts of Rehmannia complex (Di Huang) containing no Corni fructus improve early diabetic nephropathy by combining suppression on the ET-ROS axis with modulate hypoglycemic effect in rats. J Ethnopharmacol 2008 Aug 13;118(3):466-472. [doi: 10.1016/j.jep.2008.05.015] [Medline: 18585879]
33. Poon TYC, Ong KL, Cheung BMY. Review of the effects of the traditional Chinese medicine Rehmannia Six Formula on diabetes mellitus and its complications. J Diabetes 2011 Sep;3(3):184-200. [doi: 10.1111/j.1753-0407.2011.00130.x] [Medline: 21631896]
34. Liu W, Liu P, Tao S, Deng Y, Li X, Lan T, et al. Berberine inhibits aldose reductase and oxidative stress in rat mesangial cells cultured under high glucose. Arch Biochem Biophys 2008 Jul 15;475(2):128-134. [doi: 10.1016/j.abb.2008.04.022] [Medline: 18471986]
35. Lan T, Liu W, Xie X, Huang K, Peng J, Huang J, et al. Berberine suppresses high glucose-induced TGF-β1 and fibronectin synthesis in mesangial cells through inhibition of sphingosine kinase 1/AP-1 pathway. Eur J Pharmacol 2012 Dec 15;697(1-3):165-172. [doi: 10.1016/j.ejphar.2012.10.003] [Medline: 23085271]
36. Zhao L, Sun L, Nie H, Wang X, Guan G. Berberine Improves Kidney Function in Diabetic Mice via AMPK Activation. PLoS ONE 2014 Nov 19;9(11):e113398 [FREE Full text] [doi: 10.1371/journal.pone.0113398] [Medline: 25409232]
37. Bailon-Moscoso N, Cevallos-Solorzano G, Romero-Benavides JC, Orellana MIR. Natural Compounds as Modulators of Cell Cycle Arrest: Application for Anticancer Chemotherapies. Curr Genomics 2017 Apr;18(2):106-131 [FREE Full text] [doi: 10.2174/1389202917666160808125645] [Medline: 28367072]
38. Zhang HW, Lin ZX, Xu C, Leung C, Chan LS. Astragalus (a traditional Chinese medicine) for treating chronic kidney disease. Cochrane Database Syst Rev 2014 Oct 22(10):CD008369. [doi: 10.1002/14651858.CD008369.pub2] [Medline: 25335553]

39. Chen H, Lin Y, Huang J, Chen Y. Chinese herbal medicine network and core treatments for allergic skin diseases: Implications from a nationwide database. J Ethnopharmacol 2015 Jun 20;168:260-267. [doi: 10.1016/j.jep.2015.04.002] [Medline: 25865681]
40. Lu Y, Yang C, Lin Y, Hsueh J, Chen J, Yang S, et al. Identifying the Chinese Herbal Medicine Network and Core Formula for Allergic Rhinitis on a Real-World Database. Evid Based Complement Alternat Med 2020;2020:5979708 [FREE Full text] [doi: 10.1155/2020/5979708] [Medline: 33204289]
41. Chen H, Pan H, Chen Y, Chen Y, Lin Y, Yang S, et al. Traditional Chinese medicine use is associated with lower end-stage renal disease and mortality rates among patients with diabetic nephropathy: a population-based cohort study. BMC Complement Altern Med 2019 Apr 03;19(1):81 [FREE Full text] [doi: 10.1186/s12906-019-2491-y] [Medline: 30943956]
42. Hendrickson WA, Ward KB. Atomic models for the polypeptide backbones of myohemerythrin and hemerythrin. Biochemical and Biophysical Research Communications 1975 Oct 27;66(4):1349-1356. [doi: 10.1016/0006-291x(75)90508-2]
43. AAAS. The Art and Science of Traditional Medicine Part 2: Multidisciplinary Approaches for Studying Traditional Medicine. Science 2015 Jan 15;347(6219):337-337. [doi: 10.1126/science.347.6219.337-c]
44. Li S, Zhang B. Traditional Chinese medicine network pharmacology: theory, methodology and application. Chinese Journal of Natural Medicines 2013 Mar;11(2):110-120. [doi: 10.1016/S1875-5364(13)60037-0] [Medline: 23787177]
45. Zhang R, Zhu X, Bai H, Ning K. Network Pharmacology Databases for Traditional Chinese Medicine: Review and Assessment. Front Pharmacol 2019;10:123 [FREE Full text] [doi: 10.3389/fphar.2019.00123] [Medline: 30846939]
46. Zhang J, Zhang Q, Chen X, Liu Y, Xue J, Dahan A, et al. Revealing Synergistic Mechanism of Multiple Components in Gandi Capsule for Diabetic Nephropathy Therapeutics by Network Pharmacology. Evid Based Complement Alternat Med 2018;2018:6503126 [FREE Full text] [doi: 10.1155/2018/6503126] [Medline: 29853965]
47. Shi S, Cai Y, Cai X, Zheng X, Cao D, Ye F, et al. A network pharmacology approach to understanding the mechanisms of action of traditional medicine: Bushenhuoxue formula for treatment of chronic kidney disease. PLoS One 2014;9(3):e89123 [FREE Full text] [doi: 10.1371/journal.pone.0089123] [Medline: 24598793]
48. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 2008 Nov;4(11):682-690. [doi: 10.1038/nchembio.118] [Medline: 18936753]
49. Hsieh CF, Huang SL, Chen CL, Chen WT, Chang HC, Yang CC. Non-aristolochic acid prescribed Chinese herbal medicines and the risk of mortality in patients with chronic kidney disease: results from a population-based follow-up study. BMJ Open 2014 Feb 21;4(2):e004033 [FREE Full text] [doi: 10.1136/bmjopen-2013-004033] [Medline: 24561496]
50. Pan H, Li C, Chen T, Su T, Wang K. Association of polypharmacy with fall-related fractures in older Taiwanese people: age- and gender-specific analyses. BMJ Open 2014 Mar 28;4(3):e004428 [FREE Full text] [doi: 10.1136/bmjopen-2013-004428] [Medline: 24682575]
51. Chien H, Kao Yang Y, Bai JPF. Trastuzumab-Related Cardiotoxic Effects in Taiwanese Women: A Nationwide Cohort Study. JAMA Oncol 2016 Oct 01;2(10):1317-1325. [doi: 10.1001/jamaoncol.2016.1269] [Medline: 27310478]
52. Suissa S. Immortal time bias in observational studies of drug effects. Pharmacoepidemiol Drug Saf 2007 Mar;16(3):241-249. [doi: 10.1002/pds.1357] [Medline: 17252614]
53. Mansournia MA, Altman DG. Inverse probability weighting. BMJ 2016 Jan 15;352:i189. [doi: 10.1136/bmj.i189] [Medline: 26773001]
54. 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]
55. Young BA, Lin E, Von Korff M, Simon G, Ciechanowski P, Ludman EJ, et al. Diabetes complications severity index and risk of mortality, hospitalization, and healthcare utilization. Am J Manag Care 2008 Jan;14(1):15-23 [FREE Full text] [Medline: 18197741]
56. Wu C, Chen Y, Ho HJ, Hsu Y, Kuo KN, Wu M, et al. Association between nucleoside analogues and risk of hepatitis B virus–related hepatocellular carcinoma recurrence following liver resection. JAMA 2012 Nov 14;308(18):1906-1914. [doi: 10.1001/2012.jama.11975] [Medline: 23162861]
57. Lien H, Chou S, Liu J. Hospital ownership and performance: evidence from stroke and cardiac treatment in Taiwan. J Health Econ 2008 Sep;27(5):1208-1223. [doi: 10.1016/j.jhealeco.2008.03.002] [Medline: 18486978]
58. Chen H, Lin Y, Thien P, Chang S, Chen Y, Lo S, et al. Identifying core herbal treatments for children with asthma: implication from a chinese herbal medicine database in taiwan. Evid Based Complement Alternat Med 2013;2013:125943 [FREE Full text] [doi: 10.1155/2013/125943] [Medline: 24066007]
59. Lin Y, Chen Y, Hu S, Chen H, Chen J, Yang S. Identifying core herbal treatments for urticaria using Taiwan's nationwide prescription database. J Ethnopharmacol 2013 Jul 09;148(2):556-562. [doi: 10.1016/j.jep.2013.04.052] [Medline: 23684721]
60. Department of Chinese Medicine and Pharmacy of the Ministry of Health and Welfare, Taiwan. URL: https://dep.mohw.gov.tw/DOCMAP/lp-874-108.html [accessed 2021-04-27]
61. Ru J, Li P, Wang J, Zhou W, Li B, Huang C, et al. TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. J Cheminform 2014;6:13 [FREE Full text] [doi: 10.1186/1758-2946-6-13] [Medline: 24735618]
62. Huang L, Xie D, Yu Y, Liu H, Shi Y, Shi T, et al. TCMID 2.0: a comprehensive resource for TCM. Nucleic Acids Res 2018 Jan 04;46(D1):D1117-D1120 [FREE Full text] [doi: 10.1093/nar/gkx1028] [Medline: 29106634]

63. Chen CY. TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico. PLoS One 2011 Jan 06;6(1):e15939 [FREE Full text] [doi: 10.1371/journal.pone.0015939] [Medline: 21253603]
64. Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P, Kuhn M. STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic Acids Res 2016 Jan 04;44(D1):D380-D384 [FREE Full text] [doi: 10.1093/nar/gkv1277] [Medline: 26590256]
65. STITCH. URL: http://stitch.embl.de/ [accessed 2021-04-27]
66. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 2015 Jan;43(Database issue):D447-D452 [FREE Full text] [doi: 10.1093/nar/gku1003] [Medline: 25352553]
67. Fabregat A, Sidiropoulos K, Viteri G, Forner O, Marin-Garcia P, Arnau V, et al. Reactome pathway analysis: a high-performance in-memory approach. BMC Bioinformatics 2017 Mar 02;18(1):142 [FREE Full text] [doi: 10.1186/s12859-017-1559-2] [Medline: 28249561]
68. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Res 2020 Jan 08;48(D1):D498-D503 [FREE Full text] [doi: 10.1093/nar/gkz1031] [Medline: 31691815]
69. Reactome. URL: https://reactome.org/ [accessed 2021-04-27]
70. NodeXL. CodePlex. URL: https://archive.codeplex.com/?p=nodexl [accessed 2021-04-27]
71. Diabetic Nephropathy. GraphSpace. URL: https://graphspace.org/graphs/31998 [accessed 2021-04-27]
72. Emdin CA, Rahimi K, Neal B, Callender T, Perkovic V, Patel A. Blood pressure lowering in type 2 diabetes: a systematic review and meta-analysis. JAMA 2015 Feb 10;313(6):603-615. [doi: 10.1001/jama.2014.18574] [Medline: 25668264]
73. Singhania N, Bansal S, Mohandas S, Nimmatoori DP, Ejaz AA, Singhania G. Role of renin-angiotensin-aldosterone system inhibitors in heart failure and chronic kidney disease. Drugs Context 2020;9 [FREE Full text] [doi: 10.7573/dic.2020-7-3] [Medline: 33240389]
74. Brenner BM, Cooper ME, de Zeeuw D, Keane WF, Mitch WE, Parving HH, RENAAL Study Investigators. Effects of losartan on renal and cardiovascular outcomes in patients with type 2 diabetes and nephropathy. N Engl J Med 2001 Sep 20;345(12):861-869. [doi: 10.1056/NEJMoa011161] [Medline: 11565518]
75. Williams B, MacDonald TM, Morant S, Webb DJ, Sever P, McInnes G, et al. Spironolactone versus placebo, bisoprolol, and doxazosin to determine the optimal treatment for drug-resistant hypertension (PATHWAY-2): a randomised, double-blind, crossover trial. The Lancet 2015 Nov 21;386(10008):2059-2068 [FREE Full text] [doi: 10.1016/S0140-6736(15)00257-3] [Medline: 26414968]
76. Bakris GL, Agarwal R, Chan JC, Cooper ME, Gansevoort RT, Haller H, Mineralocorticoid Receptor Antagonist Tolerability Study–Diabetic Nephropathy (ARTS-DN) Study Group. Effect of Finerenone on Albuminuria in Patients With Diabetic Nephropathy: A Randomized Clinical Trial. JAMA 2015 Sep 01;314(9):884-894. [doi: 10.1001/jama.2015.10081] [Medline: 26325557]
77. Zou H, Zhou B, Xu G. SGLT2 inhibitors: a novel choice for the combination therapy in diabetic kidney disease. Cardiovasc Diabetol 2017 May 16;16(1):65 [FREE Full text] [doi: 10.1186/s12933-017-0547-1] [Medline: 28511711]
78. Ren W, Wang M, Liao J, Li L, Yang D, Yao R, et al. The Effect of Chinese Herbal Medicine Combined With Western Medicine on Vascular Endothelial Function in Patients With Hypertension: A Systematic Review and Meta-Analysis of Randomized Controlled Trials. Front Pharmacol 2020;11:823 [FREE Full text] [doi: 10.3389/fphar.2020.00823] [Medline: 32612527]
79. Lu Z, Zhong Y, Liu W, Xiang L, Deng Y. The Efficacy and Mechanism of Chinese Herbal Medicine on Diabetic Kidney Disease. J Diabetes Res 2019;2019:2697672 [FREE Full text] [doi: 10.1155/2019/2697672] [Medline: 31534972]
80. Canaud G, Brooks CR, Kishi S, Taguchi K, Nishimura K, Magassa S, et al. Cyclin G1 and TASCC regulate kidney epithelial cell G-M arrest and fibrotic maladaptive repair. Sci Transl Med 2019 Jan 23;11(476) [FREE Full text] [doi: 10.1126/scitranslmed.aav4754] [Medline: 30674655]
81. Xu J, Zhou L, Liu Y. Cellular Senescence in Kidney Fibrosis: Pathologic Significance and Therapeutic Strategies. Front Pharmacol 2020;11:601325 [FREE Full text] [doi: 10.3389/fphar.2020.601325] [Medline: 33362554]
82. Ferenbach DA, Bonventre JV. Mechanisms of maladaptive repair after AKI leading to accelerated kidney ageing and CKD. Nat Rev Nephrol 2015 May;11(5):264-276 [FREE Full text] [doi: 10.1038/nrneph.2015.3] [Medline: 25643664]
83. Ma Y, Yan R, Wan Q, Lv B, Yang Y, Lv T, et al. Inhibitor of growth 2 regulates the high glucose-induced cell cycle arrest and epithelial-to-mesenchymal transition in renal proximal tubular cells. J Physiol Biochem 2020 Aug;76(3):373-382. [doi: 10.1007/s13105-020-00743-3] [Medline: 32424454]
84. Moonen L, D'Haese PC, Vervaet BA. Epithelial Cell Cycle Behaviour in the Injured Kidney. Int J Mol Sci 2018 Jul 13;19(7) [FREE Full text] [doi: 10.3390/ijms19072038] [Medline: 30011818]
85. Johnson DG, Walker CL. Cyclins and cell cycle checkpoints. Annu Rev Pharmacol Toxicol 1999;39:295-312. [doi: 10.1146/annurev.pharmtox.39.1.295] [Medline: 10331086]
86. Yang L, Besschetnova TY, Brooks CR, Shah JV, Bonventre JV. Epithelial cell cycle arrest in G2/M mediates kidney fibrosis after injury. Nat Med 2010 May;16(5):535-43, 1p following 143 [FREE Full text] [doi: 10.1038/nm.2144] [Medline: 20436483]

87. Canaud G, Bonventre JV. Cell cycle arrest and the evolution of chronic kidney disease from acute kidney injury. Nephrol Dial Transplant 2015 Apr;30(4):575-583 [FREE Full text] [doi: 10.1093/ndt/gfu230] [Medline: 25016609]
88. Wang JG, Pan L, Wu B, Wang M. Familial characteristics of kidney-yang deficiency and cold syndrome. J Toxicol Environ Health A 2006 Nov;69(21):1939-1950. [doi: 10.1080/15287390600751322] [Medline: 16982532]
89. Ding WJ, Zeng YZ, Li WH, Zhang TE, Liu WW, Teng XK, et al. Identification of Linkage Disequilibrium SNPs from a Kidney-Yang Deficiency Syndrome Pedigree. Am J Chin Med 2009;37(3):427-438. [doi: 10.1142/S0192415X09006953] [Medline: 19606505]
90. Liu WW, Gao YX, Zhou LP, Duan A, Tan LL, Li WZ, et al. Observations on Copy Number Variations in a Kidney-yang Deficiency Syndrome Family. Evid Based Complement Alternat Med 2011;2011:548358 [FREE Full text] [doi: 10.1093/ecam/neq069] [Medline: 21811512]
91. Lai M, Lai J, Chen P, Tseng W, Chen Y, Hwang J, et al. Increased risks of chronic kidney disease associated with prescribed Chinese herbal products suspected to contain aristolochic acid. Nephrology (Carlton) 2009 Apr;14(2):227-234. [doi: 10.1111/j.1440-1797.2008.01061.x] [Medline: 19076288]

Abbreviations

ACEi: angiotensin-converting enzyme inhibitor
aHR: adjusted hazard ratio
API: application programming interface
ARB: angiotensin receptor blocker
ARM: association rule mining
CHM: Chinese herbal medicine
CKD: chronic kidney disease
CMN: Chinese Herbal Medicine Network
DKD: diabetic kidney disease
DM: diabetes mellitus
DN: diabetic nephropathy
DRI: direct renin inhibitor
ESRD: end-stage renal disease
FDR: false discovery rate
GLP-1: glucagon-like peptide-1
HF: herbal formula
HR: hazard ratio
ICD-9-CM: International Classification of Diseases 9, Clinical Modification
IHD: ischemic heart disease
IPTW: inverse probability treatment weighting
NHIRD: National Health Insurance Research Database
NSAID: nonsteroidal anti-inflammatory drug
RAAS: renin-angiotensin-aldosterone system
SGLT2i: sodium-glucose cotransporter 2 inhibitor
SH: single herb
SNA: social network analysis
TCM: Traditional Chinese medicine
WM: Western medicine

Edited by G Eysenbach; submitted 31.01.21; peer-reviewed by S Wei; comments to author 22.02.21; revised version received 01.03.21; accepted 11.04.21; published 11.05.21.

Please cite as:
Wu CW, Chen HY, Yang CW, Chen YC
Deciphering the Efficacy and Mechanisms of Chinese Herbal Medicine for Diabetic Kidney Disease by Integrating Web-Based Biochemical Databases and Real-World Clinical Data: Retrospective Cohort Study
JMIR Med Inform 2021;9(5):e27614
URL: https://medinform.jmir.org/2021/5/e27614
doi:10.2196/27614
PMID:33973855

Original Paper

Federated Learning for Thyroid Ultrasound Image Analysis to Protect Personal Information: Validation Study in a Real Health Care Environment

Haeyun Lee[1,2*], MSc; Young Jun Chai[3*], MD, PhD; Hyunjin Joo[1,4], BA; Kyungsu Lee[1,2], BS; Jae Youn Hwang[2], PhD; Seok-Mo Kim[5], MD, PhD; Kwangsoon Kim[6], MD, PhD; Inn-Chul Nam[7], MD, PhD; June Young Choi[8], MD, PhD; Hyeong Won Yu[8], MD, PhD; Myung-Chul Lee[9], MD, PhD; Hiroo Masuoka[10], MD, PhD; Akira Miyauchi[10], MD, PhD; Kyu Eun Lee[1,11], MD, PhD; Sungwan Kim[1,4,12], PhD; Hyoun-Joong Kong[1,4,13], PhD

[1]Institute of Medical & Biological Engineering, Medical Research Center, Seoul National University College of Medicine, Seoul, Republic of Korea

[2]Department of Information and Communication Engineering, Daegu Gyeongbuk Institute of Science & Technology, Daegu, Republic of Korea

[3]Department of Surgery, Seoul Metropolitan Government Seoul National University Boramae Medical Center, Seoul, Republic of Korea

[4]Transdisciplinary Department of Medicine and Advanced Technology, Seoul National University Hospital, Seoul, Republic of Korea

[5]Department of Surgery, Thyroid Cancer Center, Gangnam Severance Hospital, Seoul, Republic of Korea

[6]Department of Surgery, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea

[7]Department of Otolaryngology-Head and Neck Surgery, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea

[8]Department of Surgery, Seoul National University Bundang Hospital, Seongnam-si, Gyeonggi-do, Republic of Korea

[9]Department of Otorhinolaryngology-Head and Neck Surgery, Korea Cancer Center Hospital, Korea Institute of Radiological and Medical Science, Seoul, Republic of Korea

[10]Department of Surgery, Kuma Hospital, Kobe, Japan

[11]Department of Surgery, Seoul National University Hospital and College of Medicine, Seoul, Republic of Korea

[12]Department of Biomedical Engineering, Seoul National University College of Medicine, Seoul, Republic of Korea

[13]Department of Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Hyoun-Joong Kong, PhD
Transdisciplinary Department of Medicine and Advanced Technology
Seoul National University Hospital
Daehak-ro 101
Jongno-gu
Seoul
Republic of Korea
Phone: 82 2 2072 4492
Email: gongcop@gmail.com

Abstract

Background: Federated learning is a decentralized approach to machine learning; it is a training strategy that overcomes medical data privacy regulations and generalizes deep learning algorithms. Federated learning mitigates many systemic privacy risks by sharing only the model and parameters for training, without the need to export existing medical data sets. In this study, we performed ultrasound image analysis using federated learning to predict whether thyroid nodules were benign or malignant.

Objective: The goal of this study was to evaluate whether the performance of federated learning was comparable with that of conventional deep learning.

Methods: A total of 8457 (5375 malignant, 3082 benign) ultrasound images were collected from 6 institutions and used for federated learning and conventional deep learning. Five deep learning networks (VGG19, ResNet50, ResNext50, SE-ResNet50, and SE-ResNext50) were used. Using stratified random sampling, we selected 20% (1075 malignant, 616 benign) of the total images for internal validation. For external validation, we used 100 ultrasound images (50 malignant, 50 benign) from another institution.

Results: For internal validation, the area under the receiver operating characteristic (AUROC) curve for federated learning was between 78.88% and 87.56%, and the AUROC for conventional deep learning was between 82.61% and 91.57%. For external validation, the AUROC for federated learning was between 75.20% and 86.72%, and the AUROC curve for conventional deep learning was between 73.04% and 91.04%.

Conclusions: We demonstrated that the performance of federated learning using decentralized data was comparable to that of conventional deep learning using pooled data. Federated learning might be potentially useful for analyzing medical images while protecting patients' personal information.

(JMIR Med Inform 2021;9(5):e25869) doi:10.2196/25869

KEYWORDS

deep learning; federated learning; thyroid nodules; ultrasound image

Introduction

Deep neural networks for image classification, object detection, and semantic segmentation have been proven to be high performance, surpassing human-level performance in some fields [1]. Deep learning for computer aided diagnosis has been frequently reported using various medical imaging modalities, such as ultrasound images, computed tomography, and magnetic resonance imaging. As in other fields, the ability for deep learning using medical images to surpass human-level performance is dependent on the volume and quality of data [2,3].

There are several challenges in the implementation of deep learning in the clinical environment. To obtain a sufficient number of medical images for high performance, medical images must be collected from multiple institutions. Personal information protection may be violated during the data collection process. Heterogeneity of data between contributing institutes is another issue that can negatively influence the performance of a deep learning network. Distribution of data varies considerably between institutions in terms of disease entities, as does the volume, location, and characteristics of medical images; this influences the performance of deep learning networks.

Federated learning is a technique used to build learning networks without the need for centralized data that is hugely advantageous in a health care context where data protection and patient confidentiality are paramount. Federated learning mitigates many systemic privacy risks by sharing with each local data source only the model and trained parameters for network training, without the need to export existing medical data sets. Network parameters that are trained with data from each local data source are aggregated in one place and are updated and sent back to each local data source. The network is trained as this process is repeatedly executed.

Although federated learning does not require the exchange of local data (ie, each medical institution's data), it's performance is similar to that of conventional deep learning. Federated learning has been applied to multiple open data sets such as Modified National Institute of Standards and Technology (MNIST) [4], Canadian Institute for Advanced Research (CIFAR-10) [4], and Brain Tumor Segmentation challenge (BraTS) 2018 [5,6] data sets. Various methods [4,6] have been applied to optimize the performance of federated learning. The application of federated learning for personal health information from wearable devices has also been reported [7]. These studies [4-7] demonstrated that federated learning is similar in performance to conventional deep learning (ie, data centralized training) approaches; however, they used either general image data, or if used, medical image data were few in number (for example, open medical image data sets such as BraTS 2018 contain only a few hundred images). In addition, the images were from one institution, and only one deep learning network was used. In real-world health care environments, when deep learning is applied, data distributions are frequently unbalanced.

In this study, we collected thyroid ultrasound images from medical institutions to evaluate the feasibility and performance of federated learning.

Methods

Thyroid Nodule Clinical Data Collection

The institutional review boards at all participating institutions (Seoul Metropolitan Government Seoul National University Boramae Medical Center, Gangnam Severance Hospital, Seoul National University Bundang Hospital, Catholic University of Korea Incheon St. Mary's Hospital, Catholic University of Korea Seoul St. Mary's Hospital, and Korea Cancer Center Hospital) approved this study. Representative institutional review board approval was granted by Seoul Metropolitan Government Seoul National University Boramae Medical Center (H-10-2020-195).

Images were collected from 6 medical institutions in captured DICOM file format (Figure 1). Of the 6 institutions, 3 used iU22 systems (Philips Healthcare), one used EPIQ 5G (Philips Healthcare), one used Prosound Alpha 7 (Hitachi Aloka), and one used Aplio 500 Platinum (Toshiba Medical Systems). Experienced surgeons at each institution labeled the images as *benign* (fine-needle aspiration cytology Bethesda Category II or benign surgical histology) or *malignant* (fine-needle aspiration cytology Bethesda Category V/VI or surgical histology of thyroid carcinoma). The images were cropped into 299×299 pixels to include typical thyroid features. The images were not augmented.

Figure 1. Thyroid ultrasound image data collected from 6 medical institutions to verify federated learning.

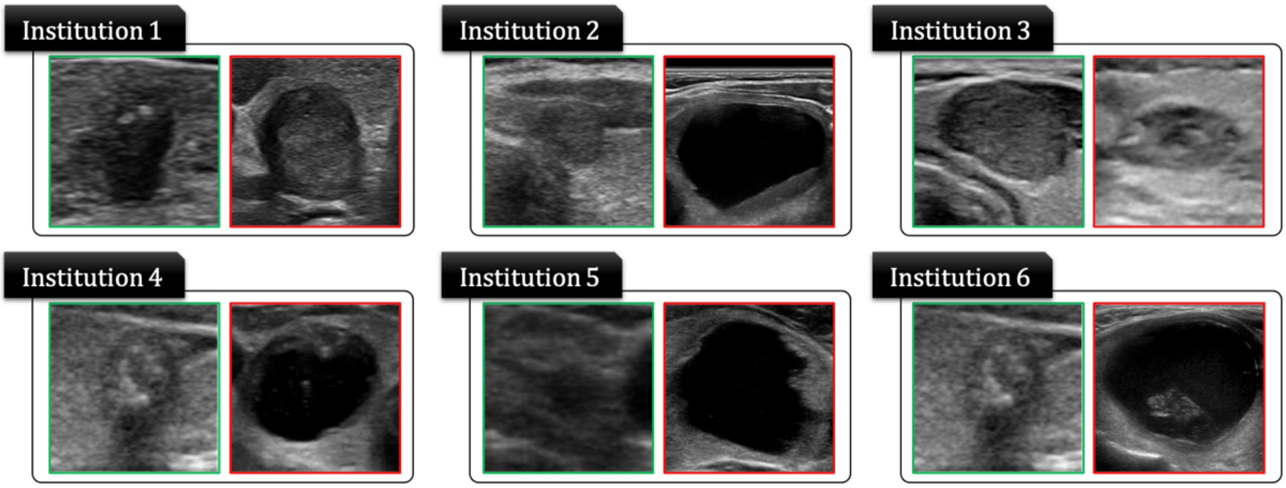


Table 1 summarizes details of the thyroid ultrasound images used in this experiment. We used 80% of each institution's data as training data and the remaining 20% as test data. We used stratified random sampling to select the test data set. There was a total of 4300 malignant images and 2465 benign images in the total training data set and a total of 1075 malignant images and 617 benign images in the test data set. For external validation, 100 thyroid ultrasound images (50 malignant image data and 50 benign) were provided by a medical institution in Japan. We were blinded to the labeling (malignant or benign) of the images.

Table 1. Thyroid ultrasound image data from 6 medical institutions used to validate federated learning.

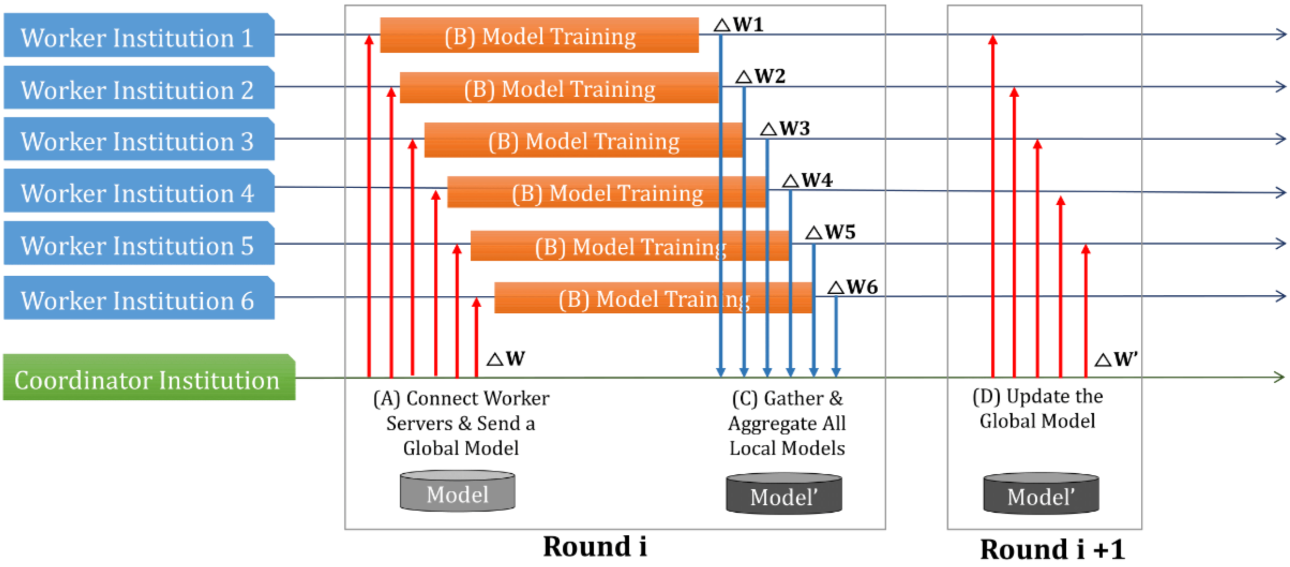
Class	Institution 1, n	Institution 2, n	Institution 3, n	Institution 4, n	Institution 5, n	Institution 6, n	Total, n
Malignant	**1233**	**3191**	**469**	**106**	**99**	**277**	**5375**
Training	986	2553	375	85	79	222	4300
Test	247	638	94	21	20	55	1075
Benign	**2257**	**291**	**10**	**100**	**100**	**324**	**3082**
Training	1806	233	8	80	80	259	2466
Test	451	58	2	20	20	65	616

In addition, to verify the performance of federated learning with external data, we collected an external test data set, which consisted of 50 malignant and 50 benign ultrasound images taken using a TUS-A500 system (Toshiba Medical System) from Kuma Hospital.

Federated Learning System Design in a Real Health Care Environment

We conducted federated learning experiments (Figure 2) with each institution's serverworker (a computer system that can train deep learning algorithms with local data in the federated learning process) and the coordinator of Seoul National University Hospital to validate federated learning in a real health care environment (serverworker system at each institution: Intel 4-core 2.3 GHz i5-8259U processor, 16 GB DDR4 RAM memory, and 11 GB Nvidia RTX 2080 Ti graphics; coordinator system: 2.3 GHz i5-8259U processor, 16 GB DDR4 RAM, and 8 GB Nvidia GTX 1080). Network training was performed on the serverworkers, and then each serverworker was configured with a high-memory graphic process unit. We configured the system using the processor and external graphics processing unit for system portability. All versions of software (Python version 3.6.5; PyTorch version 1.4.0; PySyft version 0.2.5) were identical between institutions. We installed Ubuntu 18.04 LTS version on each serverworker and the coordinator system.

Figure 2. Federated learning procedure in a real-world health care environment. (A) The serverworker from each medical institution (upper 6 medical institutions) was trained with local data from their corresponding medical institution. (B) Trained parameters were sent from each institution to the coordinator. (C) The coordinator averaged the parameters received from each institution. (D) The average value was sent back to each serverworker.



Deep Learning Algorithm

We used 5 deep neural network classifiers for thyroid ultrasound image analysis: VGG19 [8], ResNet50 [9], ResNext50 [10], SE-ResNet50, and SE-ResNext50 [11]. We also used these 5 models to verify federated learning.

Stochastic optimization (ADAM) was used with the following parameters: β_1=0.9, β_2=0.999, =10^{-8} [12]. The initial learning rate was 0.001 which was reduced by half every 30 rounds. The mini-batch size was 32. We used a binary cross-entropy loss function to train all networks. We trained the network for 120 rounds. We used PyTorch [13] and PySyft [14] to implement and train all networks with federated learning.

Conventional Deep Learning Using Pooled Data

After removing all patient identifying information, images from each participating institution were collected at Seoul National University Hospital to create a pooled data set. We used the pooled data set to conduct conventional deep learning. All settings were the same as those for federated learning, with the exception of those used in PySyft, and the same equipment, with the same specifications as those of the serverworker, was used. Only training data from each hospital used in the federated learning were pooled and used for conventional deep learning. The test data set was the same as that used for federated learning.

Results

Federated Learning Performance

For the internal test data set, consisting of 1691 images (1075 malignant and 616 benign), and federated learning–trained deep learning algorithms, the accuracies of VGG19, SE-ResNet50, ResNet50, SE-ResNext50, and ResNext50 were 79.5%, 77.9%, 77.4%, 77.2%, and 73.9%, respectively (Table 2; Table S1 in Multimedia Appendix 1). Figure 3 shows the receiver operating characteristic curve [15] of each network for the internal test data set. Area under the receiver operating characteristic (AUROC) curve values of SE-ResNext50, ResNext50, VGG19, SE-ResNet50, and ResNet50 were 87.6%, 86.0%, 82.0%, 79.9%, and 78.9%, respectively.

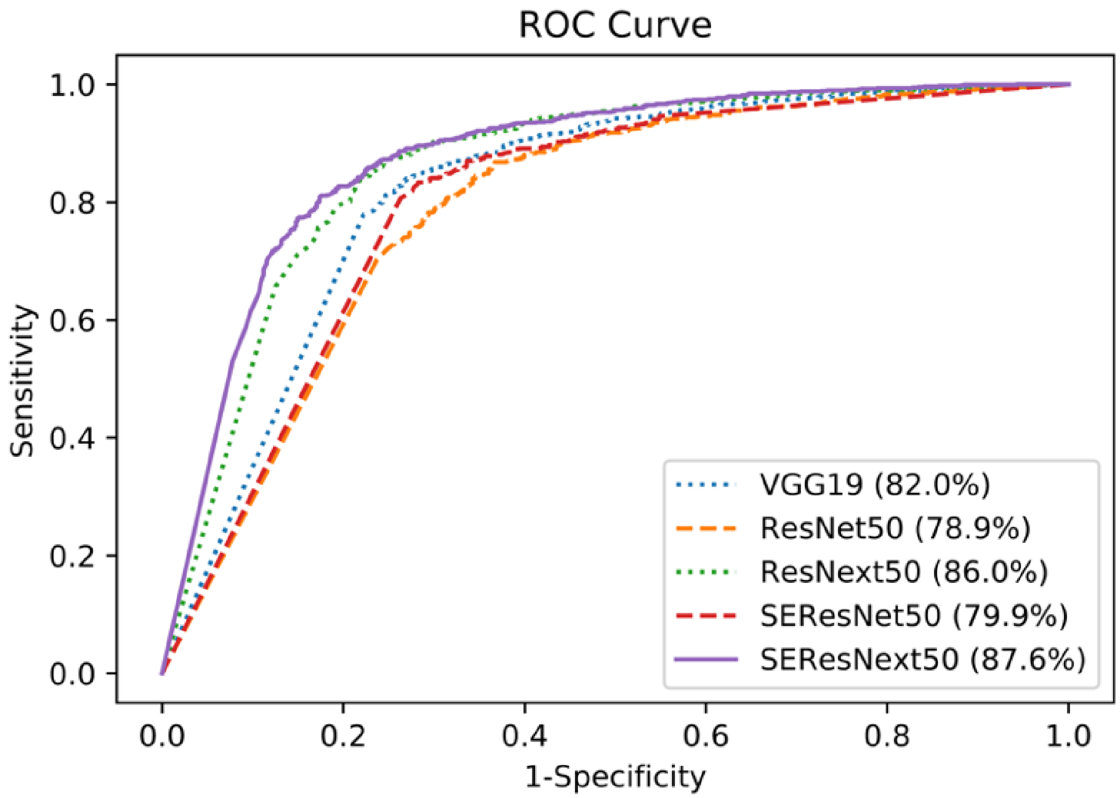
Table 2. Thyroid classification results with federated learning with internal test data.

Deep learning algorithm	Accuracy (%)	Specificity (%)	Sensitivity (%)	PPV[a] (%)	NPV[b] (%)	F1 score (%)	AUROC (%)
VGG19	79.5	64.3	88.2	81.2	75.7	84.5	82.0
ResNet50	77.4	57.8	88.6	78.6	74.3	83.3	78.9
ResNext50	73.9	31.5	98.2	71.5	91.1	82.7	86.0
SE-ResNet50	77.9	56.3	90.2	78.3	76.8	83.8	79.9
SE-ResNext50	77.2	42.1	97.3	74.6	90.0	84.4	87.6

[a] PPV: positive predictive value.

[b] NPV: negative predictive value.

Figure 3. Receiver operating characteristic curves of each deep learning network for the internal test data set.



For the external test data set and federated learning model, the accuracies of ResNet50, SE-ResNet50, VGG19, SE-ResNext50, and ResNext50 were 76.0%, 73.0%, 69.0%, 60.0%, and 56.0%, respectively (Table 3; Table S2 in Multimedia Appendix 1). AUROC curve values of SE-ResNet50, SE-ResNext50, ResNext50, ResNet50, and VGG19 were 86.7%, 83.4%, 83.0%, 81.0%, and 75.2%, respectively.

Table 3. Thyroid classification results with federated learning with external test data.

Deep learning algorithm	Accuracy (%)	Specificity (%)	Sensitivity (%)	PPV[a] (%)	NPV[b] (%)	F1 score (%)	AUROC (%)
VGG19	69.0	52.0	86.0	64.2	78.8	73.5	75.2
ResNet50	76.0	58.0	94.0	69.1	90.6	79.7	81.0
ResNext50	56.0	12.0	100	53.2	100	69.4	83.0
SE-ResNet50	73.0	48.0	98.0	65.3	96.0	78.4	86.7
SE-ResNext50	60.0	20.0	100	55.6	100	71.4	83.4

[a] PPV: positive predictive value.

[b] NPV: negative predictive value.

Performance of Conventional Deep Learning Using Pooled Data

For each deep learning algorithm trained with the pooled data, the accuracies of VGG19, ResNet50, ResNext50, SE-ResNet50, and SE-ResNext50 were 81.5%, 78.7%, 85.2%, 83.2%, and 85.2%, respectively (Table 4; Table S3 in Multimedia Appendix 1). The AUROC curve values of VGG19, ResNet50, ResNext50, SE-ResNet50, and SE-ResNext50 were 87.6%, 82.6%, 91.0%, 84.5%, and 91.5%, respectively.

For conventional deep learning using the pooled external test data set, the accuracies of VGG19, ResNet50, ResNext50, SE-ResNet50, and SE-ResNext50 were 71.0%, 77.0%, 80.0%, 66.0%, and 76.0%, respectively (Table 5; Table S4 in Multimedia Appendix 1). The AUROC curve values of VGG19, ResNet50, ResNext50, SE-ResNet50, and SE-ResNext50 were 79.3%, 81.2%, 89.7%, 73.4%, and 91.0%, respectively.

Table 4. Thyroid classification results with conventional deep learning using pooled internal test data.

Deep learning algorithm	Accuracy (%)	Specificity (%)	Sensitivity (%)	PPV[a] (%)	NPV[b] (%)	F1 score (%)	AUROC (%)
VGG19	81.5	62.0	92.7	81.0	83.0	86.5	87.6
ResNet50	78.7	62.8	87.7	80.5	74.6	83.9	82.6
ResNext50	85.2	72.5	92.5	85.5	84.7	88.8	91.0
SE-ResNet50	83.2	70.0	90.7	84.1	81.2	82.7	84.5
SE-ResNext50	85.3	70.9	93.5	84.9	86.2	89.0	91.5

[a] PPV: positive predictive value.

[b] NPV: negative predictive value.

Table 5. Thyroid classification results with conventional deep learning using pooled external test data.

Deep learning algorithm	Accuracy (%)	Specificity (%)	Sensitivity (%)	PPV[a] (%)	NPV[b] (%)	F1 score (%)	AUROC (%)
VGG19	71.0	56.0	86.0	66.2	80.0	74.8	79.3
ResNet50	77.0	72.0	82.0	74.5	80.0	78.1	81.2
ResNext50	80.0	72.0	88.0	75.9	85.7	81.5	89.7
SE-ResNet50	66.0	48.0	84.0	61.8	75.0	71.2	73.4
SE-ResNext50	76.0	58.0	94.0	69.1	90.6	79.7	91.0

[a] PPV: positive predictive value.

[b] NPV: negative predictive value.

Discussion

Principal Results

The goal of this study was to verify the performance of federated learning in a real-world health care environment. We first collected thyroid nodule data from 6 institutions and designed a federated learning system using these data. We trained each deep learning algorithm (VGG19, ResNet50, ResNext50, SE-ResNet50, and SE-ResNext50) with the federated learning system. We also trained the same deep learning algorithms using conventional deep learning techniques and compared the performance of federated learning with that of conventional deep learning.

Comparison With Prior Work

The medical vision community is currently actively conducting diagnosis using computer-aided diagnosis [16]. To improve the performance of computer-aided diagnosis, several deep learning algorithms have been developed and applied [17-20]. Various challenges for deep learning with open data sets have been identified [21,22]. In particular, due to health care data privacy regulations, most open data sets only have a small amount of data collected from a single institution. When training and validation are carried out with only a small volume of data, the performance of a deep learning model cannot be properly evaluated, and generality cannot properly be validated. Federated learning, which can train a deep learning model without centralized data, offers a training strategy that addresses these challenges.

There have been several recent reports of the use of federated learning trained with general images [4] and medical imaging [5,6]. McMahan et al [4] published a study using federated learning with federated averaging and reported that the average parameters trained from each serverworker each round performed similarly to those of conventional deep learning and better than those of federated stochastic gradient descent; however, the study used a relatively simple model and general image data sets (MNIST and CIFAR-10). Sheller et al [5] compared federated learning, institutional incremental learning (IIL), and cyclic IIL using the BraTS 2018 data set [21]. IIL is a collaborative learning process that trains a network with data from one institution and then continues training with another institution’s data successively. One disadvantage of this model is that when the network is trained using data from another institution, the patterns trained from the previous institutions’ data are disregarded. To compensate for this shortcoming, Sheller et al [5] proposed cyclic IIL which repeats the IIL process. They used U-Net architecture [17] for brain tumor segmentation with federated learning, IIL, and cyclic IIL and demonstrated that the performance of federated learning was superior to those of IIL and cyclic IIL; however, the study applied federated learning but did not address the class imbalance or data volume imbalance problems associated with federated learning. Li et al [6] also used the BraTS 2018 data set to compare federated learning and centralized data training; they found no significant difference in performance between federated learning and centralized data training. Most federated learning studies compare federated learning with conventional deep learning only, and there are no studies using clinical data from a real-world health care environment.

The application of federated learning in our study shows that this technology has substantial potential applicability in clinical environments. First, federated learning showed performance comparable with that of conventional deep learning, despite an extremely uneven distribution of data volume from each institution. The difference between the hospital with the most data and the hospital with the least data was 17.5 fold. Moreover, the distribution of benign and malignant images was also skewed. For example, the ratio of malignant to benign images was 47:1 for institution 3, whereas it was 1:2 for institution 1. Because data distributions between hospitals are diverse, the conditions presented in this study demonstrated the applicability of federated learning in the real world and its ability to facilitate collaboration between different size institutions.

In medical image analysis, if the amount of data is insufficient, overfitting (learning from noise in data) often occurs. In such cases, only the accuracy of the internal data set is high, and deep learning algorithms cannot be rigorously evaluated. We were able to overcome the issue of overfitting by collecting images from multiple institution and by performing external validation using images from an institute in a different country. We demonstrated that federated learning is able to maximize the efficiency of medical resources and generalizability of deep learning algorithms using data from different size medical institutions (with various imaging devices and different patient groups). This represents scenarios in real-world health care environments [23-26].

In our study, federated learning training took at least 4 times longer than that of conventional deep learning. The training time for federated learning varied depending on the peripheral environment such as internet speed and temperature of graphics process unit. The performance of federated learning may be enhanced with more images or data augmentation. The ideal volume of data and the distribution of data contributed by each institution for peak performance of federated learning is also not yet known. Further investigation into the optimal training environment, training time, data volume, data distribution, and state-of-the-art deep learning algorithms is required for federated learning.

As shown in Table 5, we noted that when thyroid nodules were classified by a conventional deep learning model, the number of malignant calls was extremely high. The same trend is frequently observed in the literature [20,27-29]. As shown in Table 3, we also found this trend to be prominent in federated learning. Because deep learning is a black box [30], we were unable to determine the potential reasons for this tendency, but we plan to investigate this phenomenon in the future.

Limitations

This study has several limitations. First, we presented the results of federated learning used in a specific context in terms of the number of participating institutions, and the number and ratio of benign and malignant images. Thus, the generalizability of the results in terms of the performance of federated learning is not known and warrants further investigation. We also used thyroid ultrasound images, which are relatively easy to analyze compared to those from computed tomography, magnetic resonance imaging, and histopathology sections. Results may not be generalizable across different imaging modalities. In future work, comparisons of federated learning with unequal data distribution, data augmentation, one-shot learning are required to explore the implications of data imbalance.

Conclusions

We demonstrated that the performance of federated learning using a shared training model and parameters from 6 institutions was comparable with that of conventional deep learning using pooled data. Federated learning is highly generalizable because it can effectively utilize data collected from different environments despite data heterogeneity. Federated learning has the potential to mitigate many systemic privacy risks by sharing only the model and parameters for training without the need to export existing medical data sets.

Acknowledgments

This research was supported by the Technology Innovation Program (20011878; Development of Diagnostic Medical Devices with Artificial Intelligence Based Image Analysis Technology) funded by the Ministry of Trade, Industry & Energy (Korea), and by the Ministry of Science and Information & Communications Technology (Korea), under the Information Technology Research Center support program (IITP-2021-2018-0-01833), supervised by the Institute for Information & Communications Technology Promotion.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Tables with additional information.
[DOCX File , 42 KB - medinform_v9i5e25869_app1.docx]

References

1. Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. Commun ACM 2017 May 24;60(6):84-90. [doi: 10.1145/3065386]
2. 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 Dec 02;542(7639):115-118. [doi: 10.1038/nature21056] [Medline: 28117445]

3. Abdel-Zaher AM, Eldeib AM. Breast cancer classification using deep belief networks. Expert Syst Appl 2016 Mar;46:139-144. [doi: 10.1016/j.eswa.2015.10.015]
4. McMahan B, Moore E, Ramage D, Hampson S, Aguera y Arcas B. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. 2017 Presented at: 20th International Conference on Artificial Intelligence and Statistics; April 20-22; Fort Lauderdale, FL p. 1273-1282.
5. Sheller M, Reina G, Edwards B, Martin J, Bakas S. Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation. Brainlesion 2019;11383:92-104 [FREE Full text] [doi: 10.1007/978-3-030-11723-8_9] [Medline: 31231720]
6. Li W, Milletarì F, Xu D, Rieke N, Hancox J, Zhu W, et al. Privacy-preserving federated brain tumour segmentation. In: Suk HI, Liu M, Yan P, Lian C, editors. Machine Learning in Medical Imaging Lecture Notes in Computer Science Vol 11861. Cham: Springer; 2019:133-141.
7. Chen Y, Qin X, Wang J, Yu C, Gao W. FedHealth: a federated transfer learning framework for wearable healthcare. IEEE Intell Syst 2020 Jul 1;35(4):83-93. [doi: 10.1109/mis.2020.2988604]
8. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv Preprint posted online on April 10, 2015 [FREE Full text]
9. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; June 27-30; Las Vegas, Nevada p. 770-778. [doi: 10.1109/cvpr.2016.90]
10. Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; July 21-26; Honolulu, Hawaii p. 1492-1500. [doi: 10.1109/cvpr.2017.634]
11. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell 2020 Aug;42(8):2011-2023. [doi: 10.1109/TPAMI.2019.2913372] [Medline: 31034408]
12. Yoon D, Lim HS, Jung K, Kim TY, Lee S. Deep learning-based electrocardiogram signal noise detection and screening model. Healthc Inform Res 2019 Jul;25(3):201-211 [FREE Full text] [doi: 10.4258/hir.2019.25.3.201] [Medline: 31406612]
13. Adam P, Sam G, Soumith C, Gregory C, Edward Y, Zachary D, et al. Automatic differentiation in PyTorch. 2017 Presented at: 31st Annual Conference on Neural Information Processing Systems; December 4-9; Long Beach, California.
14. Ryffel T, Trask A, Dahl M, Wagner B, Mancuso J, Rueckert D, et al. A generic framework for privacy preserving deep learning. arXiv Preprint posted online on November 13, 2018 [FREE Full text]
15. Yu JY, Jeong GY, Jeong OS, Chang DK, Cha WC. Machine learning and initial nursing assessment-based triage system for emergency department. Healthc Inform Res 2020 Jan;26(1):13-19 [FREE Full text] [doi: 10.4258/hir.2020.26.1.13] [Medline: 32082696]
16. Johnston ME, Langton KB, Haynes RB, Mathieu A. Effects of computer-based clinical decision support systems on clinician performance and patient outcome. a critical appraisal of research. Ann Intern Med 1994 Jan 15;120(2):135-142. [doi: 10.7326/0003-4819-120-2-199401150-00007] [Medline: 8256973]
17. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. 2015 Oct Presented at: International Conference on Medical Image Computing and Computer-Assisted Intervention; October 5-9; Munich, Germany p. 234-241. [doi: 10.1007/978-3-319-24574-4_28]
18. Lee H, Park J, Hwang J. Channel attention module with multiscale grid average pooling for breast cancer segmentation in an ultrasound image. IEEE Trans Ultrason Ferroelect Freq Contr 2020 Feb 10:1-1 [FREE Full text] [doi: 10.1109/tuffc.2020.2972573]
19. Youn S, Lee K, Son J, Yang IH, Hwang JY. Fully-automatic deep learning-based analysis for determination of the invasiveness of breast cancer cells in an acoustic trap. Biomed Opt Express 2020 Jun 01;11(6):2976-2995 [FREE Full text] [doi: 10.1364/BOE.390558] [Medline: 32637236]
20. Song J, Chai Y, Masuoka H, Park S, Kim S, Choi J, et al. Ultrasound image analysis using deep learning algorithm for the diagnosis of thyroid nodules. Medicine (Baltimore) 2019 Apr 12;98(15):e15133. [doi: 10.1097/md.0000000000015133]
21. Isensee F, Kickingereder P, Wick W, Bendszus M, Maier-Hein K. Brain tumor segmentation and radiomics survival prediction: contribution to the BRATS 2017 challenge. 2017 Presented at: International MICCAI Brainlesion Workshop; September 17; Quebec City, Canada p. 287-297. [doi: 10.1007/978-3-319-75238-9_25]
22. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018 Feb 22;172(5):1122-1131.e9 [FREE Full text] [doi: 10.1016/j.cell.2018.02.010] [Medline: 29474911]
23. He H, Garcia E. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2009 Sep;21(9):1263-1284. [doi: 10.1109/TKDE.2008.239]
24. Raudys S, Jain A. Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans Pattern Anal Machine Intell 1991;13(3):252-264. [doi: 10.1109/34.75512]

25. Tzeng D, Chung W, Lin C, Yang C. Effort-reward imbalance and quality of life of healthcare workers in military hospitals: a cross-sectional study. BMC Health Serv Res 2012 Sep 08;12(1):309 [FREE Full text] [doi: 10.1186/1472-6963-12-309] [Medline: 22958365]
26. Sassaroli E, Crake C, Scorza A, Kim D, Park M. Image quality evaluation of ultrasound imaging systems: advanced B-modes. J Appl Clin Med Phys 2019 Mar;20(3):115-124 [FREE Full text] [doi: 10.1002/acm2.12544] [Medline: 30861278]
27. Nguyen DT, Pham TD, Batchuluun G, Yoon HS, Park KR. Artificial intelligence-based thyroid nodule classification using information from spatial and frequency domains. J Clin Med 2019 Nov 14;8(11):1976 [FREE Full text] [doi: 10.3390/jcm8111976] [Medline: 31739517]
28. Wang J, Li S, Song W, Qin H, Zhang B, Hao A. Learning from weakly-labeled clinical data for automatic thyroid nodule classification in ultrasound images. 2018 Oct Presented at: 25th IEEE International Conference on Image Processing; October 7-10; Athens, Greece p. 3114-3118. [doi: 10.1109/icip.2018.8451085]
29. Ma J, Wu F, Zhu J, Xu D, Kong D. A pre-trained convolutional neural network based method for thyroid nodule diagnosis. Ultrasonics 2017 Jan;73:221-230. [doi: 10.1016/j.ultras.2016.09.011] [Medline: 27668999]
30. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box model. ACM Comput Surv 2019 Jan 23;51(5):1-42. [doi: 10.1145/3236009]

Abbreviations

AUROC: area under the receiver operating characteristic curve
BraTS: Brain Tumor Segmentation challenge data set
CIFAR: Canadian Institute for Advanced Research
IIL: institutional incremental learning
MNIST: Modified National Institute of Standards and Technology

Edited by G Eysenbach; submitted 19.11.20; peer-reviewed by Y Cho, S Kim; comments to author 10.12.20; revised version received 02.02.21; accepted 03.04.21; published 18.05.21.

Please cite as:
Lee H, Chai YJ, Joo H, Lee K, Hwang JY, Kim SM, Kim K, Nam IC, Choi JY, Yu HW, Lee MC, Masuoka H, Miyauchi A, Lee KE, Kim S, Kong HJ
Federated Learning for Thyroid Ultrasound Image Analysis to Protect Personal Information: Validation Study in a Real Health Care Environment
JMIR Med Inform 2021;9(5):e25869
URL: https://medinform.jmir.org/2021/5/e25869
doi:10.2196/25869
PMID:33858817

Original Paper

Predicting Prolonged Length of Hospital Stay for Peritoneal Dialysis–Treated Patients Using Stacked Generalization: Model Development and Validation Study

Guilan Kong[1,2*], PhD[‡]; Jingyi Wu[2*], MS; Hong Chu[3], MD, PhD; Chao Yang[3], MS; Yu Lin[4], MPH; Ke Lin[1], MS; Ying Shi[5], BMS; Haibo Wang[1,6], MSc, MPH, MBBS; Luxia Zhang[1,2,3], MPH, MD

[1]National Institute of Health Data Science, Peking University, Beijing, China

[2]Advanced Institute of Information Technology, Peking University, Hangzhou, China

[3]Renal Division, Department of Medicine, Peking University First Hospital, Peking University Institute of Nephrology, Beijing, China

[4]Department of Medicine and Therapeutics, LKS Institute of Health Science, The Chinese University of Hong Kong, Hong Kong, China

[5]China Standard Medical Information Research Center, Shenzhen, China

[6]Clinical Trial Unit, First Affiliated Hospital of Sun Yat-Sen University, Guangzhou, China

[‡]China Kidney Disease Network Working Group

[*]these authors contributed equally

Corresponding Author:
Luxia Zhang, MPH, MD
National Institute of Health Data Science
Peking University
No 38 Xueyuan Road
Haidian District
Beijing, 100191
China
Phone: 86 10 82806538
Email: zhanglx@bjmu.edu.cn

Abstract

Background: The increasing number of patients treated with peritoneal dialysis (PD) and their consistently high rate of hospital admissions have placed a large burden on the health care system. Early clinical interventions and optimal management of patients at a high risk of prolonged length of stay (pLOS) may help improve the medical efficiency and prognosis of PD-treated patients. If timely clinical interventions are not provided, patients at a high risk of pLOS may face a poor prognosis and high medical expenses, which will also be a burden on hospitals. Therefore, physicians need an effective pLOS prediction model for PD-treated patients.

Objective: This study aimed to develop an optimal data-driven model for predicting the pLOS risk of PD-treated patients using basic admission data.

Methods: Patient data collected using the Hospital Quality Monitoring System (HQMS) in China were used to develop pLOS prediction models. A stacking model was constructed with support vector machine, random forest (RF), and K-nearest neighbor algorithms as its base models and traditional logistic regression (LR) as its meta-model. The meta-model used the outputs of all 3 base models as input and generated the output of the stacking model. Another LR-based pLOS prediction model was built as the benchmark model. The prediction performance of the stacking model was compared with that of its base models and the benchmark model. Five-fold cross-validation was employed to develop and validate the models. Performance measures included the Brier score, area under the receiver operating characteristic curve (AUROC), estimated calibration index (ECI), accuracy, sensitivity, specificity, and geometric mean (Gm). In addition, a calibration plot was employed to visually demonstrate the calibration power of each model.

Results: The final cohort extracted from the HQMS database consisted of 23,992 eligible PD-treated patients, among whom 30.3% had a pLOS (ie, longer than the average LOS, which was 16 days in our study). Among the models, the stacking model achieved the best calibration (ECI 8.691), balanced accuracy (Gm 0.690), accuracy (0.695), and specificity (0.701). Meanwhile, the stacking and RF models had the best overall performance (Brier score 0.174 for both) and discrimination (AUROC 0.757 for the stacking model and 0.756 for the RF model). Compared with the benchmark LR model, the stacking model was superior in all performance measures except sensitivity, but there was no significant difference in sensitivity between the 2 models. The 2-sided *t* tests revealed significant performance differences between the stacking and LR models in overall performance, discrimination, calibration, balanced accuracy, and accuracy.

Conclusions: This study is the first to develop data-driven pLOS prediction models for PD-treated patients using basic admission data from a national database. The results indicate the feasibility of utilizing a stacking-based pLOS prediction model for PD-treated patients. The pLOS prediction tools developed in this study have the potential to assist clinicians in identifying patients at a high risk of pLOS and to allocate resources optimally for PD-treated patients.

(JMIR Med Inform 2021;9(5):e17886) doi:10.2196/17886

KEYWORDS

peritoneal dialysis; prolonged length of stay; machine learning; prediction model; clinical decision support

Introduction

Over the past 30 years, the United States Renal Data System has reported a rapid increase in the incidence of end-stage kidney disease (ESKD) [1]. The increasing number of patients with ESKD treated with kidney replacement therapy—including hemodialysis, peritoneal dialysis (PD), and renal transplantation—has put a large burden on the health care system. Approximately 2.6 million people worldwide received kidney replacement therapy in 2010 [2], and the prevalence of ESKD in China was 237.3 cases per million population in 2012 [3]. In 2015, the average inpatient expenditure for patients with ESKD in China was approximately ¥24,800 (US $3793) [4], and the total inpatient expenditure for patients with ESKD in China was in excess of ¥6.75 billion (US $1.03 billion). In 2016, the average expenditure on patients with ESKD in the United States was estimated to be US $50 billion, one-third of which was attributed to hospitalization costs [1]. Hospitalization remains a critical outcome for patients with ESKD, and the risk of hospitalization in patients undergoing dialysis is triple that of patients without ESKD [5]. In-hospital length of stay (LOS) is a key indicator of the efficiency of inpatient management. Prolonged LOS (pLOS) is associated not only with high resource consumption and medical expenses [6,7] but also with a high risk of complications [8]. Much attention has been given to reducing hospitalization costs [9-15], but few studies have focused on preventing pLOS for PD-treated patients. The increasing number of PD-treated patients and their consistently high hospital admission rate have placed a large burden on the health care system. An accurate pLOS prediction model can assist physicians to risk-stratify patients and optimally allocate health care resources [7,16]. Early clinical interventions and optimal management of patients at a high risk of pLOS may help reduce hospitalization expenses and improve prognosis for PD-treated patients [7,8,17]. If timely clinical interventions are not provided, patients at a high risk of pLOS may face poor prognosis and high medical expenses, which will also burden hospitals [18].

Given the increasing number of patients undergoing dialysis and the importance of optimal resource allocation, physicians need an effective LOS prediction model. However, no well-developed LOS prediction models for patients undergoing dialysis can be found in the literature. Some other risk-stratification models for patients undergoing dialysis use mortality [19-21] or cardiovascular events [22] as the end point. Wagner et al [20] used a nationwide, multicenter, prospective cohort study in the United Kingdom (the UK Renal Registry) as a data source to develop a Cox proportional hazards model for predicting long-term mortality in incident dialysis patients. They found that using basic patient characteristics, comorbid conditions, and laboratory variables to predict the 3-year mortality of incident dialysis patients had sufficient accuracy. Quinn et al [21] used a Canadian administrative health database to develop a prognostic index for 1-year mortality in patients undergoing dialysis by combining logistic regression (LR) with different variable selection methods. Matsubara et al [22] used data from the Japan Dialysis Outcomes and Practice Patterns Study to develop an LR model for predicting the incidence of cardiovascular events among patients undergoing hemodialysis. However, few models use LOS as the prediction outcome.

Meanwhile, a number of studies have explored the factors affecting the LOS of patients undergoing dialysis. Allon et al [23] explored the association of hospitalization outcomes with clinical factors and laboratory parameters in patients undergoing hemodialysis and found that infection-related hospitalization was associated with pLOS. Kshirsagar et al [24] compared the LOS of hemodialysis patients receiving care from nephrologists and internists and found that the LOS was significantly shorter for patients under the care of nephrologists than for patients under the care of internists. Rocco et al [25] studied the risk factors for hospitalization in patients receiving chronic dialysis and confirmed that the risk factors for LOS were similar to those for mortality. Other factors affecting the LOS of patients undergoing dialysis have also been explored, such as obesity [26], hemoglobin level [27], admission diagnosis [28], and comorbidities [23,29]. However, no study has built an effective model for pLOS prediction in patients undergoing dialysis.

With the exponential increase in the amount of health care data, machine learning algorithms have gained special attention for their capabilities of handling high-dimension and large-scale data. Some machine learning–based LOS prediction models have been developed for patients with other diseases. The prediction outcome of existing LOS prediction models could be classified into 2 types: (1) numeric LOS and (2) binary outcome (ie, having a pLOS or not). Moran et al [30] constructed a numeric LOS prediction model for patients in the intensive care unit (ICU) by using a traditional linear regression model. Their results suggested that their LOS prediction model performed well in predicting the average LOS of patients in the ICU but showed limited performance in predicting the LOS of individual patients. Yang et al [31] developed a numeric LOS prediction model based on the support vector machine (SVM) algorithm for burn patients at different stages and compared its prediction performance with that of the traditional linear regression model. They found that although the SVM model was more effective than the linear regression model in LOS prediction for burn patients, it yielded a high mean relative error of 43.9%. LaFaro et al [32] developed a numeric LOS prediction model based on the artificial neural network (ANN) algorithm for patients in the ICU after cardiac surgery. Their results also suggested that the ANN-based LOS prediction model outperformed the traditional linear regression model (R^2: 0.410 vs 0.200; R^2 measures the goodness of fit of the corresponding model), but the prediction performance of the ANN-based model was still limited. However, if patients are classified into 2 groups (ie, with and without pLOS), the difference in LOS patterns between patients in the 2 groups could be more obvious and easily discovered, and this classification helps identify typical LOS patterns and improve the performance of LOS prediction models [33]. In the literature, the LOS prediction models with binary outcomes achieved good performance. Ma et al [34] developed a personalized pLOS prediction model for patients in the ICU by combining just-in-time learning and one-class extreme learning machine algorithms and found that the model achieved superior performance to the traditional binary classification algorithms. Chuang et al [35] compared the performance of various supervised learning approaches with an LR model in pLOS prediction for general surgery patients and the results showed that the random forest (RF) model outperformed the LR model. Morton et al [36] used 5 machine learning algorithms to predict the pLOS of hospitalized patients with diabetes and found that the SVM model demonstrated the best prediction performance, followed closely by the RF model. However, LOS prediction models based on machine learning technologies for PD-treated patients remain to be developed.

Stacked generalization, or stacking, is a general ensemble method that combines different types of machine learning models (“base models”) through an aggregation model (“meta-model”) to maximize the prediction performance [37]. Several studies [38,39] have found that ensemble learning methods can produce a better or equal predictive performance than their component parts. Lertampaiporn et al [38] developed a heterogeneous ensemble model for microRNA precursor classification through a voting system. Their results showed that the ensemble method produced a more reliable prediction than its base classifiers. Wang et al [39] used the stacking algorithm to predict membrane protein types, and the ensemble model yielded a better overall performance than its base models. Phan et al [40] developed a stacking model to predict cancer survival and reported that this model outperformed the majority-vote model. An ensemble of various machine learning models could help reduce the bias in a single machine learning algorithm to provide a much better prediction performance than single models.

This study aimed to develop an optimal data-driven pLOS prediction model for PD-treated patients by using basic admission data from a national database. A pLOS prediction model was constructed for PD-treated patients by using the stacking method, and the Hospital Quality Monitoring System (HQMS) database in China was used for model development. An LR-based pLOS prediction model was built and considered as the benchmark model. The RF, SVM, and K-nearest neighbor (KNN) algorithms were employed as the base models because of their superior performance in constructing ensemble models [38,41], and the LR model was used as the meta-model for constructing the stacking model.

Methods

Data Set and Subjects

In this study, the HQMS database—a mandatory, patient-level national database in China—was used for data extraction and model development. The HQMS database is a large database consisting of standardized electronic inpatient discharge records, including 878 Class 3 hospitals in China [42]. The standardized electronic inpatient discharge record is a national standard medical record with a stringent standard format across different hospitals in China. The standardized electronic inpatient discharge records of patients must be filled in by clinicians who have the most comprehensive understanding of the patients’ medical conditions to ensure their validity. Strict automated data quality control was performed on the HQMS data reporting system. The completeness, accuracy, and consistency of data were assessed at the time of data submission to the HQMS. Patient demographic characteristics, clinical diagnoses, medical procedures, pathology diagnoses, and medical expenditures were included in the HQMS database.

This study was reviewed and approved by the Ethics Committee of Peking University First Hospital (2015-928). The HQMS data set used in this study spans from 2013 to 2015.

Patient records of individuals who met the following criteria were extracted from the HQMS data set: (1) aged between 18 and 100 years, and (2) treated with PD. Exclusion criteria were as follows: (1) diagnosed with acute kidney injury or kidney transplantation, and (2) died in the hospital. For patients readmitted on the same day as hospital discharge, we recalculated their LOS by merging the back-to-back admission records. The PD-treated patients were identified through admission and discharge diagnoses or in-hospital medical operations by using the International Statistical Classification of Diseases, Tenth Revision (ICD-10) codes (Multimedia Appendix 1). For PD-treated patients with several discontinuous hospitalizations, we randomly selected one record for each patient to ensure that all observations were independent and that PD-treated patients with varying severities were included for model development.

Outcome and Predictor Variables

The prediction outcome of this study was binary (ie, having a pLOS or not). LOS was defined as the period from admission to discharge. pLOS was defined as an LOS longer than the average LOS, which is 16 days for patients with ESKD in China [43]. Patients with pLOS may have serious medical situations and thus need a longer hospital stay. We adopted this pLOS definition in our study by referring to existing studies [44-46] and consulting with experienced clinicians. The pLOS prediction models developed in our study aimed to assist physicians in identifying patients at a high pLOS risk and thus to provide early and timely interventions for these high-risk patients.

Predictor variables were determined on the basis of prior studies [23,24,28,29] and variable availability on admission. Variables used as predictor variables for model development in this study included age, sex, nationality, reason for admission, specific causes of chronic kidney disease (CKD), comorbidities, admission type, number of hospitalizations within 6 months, number of emergency admissions within 6 months, admission department, planned admission or not, admission day of the week, admitted in the same hospital as last admission or not, place of residence, and insurance type. The reason for admission, specific causes of CKD, and comorbidities were extracted using ICD-10 codes. The categories of reasons for admission and comorbidities were determined after consultation with experienced clinicians. Limited by the available data set, the number of hospitalizations within 6 months and number of emergency admissions within 6 months were calculated on the basis of the data collected from Class 3 hospitals.

Model Development

RF Model

RF is a supervised ensemble learning algorithm consisting of a collection of tree-structured classifiers [47]. RF models work by generating a multitude of decision trees independently and then synthesizing the individual predictions of all trees through a voting system. Each tree in an RF model is built using a bootstrap sample of the training data set. Assuming that *M* predictor variables are included for model development, *F* of all *M* input variables are randomly selected for each node, and the split of each node is performed according to the minimal impurity principle. For each tree, a variable that was used for tree growth in the previous nodes will no longer be used in later splitting. In decision tree induction, the Gini index is a general impurity measure used to determine the splitting variables. If a data set *D* contains samples with *J* classes, the Gini index of data set *D*—Gini(*D*)—is defined as follows [48]:

where p_j is the frequency of the *j*th class in *D*. At each node, if a variable can split the parent data set *D* into 2 child data sets, D_1 and D_2, the decrease in the Gini index, *S*, for this variable is defined by the following:

The variable with a maximal decrease in the Gini index will be used for splitting at this node.

In an RF model, to classify a new case, each tree in the forest model gives a classification result for the new case as a vote, and the majority vote is declared as the final classification of the model. Twice randomization in an RF model, which involves randomly selecting training data samples and randomly selecting the attributes for each tree growth, provides the model with a strong capability of handling high-dimensional data together with a stable generalization error [49].

We used the RandomForestClassifier package in Python to construct the RF model in this study. A set of optimal parameters of the RF model was found using grid search, which is an exhaustive searching method using a manually specified subset of hyperparameter space to find the optimal parameters of a learning algorithm [50]. The RF model obtained in this study had the following parameters: the number of decision trees was 300, the number of variables (*F*) selected at each node was 10, and the maximal depth of each decision tree was 28.

SVM Model

SVMs have been used frequently in various classification problems because of their remarkably robust performance in handling noisy and nonlinearly classified data [51]. If the data set is not linearly separable, a mapping function will be used in the SVM to map the data set into a high-dimensional space. An SVM tries to find an optimal separating hyperplane (ie, the maximum-margin hyperplane) in the high-dimensional space to make a classification. Assuming that a training data set, *D*, consists of *N* labeled cases, , where x_i represents the *i*th feature vector and y_i is the label of the *i*th case. A mapping function, ø (*x*), will map the data set from the original space into a high-dimensional space. In the transformed high-dimensional space, the separating hyperplane [52] is defined as follows:

where is a normal vector determining the direction, and *b* is the bias. The training cases with minimum margins from the hyperplane are called support vectors. A support vector (x_j, y_j) satisfies:

In the high-dimensional space, the margin *M* between the support vector and the hyperplane is defined as

The hyperplane that makes the margin *M* maximum is the optimal separating hyperplane (ie, maximum-margin hyperplane). In the process of finding the optimal separating hyperplane, a kernel function is usually used to deal with the high computational cost. Commonly used kernel functions include the polynomial kernel, the linear kernel, the exponential kernel, and the radial basis function kernel.

We used the svm package in Python to construct the SVM model, and the optimal parameters of our SVM model were found using grid search. The SVM model obtained in this study had the following parameters: the kernel function was polynomial kernel, the degree of the polynomial kernel function was 2, and the penalty parameter C was 0.01.

KNN Model

KNN is a type of instance-based learning method that makes predictions based on a small number of cases that are very similar to the target observation [53]. Specifically, given a new case (x_{new}), we can find the K closest training cases, ☒ sorted by the distance to x_{new}, and then classify x_{new} using majority voting among the K neighbors. A commonly used distance metric in the KNN algorithm is the Euclidean distance. Assuming the presence of case ☒ and ☒, we define the Euclidean distance from x_i to x_j as

☒

where ☒ and ☒ denote the values of *M* input predictor variables of the 2 cases. Typically, we first normalize all the values of the variables to the range of (0,1) because different variables could be measured in different units. The KNN algorithm yields convincing results in handling various classification problems in medicine [54-56]. The model is effective on data sets where samples of 1 class have many possible patterns and the decision boundary is nonlinear [57]. The most important parameter in the KNN model is the number of neighbors, which must be selected with care. In this study, we used the KNeighborsClassifier package in Python to construct the KNN model. The optimal parameter *K* was found using grid search, and the KNN model with optimal performance was obtained with the parameter *K*=130.

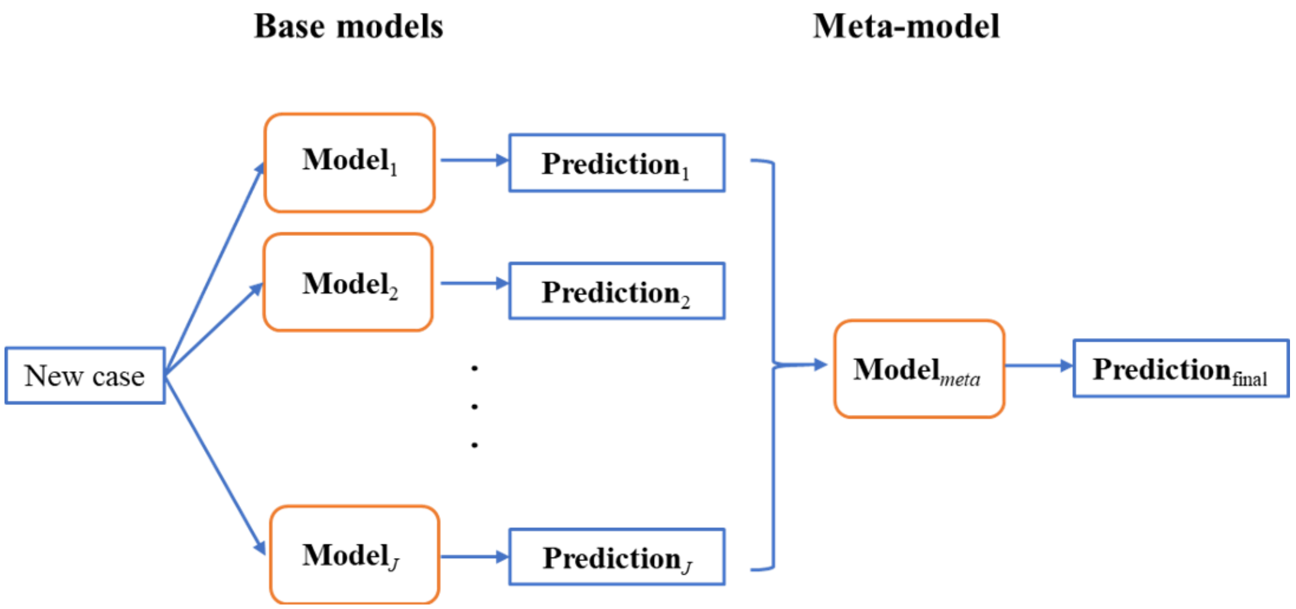
Stacked Generalization

Stacked generalization, or stacking, is an ensemble model that can combine the predictions of several primary machine learning models [37]. There are 2 types of models in a stacking framework: several base models (level-0 models) and 1 meta-model (level-1 model). The meta-model is employed to combine the base models. In general, a stacking framework can obtain a more accurate prediction result than any single base model. Different models may complement each other, and the meta-algorithm can combine the advantages of these base models.

The stacking model is trained as follows. Given a data set ☒ we define D_k and $D_{-k} = D - D_k$ as the training and test data sets, respectively, in the *k*th round of model training. We assume that the stacking model has *J* base models ($Model_1$, $Model_2$, ... , $Model_j$, ... $Model_J$) and that each base model is trained using D_k. Let ☒ denote the prediction outcome produced by $Model_j$ for training case (x_i, y_i). The outputs of all *J* base models are assembled as the input of the meta-model. Let ☒ denote the set of outputs produced by all of the *J* base models for (x_i, y_i). The meta-model is then trained using data set ☒.

For a new input case, the output of the meta-model is the final prediction outcome produced by the stacking model for the case. How the base models are assembled in the stacking method and how the prediction outcome for a new input case is generated by the stacking model are shown in Figure 1.

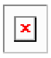
Figure 1. Stacked generalization, where Predictionj denotes the prediction outcome produced by the model (Modelj) for a new case.



Given that the level-0 base models have already completed most of the prediction work, the level-1 meta-model could be rather simple [58]. The LR model is commonly used as the meta-model. Existing studies [37,59] suggested that increasing diversity of the base models could help improve the performance of the stacking model. In this study, the RF, SVM, and KNN models were employed as the base models and the LR model was used as the meta-model.

Statistical Analysis

Two-sided *t* tests and chi-square tests were used for comparisons of patient demographics. In model development and comparisons, we employed 5-fold cross-validation. In performance comparisons, the Brier score [60], area under the receiver operating characteristic curve (AUROC) [60], estimated calibration index (ECI) [61], accuracy, sensitivity, specificity, and geometric mean (Gm) [62] were employed as performance measures. Considering that other performance metrics, such as positive and negative predictive values and likelihood ratios, can be calculated from sensitivity and specificity, we did not employ them in performance comparisons. Brier score is an overall performance measure, with a lower Brier score suggesting a superior overall prediction performance. AUROC measures the discrimination power of a prediction model, representing the ability to distinguish positive samples from negative samples. ECI measures the calibration power of a model, representing the average difference between the predicted probabilities of individual patients and the observed probability in that patient population. ECI ranges between 0 and 100, with a lower ECI suggesting a stronger calibration power of the corresponding model. Gm is considered a balanced accuracy measure because it incorporates sensitivity and specificity, and it is defined as follows:



Gm measures the balance of the classification performance for the majority and minority classes. The optimal cutoff value for each model was obtained according to its corresponding receiver operating characteristic curve, and then accuracy, sensitivity, specificity, and Gm were calculated. Performance differences between different models were assessed using 2-sided *t* tests. Furthermore, we used the calibration plot [60] to demonstrate the calibration power of each model in different patient groups with pLOS risk from low to high. In the calibration plot, patients were divided into 10 groups according to their predicted pLOS probabilities. The x-axis shows the observed pLOS probability of each patient group, and the y-axis shows the averaged predicted pLOS probability of each group. The ideal calibration curve for a perfect model is a diagonal, which suggests that the predicted probabilities are exactly consistent with the observed probabilities.

Statistical analysis and calculations were performed using Python 3. Less than 15% of records in the HQMS database had missing values for the nationality and admission type variables, and the missing values were considered as a special category in the analysis.

Results

A total of 23,992 eligible patients receiving PD were included in our study, of whom 30.3% had a pLOS. Characteristics of the PD-treated patients are displayed in Table 1. The proportion of male patients was 55.6% (13,351/23,992), and the average age of all patients was 52.1 (SD 15.0) years. The 2-sided *t* tests showed that the differences in age, place of residence, and insurance type between PD-treated patients with a pLOS and those without a pLOS were statistically significant. The histogram of the LOS distribution of the PD-treated patients is displayed in Figure 2.

Table 1. Characteristics of peritoneal dialysis–treated patients in the study.

Characteristic	All patients	Patients with pLOS[a]	Patients without pLOS	*P* value
Number of patients (%)	23,992 (100)	7270 (30.3)	16,722 (69.7)	
Age (years), mean (SD)	52.1 (15.0)	53.6 (15.4)	51.5 (14.8)	<.001
Sex, n (%)				.63
Female	10,641 (44.4)	3242 (44.6)	7399 (44.2)	
Male	13,351 (55.6)	4028 (55.4)	9323 (55.8)	
Place of residence, n (%)				<.001
East China	9425 (39.3)	2565 (35.3)	6860 (41.0)	
North China	2318 (9.7)	902 (12.4)	1416 (8.5)	
Central China	3416 (14.2)	1157 (15.9)	2259 (13.5)	
South China	3978 (16.6)	1261 (17.3)	2717 (16.2)	
Southwest China	2933 (12.2)	849 (11.7)	2084 (12.5)	
Northwest China	1067 (4.4)	225 (3.1)	842 (5.0)	
Northeast China	855 (3.6)	311 (4.3)	544 (3.3)	
Insurance, n (%)				.005
UEBMI[b]	9100 (37.9)	2714 (37.3)	6386 (38.2)	
URBMI[c]	2192 (9.1)	705 (9.7)	1487 (8.9)	
NRCMS[d]	6082 (25.4)	1931 (26.6)	4151 (24.8)	
Free medical care	334 (1.4)	101 (1.4)	233 (1.4)	
Self-paid treatment	3493 (14.6)	997 (13.7)	2496 (14.9)	
Other	2791 (11.6)	822 (11.3)	1969 (11.8)	

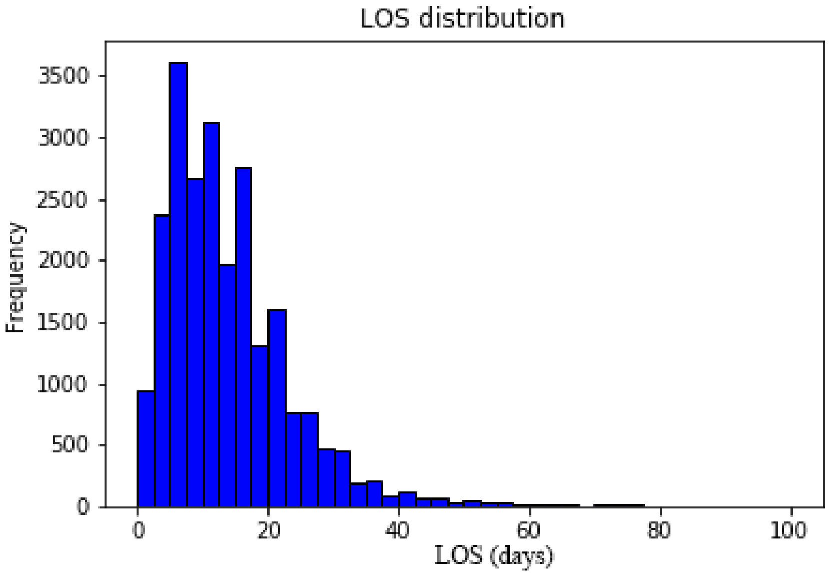
[a] pLOS: prolonged length of stay.

[b] UEBMI: urban employee basic medical insurance.

[c] URBMI: urban resident basic medical insurance.

[d] NRCMS: new rural cooperative medical system.

Figure 2. Histogram of length of stay (LOS) distribution of peritoneal dialysis–treated patients.



A comparison of the prediction performance of the stacking model, its 3 base models, and the benchmark LR model in terms of the Brier score, AUROC, ECI, Gm, accuracy, sensitivity, and specificity is shown in Table 2. Among these models, the stacking model achieved the best calibration (ECI 8.691), balanced accuracy (Gm 0.690), accuracy (0.695), and specificity (0.701). Meanwhile, the stacking and RF models had the best overall performance (Brier score 0.174 for both) and discrimination (AUROC 0.757 for the stacking model and 0.756 for the RF model). Compared with the benchmark LR model, the stacking model was superior in all performance measures except sensitivity, but there was no significant difference in sensitivity between the 2 models. The 2-sided *t* tests revealed significant performance differences between the stacking and LR models in overall performance, discrimination, calibration, balanced accuracy, and accuracy.

Table 2. Prediction performance of the 5 models.

Model	Brier score	AUROC[a] (95% CI)	ECI[b]	Gm[c]	Accuracy	Sensitivity	Specificity
LR[d]	0.178	0.742 (0.731-0.753)	8.911	0.677	0.675	0.683	0.671
KNN[e]	0.188*	0.721 (0.703-0.740)*	9.386*	0.661*	0.666	0.666	0.657
SVM[f]	0.187*	0.730 (0.720-0.739)*	9.342*	0.673	0.680	0.656	0.690
RF[g]	0.174*	0.756 (0.748-0.765)*	8.722*	0.689*	0.691*	0.686	0.693
Stacking	0.174*	0.757 (0.748-0.765)*	8.691*	0.690*	0.695*	0.680	0.701

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

[b] ECI: estimated calibration index.

[c] Gm: geometric mean.

[d] LR: logistic regression.

[e] KNN: K-nearest neighbor.

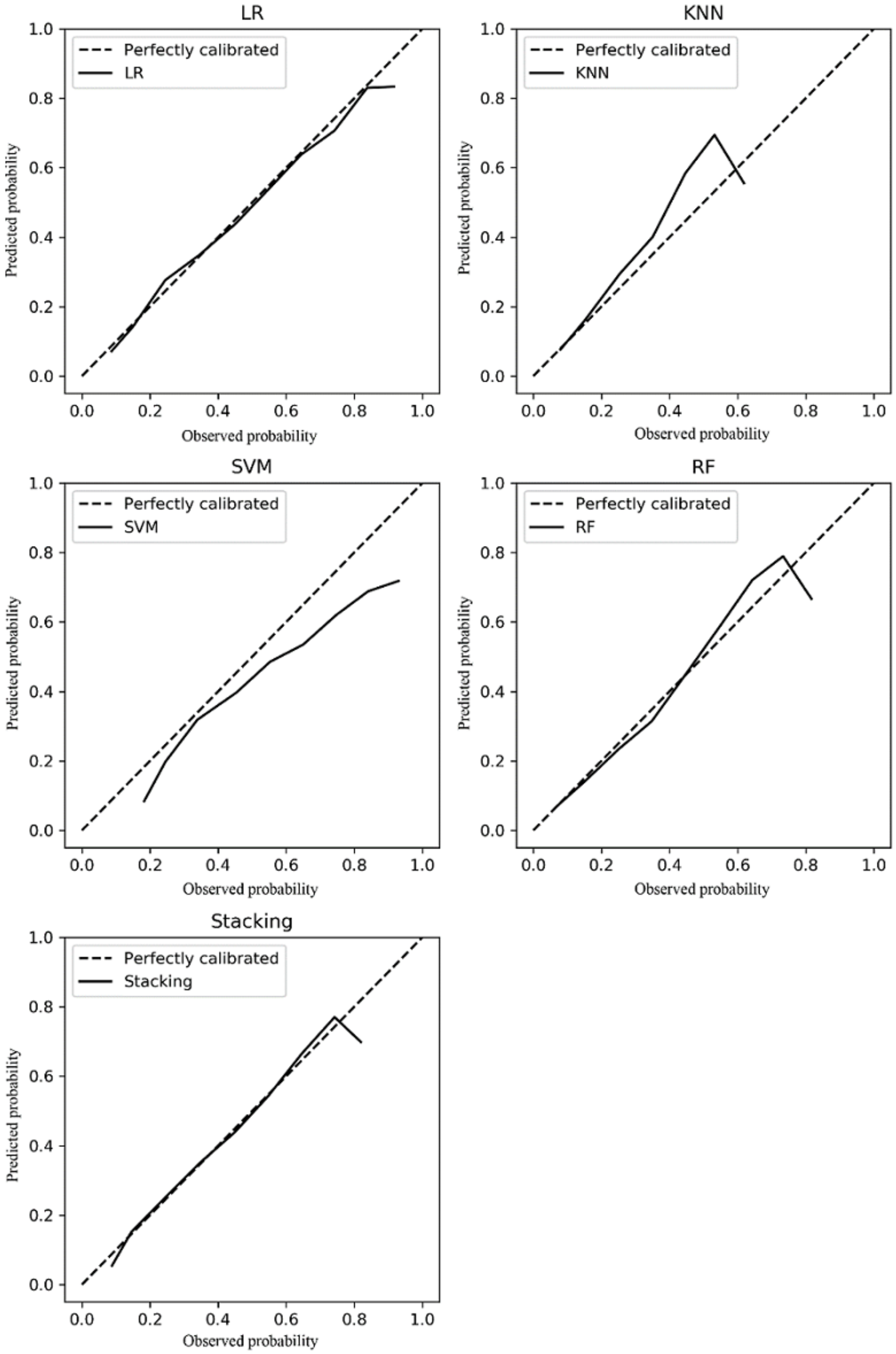
[f] SVM: support vector machine.

[g] RF: random forest.

**P*<.05 in 2-sided *t* test when compared with the LR model.

Figure 3 demonstrates the calibration plots of the 5 models. The calibration curve of the stacking model was the optimal fitting curve among the 5 models. The SVM model underestimated the pLOS probabilities for most patients, whereas the KNN model overestimated the pLOS probabilities for most patients. The RF model underestimated the pLOS probabilities for most patients at low risk and overestimated the probabilities for most patients at high risk.

Figure 3. Calibration plots of the 5 models. KNN: K-nearest neighbor; LR: logistic regression; RF: random forest; SVM: support vector machine.



Discussion

Principal Findings

The main objective of this study was to develop an optimal data-driven model for predicting the pLOS risk of PD-treated patients using basic admission data. To the best of our knowledge, this study is the first to develop such pLOS prediction models for PD-treated patients by using data from a national database. Our study constructed a pLOS prediction model for PD-treated patients based on a stacking method with KNN, SVM, and RF as its base models and LR as its meta-model. The prediction performance of the stacking model was compared with those of a benchmark LR model and its 3 base models. A pragmatic pLOS prediction model for PD-treated patients would be useful in family consultation and has the potential to assist physicians in making optimal clinical decisions. Considering that medical expenses are highly associated with LOS [6,7], the pLOS prediction model could help estimate the medical expenses for PD-treated patients. The degree of satisfaction may increase if patients and their families know more about their LOS and medical expenses on hospital admission. In addition, the pLOS prediction models could be integrated into hospital information systems, providing physicians with real-time suggestions about the LOS of patients and helping physicians to identify PD-treated patients at a high risk of pLOS and give timely individualized intervention.

In this study, the RF, SVM, and KNN models were employed as base models for stacking because they have different learning mechanisms and have advantages in different aspects. RF is an ensemble learning algorithm consisting of a collection of tree-structured classifiers. The twice randomization in an RF model provides the model with a strong capability of handling high-dimensional data together with a stable generalizability [49]. However, RF models are sensitive to noise data. SVM models make classifications by mapping data into a high-dimensional space and finding an optimal separating hyperplane in the high-dimensional space. SVM models show remarkably robust performance in handling noisy and nonlinearly classified data but have limitations in handling high-dimensional data [51]. KNN is an instance-based learning method that makes predictions depending on a small number of cases that are strongly similar to the target observation. KNNs are effective on nonlinearly separable data sets and data sets where samples of one class have different patterns [57]. KNNs are insensitive to noise data but have limited accuracy in unbalanced data. In addition, an existing study [38] showed that the ensemble of the 3 models demonstrated superior prediction performance in dealing with classification problems. Moreover, the literature states that the 3 classifiers are suitable for pLOS prediction problems. All 3 classifiers have shown superior performance in predicting pLOS for patients. Chuang et al [35] employed the SVM and RF models for pLOS prediction in patients who underwent general surgery, and both models achieved a high AUROC. Steele and Thompson [63] developed a KNN-based pLOS prediction model for general patients and achieved an AUROC of 0.847. KNN was included as the base model in our study because it has shown superior performance in pLOS prediction in existing studies [63,64]. Given that its learning mechanism is different from the learning mechanisms of the 2 other base models (SVM and RF), KNN was expected to improve the prediction performance of the stacking model in dealing with data sets with various characteristics [37,59]. We also attempted to construct stacking models with combinations of any 2 base models of RF, SVM, and KNN. We found that the stacking model with SVM and KNN as its base models had the worst performance, while the stacking model with 3 base models and the stacking models with the other 2 combinations (SVM and RF, and KNN and RF) had similar overall performances. Considering the diversity and respective advantages of the base models, and the generalizability of the stacking model in dealing with data sets with different characteristics, we selected the stacking model with 3 base models.

The performance comparison results showed that the stacking model was the best among the 5 models in terms of overall performance (Brier score), discrimination (AUROC), calibration (ECI), balanced accuracy (Gm), accuracy, and specificity. The RF model showed the best prediction performance among the 3 base models, and it had a similar overall performance and discrimination power as the stacking model. The good prediction performance of the stacking and RF models may be due to the fact that both models are ensemble learning models. Our study results are consistent with previous studies showing that the ensemble model is almost always superior to single learning models [38,39]. A stacking model can exploit its base models by combining the output of each model via a meta-model, thus reducing the bias that tends to occur with a single classifier. An RF model can exploit its base tree models by combining the output of each model via a voting system. The stacking model was slightly superior to the RF model in most performance measures for 2 possible reasons. First, the prediction performance of a stacking model is usually similar to its best base model [40,41]. Second, compared with an RF model, a stacking model has more diverse base models that can complement each other.

The calibration curves of the 5 models further suggest that the stacking model had the optimal calibration power in different patient groups. ECI measures the overall calibration power of a model, whereas the calibration curve visually shows the calibration power of a model in patient groups with pLOS risk from low to high. The ECI and calibration curve demonstrated that the stacking model had superior calibration power. The calibration curve showed that the averaged predicted pLOS probability of the stacking model had high consistency with the observed outcome across different pLOS risk groups. Meanwhile, the calibration curve showed that the RF model underestimated the pLOS probabilities of most patients at low risk and overestimated the probabilities of most patients at high risk. This feature can help the RF model expand the difference of predicted probabilities between patients with different pLOS risks and thus discriminate the patients at a high pLOS risk from those at a low risk. This probably explained why the RF model showed similar discrimination but worse calibration power than the stacking model.

We also attempted to develop numeric LOS prediction models for PD-treated patients, but the corresponding prediction performance of the models was limited, which was similar to that of existing numeric LOS prediction models. Numeric LOS prediction models focused on mining different LOS patterns for patients with different LOSs (even 1 day apart), but the difference in LOS patterns between patients with different LOSs, especially those LOSs with 1 or 2 days apart, may be slight and was difficult to identify. The pLOS prediction models with binary outcomes had a much better performance.

Regarding data exclusion, the PD-treated patients who died in the hospital were excluded in our study because the LOS pattern of the decedents might be different from that of patients who survived in the hospital [65,66]. Based on our consultations with experienced clinicians, we knew that there was uncertainty in the LOS pattern of patients who died in the hospital. Specifically, deceased patients could die quickly after hospital admission and have a short LOS or die after a long period of treatment and have a long LOS. In fact, the proportion of

PD-treated patients who died in the hospital was only 0.8% in our study. Selection bias might have occurred when we excluded those PD-treated patients who died in the hospital, and the pLOS prediction model developed in our study may not apply to those patients who have a high risk of in-hospital mortality.

In our study, some PD-treated patients were hospitalized more than once; they can be classified into 2 types: (1) patients readmitted on the same day as discharge, and (2) patients with several discontinuous hospitalizations. Some hospitals in China may discharge patients with a potential pLOS first and then readmit them on the same day to reduce the average LOS, which is an important indicator in hospital evaluation. Therefore, for the PD-treated patients readmitted on the same day as discharge, we recalculated their actual LOS by merging the back-to-back admission records in this study. To deal with the situation of PD-treated patients with several discontinuous hospitalizations, we examined 2 approaches that were employed in the literature: (1) selecting the first hospitalization record, or (2) randomly selecting 1 record among multiple hospitalization records. Compared with the former approach, the latter approach may help include patients with varying severities [67]. Thus, we employed the second approach and randomly selected 1 record for each patient to ensure that all observations were independent and PD-treated patients with varying severities were included in model development.

Definition of pLOS

In this study, pLOS was defined as an LOS longer than the average LOS by referring to existing studies [44-46] and consulting with experienced clinicians. In the literature, there is no consensus on the definition of pLOS for general patients or PD-treated patients. Existing studies have defined pLOS as an LOS longer than the average LOS [44-46], longer than the median LOS [68], or longer than a specific LOS according to experiences [69]. After consulting with experienced clinicians, we know that the average LOS is a more important metric for PD-treated patients, and it is also a more commonly used metric in assessing medical efficiency around the world. In addition, pLOS has been defined as an LOS longer than the average LOS in various medical fields by researchers from different countries [44-46]. Among the 3 cited references that defined pLOS as an LOS longer than the average LOS, one study [44] was of trauma patients in the United States, another study [45] was of critically ill patients in Switzerland, and the third study [46] was of surgery patients in China. Therefore, the definition of pLOS as longer than the average LOS may help our models achieve good generalizability to some extent.

Diagnosis Codes

The use of diagnosis codes to identify patients with specific diseases may miss some target patients because clinicians tend to focus on the main diagnosis related to admission reasons and overlook the diagnosis of other diseases. To address this problem, we employed ICD-10 codes associated with all admission and discharge diagnoses and in-hospital medical operations to identify PD-treated patients. We also used ICD-10 codes associated with admission and discharge diagnoses to identify patients' comorbidities.

Strengths and Limitations of the Study

This study has several strengths. First, a large nationwide database with a relatively representative population was used to derive the prediction models. Second, all of the predictor variables are available at admission, which ensures the feasibility of applying the developed models in clinical practice to assist clinical decision making. Third, 5-fold cross-validation was employed to achieve reliable performance results.

However, this study has some limitations. First, the models were derived from a nationwide data set in China. Some of the variables included in the models, such as nationality and insurance type, are region specific. The generalizability and validity of our prediction models need to be validated using a data set from different regions. Second, other potentially important variables, such as some laboratory markers, that reportedly affect LOS [27,70] were not available in the studied data set. Third, only patient data from Class 3 hospitals were included in the studied data set. Class 3 hospitals in China provide the best medical services for patients, and patients admitted to Class 3 hospitals in China may be suffering from serious diseases. Thus, our pLOS prediction models may not be applicable to the PD-treated patients in the primary or Class 2 hospitals in China, considering that patients admitted to those hospitals may have only minor or moderate diseases.

Conclusion

This study was the first to develop data-driven automated pLOS prediction models for PD-treated patients using basic admission data from a national database. The results of our study indicate the feasibility of utilizing a stacking-based model for PD-treated patients. The developed pLOS prediction models have the potential to help clinicians identify PD-treated patients at a high risk of pLOS and then provide optimal patient management. The pLOS prediction tools developed in this study have the potential to assist clinicians in identifying patients at a high risk of pLOS and to allocate resources optimally for PD-treated patients. The generalizability and validity of the developed pLOS prediction models need to be externally validated, and the clinical utility of the models needs further validation before they are used in clinical practice. The pLOS prediction models developed in our study are purely theoretical so far, and we plan to integrate them into the information system of a pilot hospital for prospective validation.

Acknowledgments

The authors thank the Bureau of Medical Administration and Medical Service Supervision, National Health Commission of China for the support of this study.

This study was supported by grants from the National Natural Science Foundation of China (Grant Nos. 81771938, 91846101, 82003529), Beijing Municipal Science & Technology Commission (Grant No. 7212201), Beijing Nova Programme Interdisciplinary Cooperation Project (Z191100001119008), Chinese Scientific and Technical Innovation Project 2030 (2018AAA0102100), the University of Michigan Health System-Peking University Health Science Center Joint Institute for Translational and Clinical Research (BMU2020JI011, BMU2018JI012, BMU2019JI005), and PKU-Baidu Fund (2019BD017).

Authors' Contributions

Research idea and study design: GK, JW, LZ; data acquisition and preprocessing: YS, HW; data analysis and statistical analysis: JW, YL, CY, KL; the methodology for extracting patients and identifying disease: HC; supervision or mentorship: GK, LZ; manuscript writing: GK, JW, LZ. Each author contributed important intellectual content during manuscript drafting and accepts accountability for the overall work by ensuring that questions pertaining to the accuracy or integrity of any portion of the work are appropriately investigated and resolved.

Conflicts of Interest

None declared.

Multimedia Appendix 1
International Statistical Classification of Diseases, Tenth Revision (ICD-10) codes for identifying peritoneal dialysis patients.
[DOCX File , 16 KB - medinform_v9i5e17886_app1.docx]

References

1. Saran R, Robinson B, Abbott KC, Agodoa LYC, Bragg-Gresham J, Balkrishnan R, et al. US Renal Data System 2018 Annual Data Report: Epidemiology of Kidney Disease in the United States. Am J Kidney Dis 2019 Mar;73(3 Suppl 1):A7-A8 [FREE Full text] [doi: 10.1053/j.ajkd.2019.01.001] [Medline: 30798791]
2. Liyanage T, Ninomiya T, Jha V, Neal B, Patrice HM, Okpechi I, et al. Worldwide access to treatment for end-stage kidney disease: a systematic review. Lancet 2015 May 16;385(9981):1975-1982. [doi: 10.1016/S0140-6736(14)61601-9] [Medline: 25777665]
3. Zhang L, Zuo L. Current burden of end-stage kidney disease and its future trend in China. Clin Nephrol 2016;86 (2016)(13):27-28. [doi: 10.5414/CNP86S104] [Medline: 27469147]
4. Zhang L, Zhao M, Zuo L, Wang Y, Yu F, Zhang H, CK-NET Work Group. China Kidney Disease Network (CK-NET) 2015 Annual Data Report. Kidney Int Suppl (2011) 2019 Mar;9(1):e1-e81 [FREE Full text] [doi: 10.1016/j.kisu.2018.11.001] [Medline: 30828481]
5. Chan KE, Lazarus JM, Wingard RL, Hakim RM. Association between repeat hospitalization and early intervention in dialysis patients following hospital discharge. Kidney Int 2009 Aug;76(3):331-341 [FREE Full text] [doi: 10.1038/ki.2009.199] [Medline: 19516243]
6. Higgins TL, McGee WT, Steingrub JS, Rapoport J, Lemeshow S, Teres D. Early indicators of prolonged intensive care unit stay: impact of illness severity, physician staffing, and pre-intensive care unit length of stay. Crit Care Med 2003 Jan;31(1):45-51. [doi: 10.1097/00003246-200301000-00007] [Medline: 12544992]
7. Lorentz CA, Leung AK, DeRosa AB, Perez SD, Johnson TV, Sweeney JF, et al. Predicting Length of Stay Following Radical Nephrectomy Using the National Surgical Quality Improvement Program Database. J Urol 2015 Oct;194(4):923-928. [doi: 10.1016/j.juro.2015.04.112] [Medline: 25986510]
8. Yu T, He Z, Zhou Q, Ma J, Wei L. Analysis of the factors influencing lung cancer hospitalization expenses using data mining. Thorac Cancer 2015 May;6(3):338-345 [FREE Full text] [doi: 10.1111/1759-7714.12147] [Medline: 26273381]
9. Lu M, Sajobi T, Lucyk K, Lorenzetti D, Quan H. Systematic review of risk adjustment models of hospital length of stay (LOS). Med Care 2015 Apr;53(4):355-365. [doi: 10.1097/MLR.0000000000000317] [Medline: 25769056]
10. Arora P, Kausz AT, Obrador GT, Ruthazer R, Khan S, Jenuleson CS, et al. Hospital utilization among chronic dialysis patients. J Am Soc Nephrol 2000 Apr;11(4):740-746 [FREE Full text] [Medline: 10752533]
11. Mathew AT, Strippoli GF, Ruospo M, Fishbane S. Reducing hospital readmissions in patients with end-stage kidney disease. Kidney Int 2015 Dec;88(6):1250-1260 [FREE Full text] [doi: 10.1038/ki.2015.307] [Medline: 26466320]
12. Bruns FJ, Seddon P, Saul M, Zeidel ML. The cost of caring for end-stage kidney disease patients: an analysis based on hospital financial transaction records. J Am Soc Nephrol 1998 May;9(5):884-890 [FREE Full text] [Medline: 9596087]
13. Chazan JA, London MR, Pono L. The impact of diagnosis-related groups on the cost of hospitalization for end-stage renal disease patients at Rhode Island Hospital from 1987 to 1990. Am J Kidney Dis 1992 Jun;19(6):523-525. [doi: 10.1016/s0272-6386(12)80829-8] [Medline: 1595699]
14. Goldstein SL, Smith CM, Currier H. Noninvasive interventions to decrease hospitalization and associated costs for pediatric patients receiving hemodialysis. J Am Soc Nephrol 2003 Aug;14(8):2127-2131 [FREE Full text] [doi: 10.1097/01.asn.0000076077.05508.7e] [Medline: 12874467]

15. Li B, Cairns J, Fotheringham J, Ravanan R, ATTOM Study Group. Predicting hospital costs for patients receiving renal replacement therapy to inform an economic evaluation. Eur J Health Econ 2016 Jul;17(6):659-668. [doi: 10.1007/s10198-015-0705-x] [Medline: 26153418]
16. Menéndez R, Cremades M, Martínez-Moragón E, Soler J, Reyes S, Perpiñá M. Duration of length of stay in pneumonia: influence of clinical factors and hospital type. Eur Respir J 2003 Oct;22(4):643-648 [FREE Full text] [doi: 10.1183/09031936.03.00026103] [Medline: 14582918]
17. Abdelaziz TS, Fouda R, Hussin WM, Elyamny MS, Abdelhamid YM. Preventing acute kidney injury and improving outcome in critically ill patients utilizing risk prediction score (PRAIOC-RISKS) study. A prospective controlled trial of AKI prevention. J Nephrol 2020 Apr;33(2):325-334. [doi: 10.1007/s40620-019-00671-6] [Medline: 31712987]
18. Hachesu PR, Ahmadi M, Alizadeh S, Sadoughi F. Use of data mining techniques to determine and predict length of stay of cardiac patients. Healthc Inform Res 2013 Jun;19(2):121-129 [FREE Full text] [doi: 10.4258/hir.2013.19.2.121] [Medline: 23882417]
19. Douma CE, Redekop WK, van der Meulen JH, van Olden RW, Haeck J, Struijk DG, et al. Predicting mortality in intensive care patients with acute renal failure treated with dialysis. J Am Soc Nephrol 1997 Jan;8(1):111-117 [FREE Full text] [Medline: 9013455]
20. Wagner M, Ansell D, Kent DM, Griffith JL, Naimark D, Wanner C, et al. Predicting mortality in incident dialysis patients: an analysis of the United Kingdom Renal Registry. Am J Kidney Dis 2011 Jun;57(6):894-902 [FREE Full text] [doi: 10.1053/j.ajkd.2010.12.023] [Medline: 21489668]
21. Quinn R, Laupacis A, Hux J, Oliver M, Austin PC. Predicting the risk of 1-year mortality in incident dialysis patients: accounting for case-mix severity in studies using administrative data. Med Care 2011 Mar;49(3):257-266. [doi: 10.1097/MLR.0b013e318202aa0b] [Medline: 21301370]
22. Matsubara Y, Kimachi M, Fukuma S, Onishi Y, Fukuhara S. Development of a new risk model for predicting cardiovascular events among hemodialysis patients: Population-based hemodialysis patients from the Japan Dialysis Outcome and Practice Patterns Study (J-DOPPS). PLoS One 2017;12(3):e0173468 [FREE Full text] [doi: 10.1371/journal.pone.0173468] [Medline: 28273175]
23. Allon M, Radeva M, Bailey J, Beddhu S, Butterly D, Coyne D, HEMO Study Group. The spectrum of infection-related morbidity in hospitalized haemodialysis patients. Nephrol Dial Transplant 2005 Jun;20(6):1180-1186. [doi: 10.1093/ndt/gfh729] [Medline: 15769823]
24. Kshirsagar AV, Hogan SL, Mandelkehr L, Falk RJ. Length of stay and costs for hospitalized hemodialysis patients: nephrologists versus internists. J Am Soc Nephrol 2000 Aug;11(8):1526-1533 [FREE Full text] [Medline: 10906167]
25. Rocco MV, Soucie JM, Reboussin DM, McClellan WM. Risk factors for hospital utilization in chronic dialysis patients. Southeastern Kidney Council (Network 6). J Am Soc Nephrol 1996 Jun;7(6):889-896 [FREE Full text] [Medline: 8793798]
26. Fleischmann E, Teal N, Dudley J, May W, Bower JD, Salahudeen AK. Influence of excess weight on mortality and hospital stay in 1346 hemodialysis patients. Kidney Int 1999 Apr;55(4):1560-1567 [FREE Full text] [doi: 10.1046/j.1523-1755.1999.00389.x] [Medline: 10201023]
27. Ofsthun N, Labrecque J, Lacson E, Keen M, Lazarus JM. The effects of higher hemoglobin levels on mortality and hospitalization in hemodialysis patients. Kidney Int 2003 May;63(5):1908-1914 [FREE Full text] [doi: 10.1046/j.1523-1755.2003.00937.x] [Medline: 12675871]
28. Matas AJ, Gillingham KJ, Elick BA, Dunn DL, Gruessner RW, Payne WD, et al. Risk factors for prolonged hospitalization after kidney transplants. Clin Transplant 1997 Aug;11(4):259-264. [Medline: 9267712]
29. Viglino G, Cancarini G, Catizone L, Cocchi R, De Vecchi A, Lupo A, et al. Ten years experience of CAPD in diabetics: comparison of results with non-diabetics. Italian Cooperative Peritoneal Dialysis Study Group. Nephrol Dial Transplant 1994;9(10):1443-1448. [Medline: 7816258]
30. Moran JL, Solomon PJ, ANZICS Centre for OutcomeResource Evaluation (CORE) of the Australian New Zealand Intensive Care Society (ANZICS). A review of statistical estimators for risk-adjusted length of stay: analysis of the Australian and new Zealand Intensive Care Adult Patient Data-Base, 2008-2009. BMC Med Res Methodol 2012 May 16;12:68 [FREE Full text] [doi: 10.1186/1471-2288-12-68] [Medline: 22591115]
31. Yang C, Wei C, Yuan C, Schoung J. Predicting the length of hospital stay of burn patients: Comparisons of prediction accuracy among different clinical stages. Decis Support Syst 2010 Dec;50(1):325-335. [doi: 10.1016/j.dss.2010.09.001]
32. LaFaro RJ, Pothula S, Kubal KP, Inchiosa ME, Pothula VM, Yuan SC, et al. Neural Network Prediction of ICU Length of Stay Following Cardiac Surgery Based on Pre-Incision Variables. PLoS One 2015;10(12):e0145395 [FREE Full text] [doi: 10.1371/journal.pone.0145395] [Medline: 26710254]
33. Wolff J, McCrone P, Patel A, Kaier K, Normann C. Predictors of length of stay in psychiatry: analyses of electronic medical records. BMC Psychiatry 2015 Oct 07;15:238 [FREE Full text] [doi: 10.1186/s12888-015-0623-6] [Medline: 26446584]
34. Ma X, Si Y, Wang Z, Wang Y. Length of stay prediction for ICU patients using individualized single classification algorithm. Comput Methods Programs Biomed 2020 Apr;186:105224. [doi: 10.1016/j.cmpb.2019.105224] [Medline: 31765937]
35. Chuang M, Hu Y, Lo C. Predicting the prolonged length of stay of general surgery patients: a supervised learning approach. Int Tran Oper Res 2016 May 30;25(1):75-90. [doi: 10.1111/itor.12298]

36. Morton A, Marzban E, Giannoulis G, Patel A, Aparasu R, Kakadiaris IA. A comparison of supervised machine learning techniques for predicting short-term in-hospital length of stay among diabetic patients. 2014 Presented at: International Conference on Machine Learning and Applications; Dec 3-6, 2014; Detroit, MI, USA p. 428-431. [doi: 10.1109/ICMLA.2014.76]
37. Wolpert DH. Stacked generalization. Neural Networks 1992 Jan;5(2):241-259. [doi: 10.1016/s0893-6080(05)80023-1]
38. Lertampaiporn S, Thammarongtham C, Nukoolkit C, Kaewkamnerdpong B, Ruengjitchatchawalya M. Heterogeneous ensemble approach with discriminative features and modified-SMOTEbagging for pre-miRNA classification. Nucleic Acids Res 2013 Jan 07;41(1):e21. [doi: 10.1093/nar/gks878] [Medline: 23012261]
39. Wang S, Yang J, Chou K. Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition. J Theor Biol 2006 Oct 21;242(4):941-946. [doi: 10.1016/j.jtbi.2006.05.006] [Medline: 16806277]
40. Phan J, Hoffman R, Kothari S, Wu P, Wang MD. Integration of multi-modal biomedical data to predict cancer grade and patient survival. In: IEEE EMBS Int Conf Biomed Health Inform.: IEEE; 2016 Feb Presented at: 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI); 24-27 Feb. 2016; Las Vegas, NV, USA p. 577-580 URL: http://europepmc.org/abstract/MED/27493999 [doi: 10.1109/BHI.2016.7455963]
41. Pourhoseingholi M, Kheirian S, Zali M. Comparison of basic and ensemble data mining methods in predicting 5-year survival of colorectal cancer patients. Acta Inform Med 2017 Dec;25(4):254-258. [doi: 10.5455/aim.2017.25.254-258] [Medline: 29284916]
42. Zhang L, Wang H, Long J, Shi Y, Bai K, Jiang W, et al. China Kidney Disease Network (CK-NET) 2014 Annual Data Report. Am J Kidney Dis 2017 Jun;69(6S2):A4 [FREE Full text] [doi: 10.1053/j.ajkd.2016.06.011] [Medline: 28532549]
43. National Health Family Planning Commission. National Medical Service and Quality Safety Report. Beijing: People's Medical Publishing House; 2016.
44. Brasel KJ, Lim HJ, Nirula R, Weigelt JA. Length of stay: an appropriate quality measure? Arch Surg 2007 May;142(5):461-5; discussion 465. [doi: 10.1001/archsurg.142.5.461] [Medline: 17515488]
45. Zoller B, Spanaus K, Gerster R, Fasshauer M, Stehberger PA, Klinzing S, et al. ICG-liver test versus new biomarkers as prognostic markers for prolonged length of stay in critically ill patients - a prospective study of accuracy for prediction of length of stay in the ICU. Ann Intensive Care 2014;4:19 [FREE Full text] [doi: 10.1186/s13613-014-0019-7] [Medline: 25045579]
46. Song X, Xia C, Li Q, Yao C, Yao Y, Chen D, et al. Perioperative predictors of prolonged length of hospital stay following total knee arthroplasty: a retrospective study from a single center in China. BMC Musculoskelet Disord 2020 Jan 31;21(1):62 [FREE Full text] [doi: 10.1186/s12891-020-3042-x] [Medline: 32005208]
47. Breiman L. Random forests. Machine Learning 2001;45(1):5-32. [doi: 10.1023/A:1010933404324]
48. Kulkarni VY, Sinha PK, Petare MC. Weighted hybrid decision tree model for random forest classifier. J Inst Eng India Ser B 2015 Jan 3;97(2):209-217. [doi: 10.1007/s40031-014-0176-y]
49. Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, et al. Data mining in the life sciences with random forest: A walk in the park or lost in the jungle? Brief Bioinform 2013 May;14(3):315-326 [FREE Full text] [doi: 10.1093/bib/bbs034] [Medline: 22786785]
50. Hsu C-W, Chang C-C, Lin C-J. A practical guide to support vector classification. 2016. URL: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf [accessed 2021-04-14]
51. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000 Oct;16(10):906-914. [doi: 10.1093/bioinformatics/16.10.906] [Medline: 11120680]
52. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifier. New York, NY, United States: Association for Computing Machinery; 1992 Presented at: COLT92: 5th Annual Workshop on Computational Learning Theory; July 1992; Pittsburgh, Pennsylvania p. 144-152. [doi: 10.1145/130385.130401]
53. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inform Theory 1967 Jan;13(1):21-27. [doi: 10.1109/tit.1967.1053964]
54. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, et al. MicroRNA expression profiles classify human cancers. Nature 2005 Jun 09;435(7043):834-838. [doi: 10.1038/nature03702] [Medline: 15944708]
55. Sarasin-Filipowicz M, Oakeley EJ, Duong FHT, Christen V, Terracciano L, Filipowicz W, et al. Interferon signaling and treatment outcome in chronic hepatitis C. Proc Natl Acad Sci U S A 2008 May 13;105(19):7034-7039 [FREE Full text] [doi: 10.1073/pnas.0707882105] [Medline: 18467494]
56. Wang X, Yu J, Sreekumar A, Varambally S, Shen R, Giacherio D, et al. Autoantibody signatures in prostate cancer. N Engl J Med 2005 Sep 22;353(12):1224-1235. [doi: 10.1056/NEJMoa051931] [Medline: 16177248]
57. Hastie T, Tibshirani R, Friedman J. Prototype methods and nearest-neighbors. In: The Elements of Statistical Learning. New York, NY: Springer Science+Business Media; Jan 01, 2009:463-471.
58. Witten I, Frank E, Hall MA. Data Mining: Practical Machine Learning Tools and Techniques. San Francisco, CA: Morgan Kaufmann Publishers; Jan 01, 2011.

59. Wang R. Significantly improving the prediction of molecular atomization energies by an ensemble of machine learning algorithms and rescanning input space: A stacked generalization approach. J Phys Chem C 2018 Apr 12;122(16):8868-8873. [doi: 10.1021/acs.jpcc.8b03405]
60. Steyerberg E, Vickers A, Cook N, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010 Jan;21(1):128-138 [FREE Full text] [doi: 10.1097/EDE.0b013e3181c30fb2] [Medline: 20010215]
61. Van Hoorde K, Van Huffel S, Timmerman D, Bourne T, Van Calster B. A spline-based tool to assess and visualize the calibration of multiclass risk predictions. J Biomed Inform 2015 Apr;54:283-293 [FREE Full text] [doi: 10.1016/j.jbi.2014.12.016] [Medline: 25579635]
62. Hao M, Wang Y, Bryant SH. An efficient algorithm coupled with synthetic minority over-sampling technique to classify imbalanced PubChem BioAssay data. Anal Chim Acta 2014 Jan 02;806:117-127 [FREE Full text] [doi: 10.1016/j.aca.2013.10.050] [Medline: 24331047]
63. Steele R, Thompson B. Data mining for generalizable pre-admission prediction of elective length of stay. : IEEE; 2019 Presented at: IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC); January 7-9, 2019; Las Vegas, NV p. 0127-0133. [doi: 10.1109/CCWC.2019.8666598]
64. Kumar A, Anjomshoa H. A two-stage model to predict surgical patients' lengths of stay from an electronic patient database. IEEE J Biomed Health Inform 2019 Mar;23(2):848-856. [doi: 10.1109/JBHI.2018.2819646] [Medline: 29993793]
65. Liew D, Liew D, Kennedy MP. Emergency department length of stay independently predicts excess inpatient length of stay. Med J Aust 2003 Nov 17;179(10):524-526. [doi: 10.5694/j.1326-5377.2003.tb05676.x] [Medline: 14609414]
66. Verburg IWM, de Keizer NF, de Jonge E, Peek N. Comparison of regression methods for modeling intensive care length of stay. PLoS One 2014;9(10):e109684 [FREE Full text] [doi: 10.1371/journal.pone.0109684] [Medline: 25360612]
67. Harel Z, Wald R, McArthur E, Chertow GM, Harel S, Gruneir A, et al. Rehospitalizations and emergency department visits after hospital discharge in patients receiving maintenance hemodialysis. J Am Soc Nephrol 2015 Dec;26(12):3141-3150 [FREE Full text] [doi: 10.1681/ASN.2014060614] [Medline: 25855772]
68. Aga Z, Machina M, McCluskey S. Greater intravenous fluid volumes are associated with prolonged recovery after colorectal surgery: A retrospective cohort study. Br J Anaesth 2016 Jun;116(6):804-810 [FREE Full text] [doi: 10.1093/bja/aew125] [Medline: 27199312]
69. Farjah F, Lou F, Rusch VW, Rizk NP. The quality metric prolonged length of stay misses clinically important adverse events. Ann Thorac Surg 2012 Sep;94(3):881-7; discussion 887. [doi: 10.1016/j.athoracsur.2012.04.082] [Medline: 22742847]
70. Noh H, Lee SW, Kang SW, Shin SK, Choi KH, Lee HY, et al. Serum C-reactive protein: a predictor of mortality in continuous ambulatory peritoneal dialysis patients. Perit Dial Int 1998;18(4):387-394. [Medline: 10505560]

Abbreviations

ANN: artificial neural network
AUROC: area under the receiver operating characteristic curve
CKD: chronic kidney disease
ECI: estimated calibration index
ESKD: end-stage kidney disease
Gm: geometric mean
HQMS: Hospital Quality Monitoring System
ICD-10: International Statistical Classification of Diseases, Tenth Revision
ICU: intensive care unit
KNN: K-nearest neighbor
LOS: length of stay
LR: logistic regression
PD: peritoneal dialysis
pLOS: prolonged length of stay
RF: random forest
SVM: support vector machine

Edited by C Lovis; submitted 22.01.20; peer-reviewed by YJ Tseng, A James, W Pian, S Sarbadhikari; comments to author 07.06.20; revised version received 10.08.20; accepted 07.03.21; published 19.05.21.

Please cite as:
Kong G, Wu J, Chu H, Yang C, Lin Y, Lin K, Shi Y, Wang H, Zhang L
Predicting Prolonged Length of Hospital Stay for Peritoneal Dialysis–Treated Patients Using Stacked Generalization: Model Development and Validation Study
JMIR Med Inform 2021;9(5):e17886
URL: https://medinform.jmir.org/2021/5/e17886
doi:10.2196/17886
PMID:34009135

Original Paper

Automatically Diagnosing Disk Bulge and Disk Herniation With Lumbar Magnetic Resonance Images by Using Deep Convolutional Neural Networks: Method Development Study

Qiong Pan[1,2*], MS; Kai Zhang[3,4*], PhD; Lin He[3], BSc; Zhou Dong[5], MS; Lei Zhang[3], BSc; Xiaohang Wu[6], PhD; Yi Wu[7], MD; Yanjun Gao[8], PhD

[1]School of Telecommunications Engineering, Xidian University, Xi’an, China

[2]College of Science, Northwest A&F University, Yangling, China

[3]School of Computer Science and Technology, Xidian University, Xi'an, China

[4]SenseTime Group Limited, Shanghai, China

[5]School of Computer Science, Northwestern Polytechnical University, Xi'an, China

[6]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China

[7]Medical Imaging Department, The Affiliated Hospital of Northwest University Xi’an Number 3 Hospital, Xi’an, China

[8]Xi’an Key Laboratory of Cardiovascular and Cerebrovascular Diseases, The Affiliated Hospital of Northwest University Xi’an Number 3 Hospital, Xi'an, China

[*]these authors contributed equally

Corresponding Author:
Yanjun Gao, PhD
Xi’an Key Laboratory of Cardiovascular and Cerebrovascular Diseases
The Affiliated Hospital of Northwest University Xi’an Number 3 Hospital
No 10 Eastern section of the third FengCheng Rd
WeiYang District
Xi'an
China
Phone: 86 61816169
Fax: 86 61816100
Email: nige.001@stu.xjtu.edu.cn

Abstract

Background: Disk herniation and disk bulge are two common disorders of lumbar intervertebral disks (IVDs) that often result in numbness, pain in the lower limbs, and lower back pain. Magnetic resonance (MR) imaging is one of the most efficient techniques for detecting lumbar diseases and is widely used for making clinical diagnoses at hospitals. However, there is a lack of efficient tools for effectively interpreting massive amounts of MR images to meet the requirements of many radiologists.

Objective: The aim of this study was to present an automatic system for diagnosing disk bulge and herniation that saves time and can effectively and significantly reduce the workload of radiologists.

Methods: The diagnosis of lumbar vertebral disorders is highly dependent on medical images. Therefore, we chose the two most common diseases—disk bulge and herniation—as research subjects. This study is mainly about identifying the position of IVDs (lumbar vertebra [L] 1 to L2, L2-L3, L3-L4, L4-L5, and L5 to sacral vertebra [S] 1) by analyzing the geometrical relationship between sagittal and axial images and classifying axial lumbar disk MR images via deep convolutional neural networks.

Results: This system involved 4 steps. In the first step, it automatically located vertebral bodies (including the L1, L2, L3, L4, L5, and S1) in sagittal images by using the faster region-based convolutional neural network, and our fourfold cross-validation showed 100% accuracy. In the second step, it spontaneously identified the corresponding disk in each axial lumbar disk MR image with 100% accuracy. In the third step, the accuracy for automatically locating the intervertebral disk region of interest in axial MR images was 100%. In the fourth step, the 3-class classification (normal disk, disk bulge, and disk herniation) accuracies for the L1-L2, L2-L3, L3-L4, L4-L5, and L5-S1 IVDs were 92.7%, 84.4%, 92.1%, 90.4%, and 84.2%, respectively.

Conclusions: The automatic diagnosis system was successfully built, and it could classify images of normal disks, disk bulge, and disk herniation. This system provided a web-based test for interpreting lumbar disk MR images that could significantly improve diagnostic efficiency and standardized diagnosis reports. This system can also be used to detect other lumbar abnormalities and cervical spondylosis.

(JMIR Med Inform 2021;9(5):e14755) doi:10.2196/14755

KEYWORDS

deep learning; object localization; disk herniation; disk bulge; image classification

Introduction

Magnetic resonance imaging (MRI) is a widely used technique for detecting lumbar disorders, and its advantages include high image quality and noninvasive and ionization-free radiation. Disk herniation and disk bulge are two common types of lumbar intervertebral disk (IVD) injuries that often result in low back pain and tingling and numbness in the legs [1,2]. The diagnosis of disk disorders is highly dependent on radiology methods such as MRI. The leading question is as follows: how can radiologists interpret massive amounts of magnetic resonance (MR) images quickly and accurately for real-world applications? Motivated by machine learning– and deep learning–based clinical practice [3-6], we propose an automatic diagnosis system for diagnosing disk bulge and disk herniation with MR images via deep convolutional neural networks (CNNs), which can reduce radiologists' workload and provide the consistency required to produce standardized diagnosis reports.

Koh et al [7] proposed a computer-aided framework that uses several heterogeneous classifiers (ie, a perceptron classifier, a least mean squares classifier, a support vector machine classifier, and a k-means classifier) to construct a 2-level classification scheme for disk herniation diagnosis, which achieved 99% accuracy for 70 subjects. A probability classifier based on Gaussian models was proposed to detect abnormal IVDs. This model used the following three features: appearance, location, and context [8]. A study [9] on texture features that were obtained from IVD MR images used three different classifiers (ie, the back-propagation neural network, k-nearest neighbor, and support vector machine classifiers) to classify normal disks and IVDs and achieved a maximum accuracy of 83.33%. Additionally, many other methods have been proposed to automatically diagnose IVD diseases based on MR images [10-13]. Most of these models are for sagittal MR images, and there are very few studies that have used axial lumbar MR images, which are even more important in real clinical scenarios to identify disk bulge and herniation [13]. Most previous studies have mainly focused on binary classification (disease and normal) [7-9,11,12], as it is rare to study 2 diseases at the same time. In this study, we present a deep CNN–based diagnosis system for diagnosing lumbar disk bulge and disk herniation based on axial MR images. CNN analysis has proven to be an efficient method that is widely used to solve various image problems and has achieved huge success in many applicable fields [14-18].

This study aimed to develop a clinical applicable system that requires as little information from doctors as possible for diagnosing disk bulge and disk herniation via deep learning methods [19-21].

Methods

Data Set

In this study, lumbar MR Images and clinical diagnosis reports were collected from the Medical Imaging Department of Xi'an Number 3 Hospital, which is a large-scale grade 3A general hospital in Xi'an, China. The sagittal and axial T_2-weighted lumbar MR images of 500 patients were acquired by using a Philips Ingenia 3.0T scanner and exported in the Digital Imaging and Communications in Medicine (DICOM) format. The main diagnosis was based on axial images, as they display the morphology of IVDs more clearly than other images. For each subject, midsagittal images were used to locate IVDs in axial images. A total of 3555 axial images were used in this study. These images were labeled as normal disk, disk bulge, and disk herniation according to diagnosis reports and rechecked by an experienced radiologist, as shown in Table 1. Examples of midsagittal lumbar images and axial images of normal disks, disk bulge, and disk herniation are shown in Figure 1.

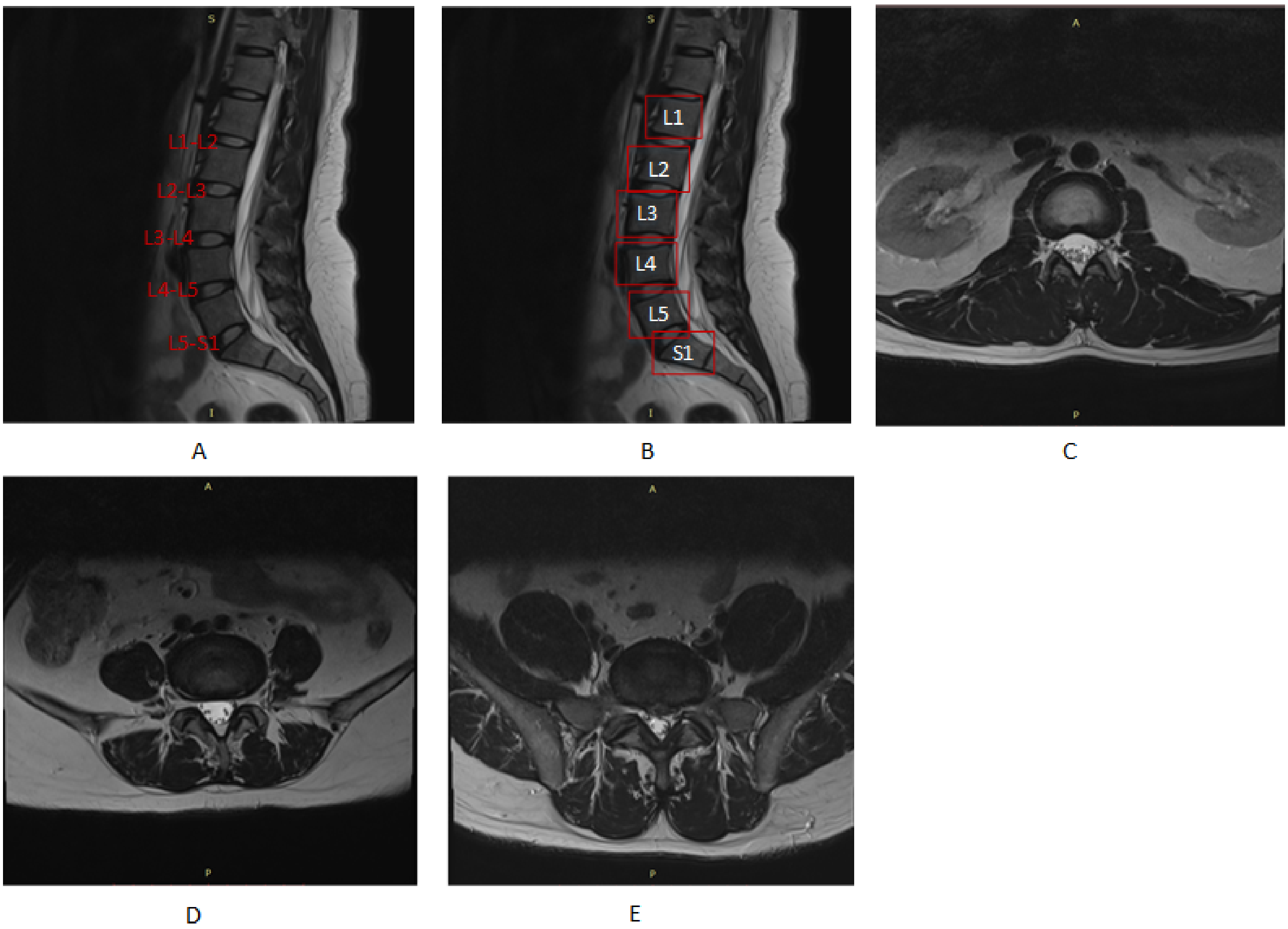
Table 1. The number of axial images in each category.

Intervertebral disk	Normal images, n	Bulge images, n	Herniation images, n	Total, n
L1-L2[a]	593	37	36	666
L2-L3	549	120	30	699
L3-L4	347	284	86	717
L4-L5	158	413	178	749
L5-S1[b]	238	242	244	724
All intervertebral disks	1885	1096	574	3555

[a] L: lumber vertebra.

[b] S: sacral vertebra.

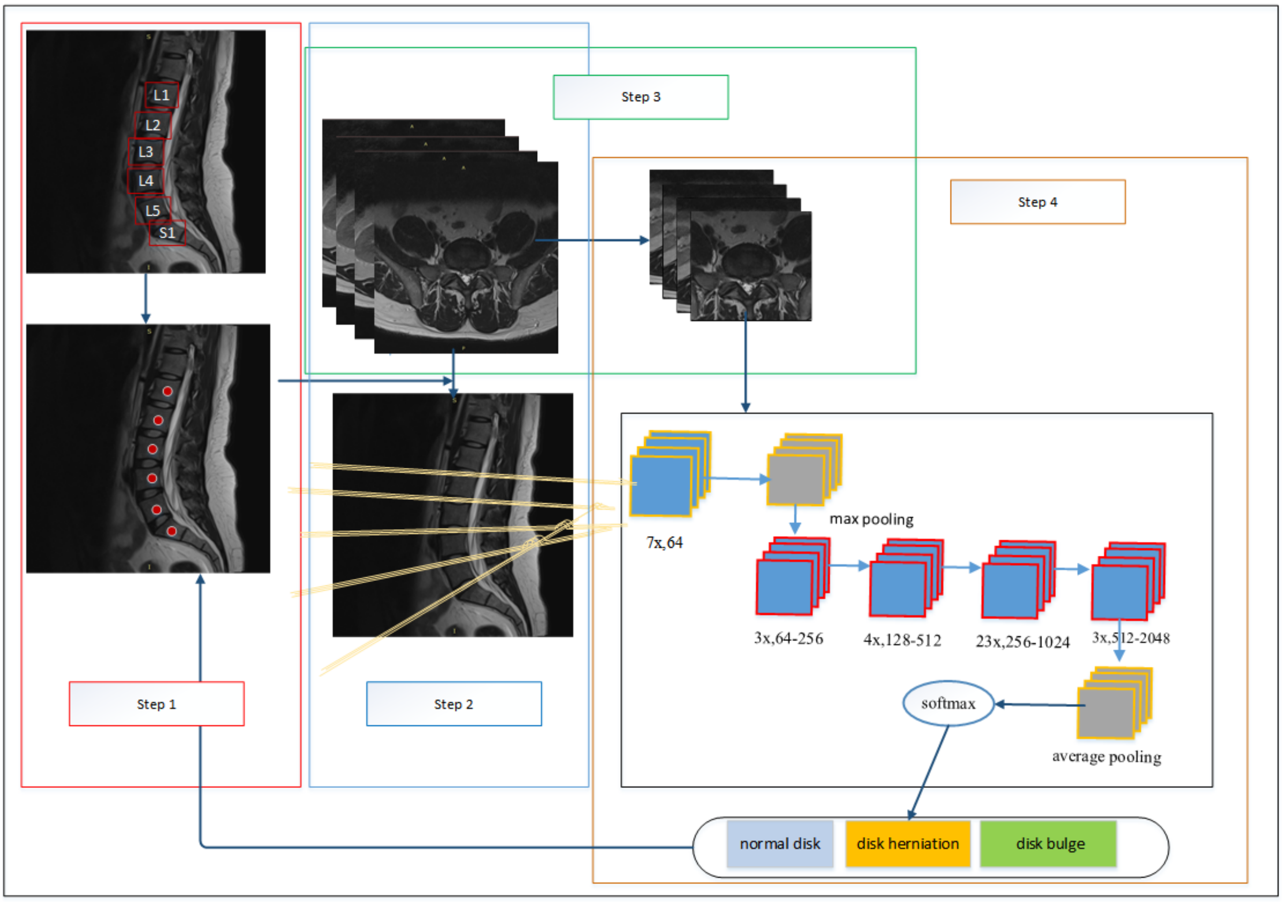
Figure 1. Examples of lumbar MR images. (A) A sagittal lumbar MR image in which 5 IVDs are labeled. (B) A sagittal lumbar MR image in which 6 vertebral bodies are enclosed in boxes. (C) An axial lumbar MR image of a normal disk. (D) An axial lumbar MR image of disk bulge. (E) An axial lumbar MR image of disk herniation. L: lumbar vertebra; MR: magnetic resonance; S: sacral.



Overall Diagnosis System

Our system consists of 4 steps, as shown in Figure 2. In the first step, the six lumbar vertebral bodies (lumbar vertebra [L] 1, L2, L3, L4, L5, and sacral vertebra [S] 1) in midsagittal images were detected and located. The second step was to identify the corresponding IVDs in each axial MR image. Afterward, these axial images were grouped into five categories (L1-L2, L2-L3, L3-L4, L4-L5, and L5-S1). In the third step, the IVD regions of interest (ROIs) in axial images were segmented to decrease the noise of the images. In the fourth step, each ROI image that included the five IVDs was classified as normal disk, disk bulge, or disk herniation.

Figure 2. Overall diagnosis system. This system consists of 4 steps. First, vertebral bodies (L1, L2, L3, L4, and L5) in sagittal lumbar magnetic resonance images were automatically located by using the faster R-CNN, and the middle point of each vertebral body was calculated. Second, the axial images were grouped into 5 categories. Each category corresponded to an intervertebral disk (ie, the L1-L2, L2-L3, L3-L4, L4-L5, and L5-S1 intervertebral disks). Third, the intervertebral disk regions of interest in each axial MR image were segmented using the faster R-CNN. Finally, in each category, the region-of-interest images were classified as images of normal disks, disk bulge, and disk herniation using ResNet101. L: lumbar vertebra; R-CNN: region-based convolutional neural network; S: sacral.



Automatically Locating Vertebral Bodies in Midsagittal Images

The faster region-based CNN (R-CNN) [19] was developed from the R-CNN [22] and the fast R-CNN [23], which unifies the target detection process (including candidate region generation, feature extraction, classification, and position refinement) into 1 deep network framework and greatly improves operational speed. In step 1, the faster R-CNN was used to locate the vertebral bodies in sagittal MR images.

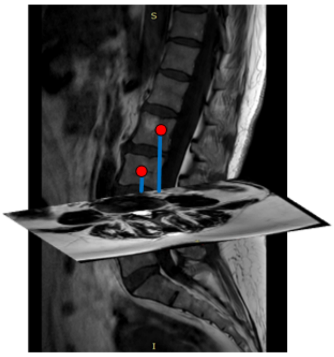
First, the six vertebral bodies (L1-S1) in 200 midsagittal images were manually located under the guidance of a radiologist. Second, the faster R-CNN was trained to detect and locate each vertebral body. We detected vertebral bodies instead of disks because they were easier to manually locate. Finally, the middle point coordinate of each vertebral body was calculated based on bounding box coordinates, as the precise location of the vertebral bodies would be used to locate the vertebrae in axial MR images, as shown in Figure 1 (step 1).

The faster R-CNN was implemented with Caffe [24] (Berkeley Vision and Learning Center deep learning framework) and trained in parallel on 4 Nvidia Titan X graphics processing units. Accuracy, sensitivity, and specificity [25,26] were analyzed to comprehensively evaluate the performance of this system.

Identifying the Corresponding IVD in Each Axial MR Image

For each subject, 15 axial slices were needed to identify the corresponding IVDs (L1-L2, L2-L3, L3-L4, L4-L5, and L5-S1) in each axial MR image. In step 1, the center point coordinates of the six vertebral bodies in the sagittal images were calculated. The directed distances from these center points to each axial image were calculated for each subject based on the spatial location relationship between sagittal images and axial images. The directed distances indicated which IVDs were closer to the corresponding IVDs in each axial image and which IVDs were located above or below the corresponding IVDs, as shown in Figure 3. Based on these distances, the axial slices were classified into 5 categories (L1-L2, L2-L3, L3-L4, L4-L5, and L5-S1). The conversion from DICOM patient-based coordinates to 2D computer coordinates was conducted in order to establish the relationship between the primitively processed images and the 3D DICOM coordinates. The detailed procedures are depicted in Multimedia Appendix 1.

Figure 3. The intervertebral disks (from L1-L2 to L5-S1) in each axial image were located by calculating directed distances. The red dot shows the middle point of each vertebral body in a sagittal image. The blue line depicts the directed distance from the red dot to a specific axial image. L: lumbar vertebra; S: sacral.



Locating IVD ROIs in Axial MR Images

Axial lumbar MR images contain large amounts of unrelated areas. In order to focus on IVDs and extract more relevant features, IVD areas were labeled manually in 1237 axial images, including normal disk areas, bulging disk areas, disk herniation areas, and the L1-L2 to L5-S1 IVD areas. The IVD areas of each ROI image needed to be located to train the faster R-CNN, and our fourfold cross-validation showed 100% accuracy. Afterward, the ROIs in each axial lumbar image were detected and extracted using the faster R-CNN, as shown in Figure 2 (step 3). We reserved a larger area for the components surrounding IVDs, as they may also help with identifying the condition of the disks (eg, the compression of the spinal canal).

Classification of ROI Images

It is worth mentioning that the degradation problem of the ultradeep CNN may result in reduced classification accuracy as the depth of the CNN increases. He et al [27] proposed a deep residual network framework that can solve this problem by using the residual block method, and this was proven to have significant accuracy for the ImageNet validation set [27-29]. The residual architecture of ResNet101 is shown in Figure 2 (step 4).

According to the diagnosis reports, in every category (L1-L2 to L5-S1), a total of 3555 axial MR images were labeled as normal disk, disk bulge, or disk herniation. All 3555 ROI images were reviewed by an expert radiologist to confirm whether the images conformed to the labels. Afterward, ResNet101 was used to conduct the 3-class classification for each category, and our fourfold cross-validation showed classification accuracies of 92.7%, 84.4%, 92.1%, 90.4% and 84.2% for the L1-L2, L2-L3, L3-L4, L4-L5, and L5-S1 IVDs, respectively. In this step, a cost-sensitive CNN was used to test for imbalances in the 3-class classification data set [30]. Relevant mathematical theory is provided in Multimedia Appendix 1.

Results

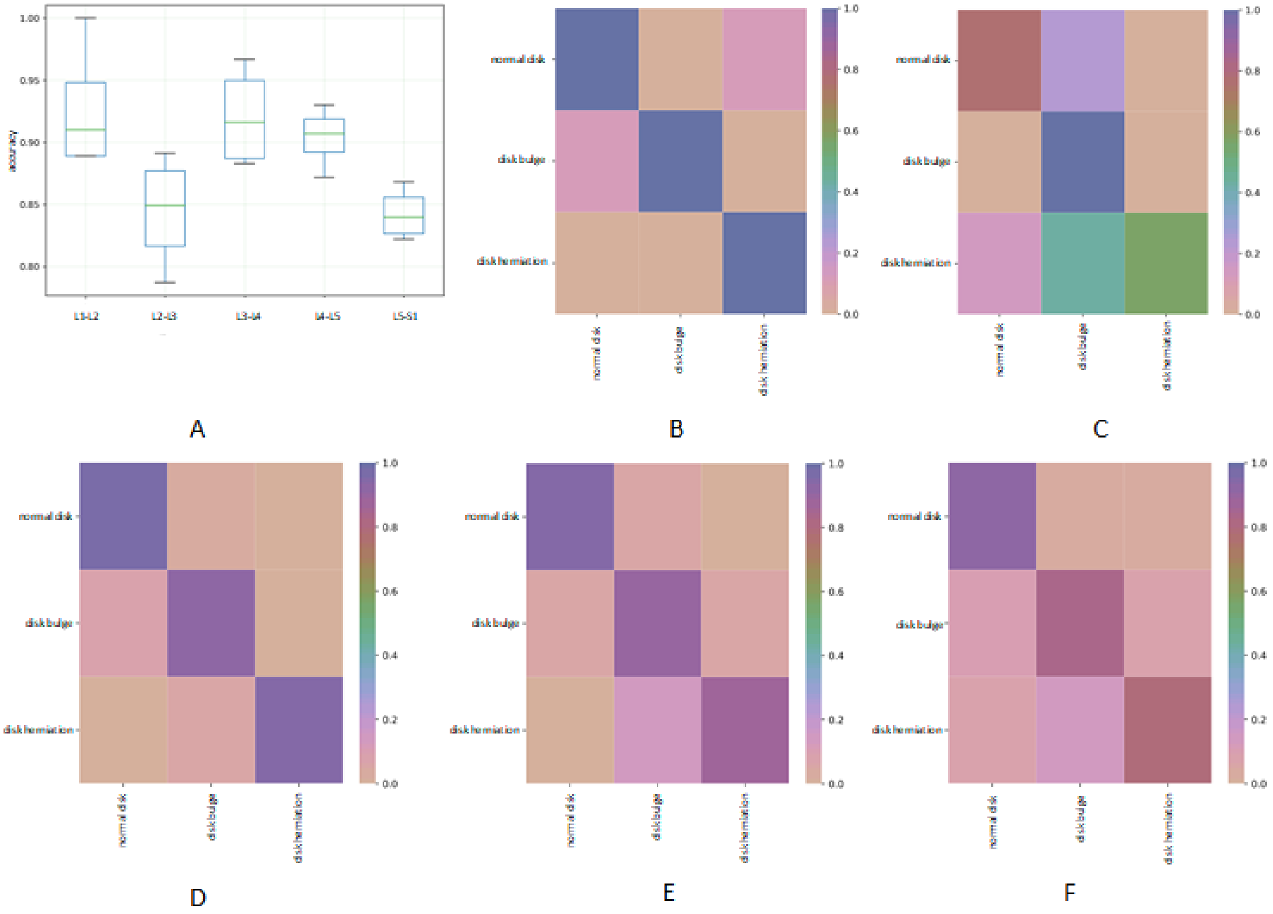
We focused on images that showed disk bulge, disk herniation, and normal disks. From Table 1, we can see that the probabilities of disk bulge and disk herniation in the L1-L2 and L2-L3 IVDs are low, and disk bulge tended to occur more commonly in the L3-L4, L4-L5, and L5-S1 IVDs. The L5-S1 IVD is the most common location of disk herniation. This is probably because it bears more weight and pressure than the other locations.

Discussion

Principal Findings

Our system is comprised of 4 steps. First, the system automatically located vertebral bodies (from L1 to S1) in sagittal images by using the faster R-CNN, which was trained on 200 manually cropped images. Our fourfold cross-validations showed 100% accuracy. This high location accuracy shows that the faster R-CNN method can more accurately locate vertebral bodies than many other methods, such as the Gabor filter bank method [31], which is a method based on measurements of disk signal intensity and structure [7]. Second, the disk positions (from L1-L2 to L5-S1) in each axial image were calculated based on the equations for coordinate conversion. We achieved an accuracy of 100%. Third, the system automatically segmented IVD ROIs in axial MR images by using the faster R-CNN, which was trained on 1300 manually boxed images that included all five types of disks (from L1-L2 to L5-S1) and the disk conditions (normal, herniation, and bulge). The mean average precision [21] reached 100%. This high accuracy was the result of the excellent performance of the faster R-CNN. Finally, all ROI images were classified as normal, bulge, and herniation by using ResNet101. The average accuracies for the 3-class classification of the L1-L2, L2-L3, L3-L4, L4-L5, and L5-S1 IVDs were 92.7%, 84.4%, 92.1%, 90.4%, and 84.2%, respectively. All relevant results are shown in Figure 4. Previous studies have mainly focused on comparing IVDs affected by 1 disease (disk bulge or herniation) with normal IVDs. This is known as a binary classification. For example, the performance value of one IVD classification system was 86.5%, and this was based on a sparse shape reconstruction from a statistical shape model [32]. Additionally, an accuracy of 92.78% was reported by a study that classified normal disks and disk bulge by using a program called IVD Descriptor [13]. Compared to the accuracies of these previous studies, our accuracies were roughly the same or slightly inferior. This was mainly because a 3-class classification system is often less accurate than a binary classification system.

Figure 4. Results of the 3-class classification (normal disk, disk bulge, and disk herniation). (A) The average accuracies of the classification system (calculated using ResNet101) for the following five categories: L1-L2, L2-L3, L3-L4, L4-L5 and L5-S1. The rows and columns of all heat maps represent ground truth labels and predicted labels, respectively. The x-axis shows the five intervertebral disks. (B) A heat map of the classification accuracies for category L1-L2. The color scale expresses the accuracy. (C) A heat map of the classification accuracies for category L2-L3. The color scale expresses the accuracy. (D) A heat map of the classification accuracies for category L3-L4. The color scale expresses the accuracy. (E) A heat map of the classification accuracies for category L4-L5. The color scale expresses the accuracy. (F) A heat map of the classification accuracies for category L5-S1. The color scale expresses the accuracy. L: lumbar vertebra; S: sacral.

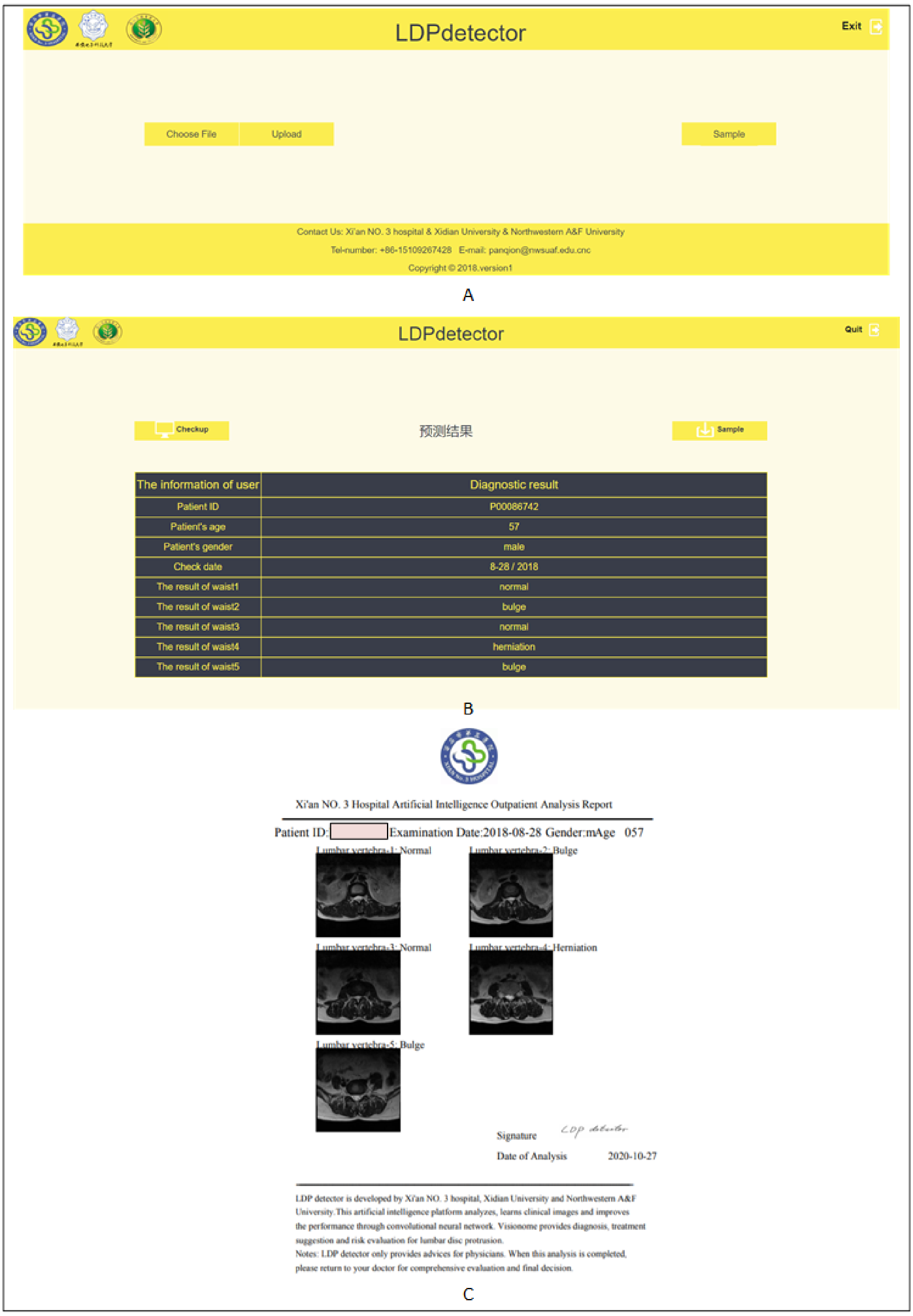


Based on our results, the classification accuracies for the L2-L3 and L5-S1 IVDs were lower than those for other disks. The shape of a normal disk is somewhat different from the L1-L2 to L5-S1 IVDs. With regard to the L2-L3 disks, several images were blurry, and it was difficult to identify subtle differences. This, coupled with our small sample of herniated disks, had a considerable impact on our classification accuracy. Data quality may become a crucial factor that could restrict the performance of algorithms used in research [33]. With regard to the L5-S1 disks, the normal disks were similar in shape to that of bulged disks in axial images. There were also a few images that were wrongfully classified by our system, which resulted in a lower classification accuracy.

Web-Based Diagnosis System

We used the Django framework [34] to develop an automatic diagnosis system for radiologists that could analyze inputted medical images and show results as normalized diagnosis reports (a PDF file). The appearance and functions of the reports are shown in Figure 5. This system can be deployed in multiple radiology departments to analyze patients' lumbar MR images and collect more images to improve radiologists' IVD interpretation performance. This system is freely available [35].

Figure 5. Appearance and functions of the reports of the web-based automatic diagnostic system. (A) This is the page for uploading a folder. (B) Diagnostic results in tabular form. (C) The diagnostic report in the Unified format. LDP: lumbar disk protrusion.



Contact Us: Xi'an NO. 3 hospital & Xidian University & Northwestern A&F University
Tel-number: +86-15109267428 E-mail: panqion@nwsuaf.edu.cnc
Copyright © 2018.version1

The information of user	Diagnostic result
Patient ID	P00086742
Patient's age	57
Patient's gender	male
Check date	8-28 / 2018
The result of waist1	normal
The result of waist2	bulge
The result of waist3	normal
The result of waist4	herniation
The result of waist5	bulge

Xi'an NO. 3 Hospital Artificial Intelligence Outpatient Analysis Report

Patient ID: Examination Date:2018-08-28 Gender:mAge 057

Lumbar vertebra-1: Normal

Lumbar vertebra-2: Bulge

Lumbar vertebra-3: Normal

Lumbar vertebra-4: Herniation

Lumbar vertebra-5: Bulge

Signature LDP detector

Date of Analysis 2020-10-27

LDP detector is developed by Xi'an NO. 3 hospital, Xidian University and Northwestern A&F University. This artificial intelligence platform analyzes, learns clinical images and improves the performance through convolutional neural network. Visionome provides diagnosis, treatment suggestion and risk evaluation for lumbar disc protrusion.

Notes: LDP detector only provides advices for physicians. When this analysis is completed, please return to your doctor for comprehensive evaluation and final decision.

In this paper, we present an automatic diagnosis system for diagnosing disk bulge and disk herniation with axial MR images via deep convolutional neural networks. This system can automatically determine the position and the condition of IVDs in axial MR images. Therefore, this system could help reduce the workloads of radiologists by analyzing lumbar MR images via a standardized method. In addition, this system can be expanded to analyze other types of lumbar diseases, such as cervical spondylosis. However, there are some limitations to using this system. Data from this system could be fundamentally limited by the quality of images (eg, when the image is blurry), making it difficult to identify subtle differences. The system is also limited by the size of the total data set, as it is relatively small for deep convolutional neural networks. Our future work will focus on the following two aspects: (1) developing this system by using a more targeted method that analyzes the specific features of MR images, and (2) gathering more MR images to train a more practical and complete automatic diagnosis system.

Acknowledgments

This study was funded by the National Key Research and Development Program of China (grant 2018YFC0116500).

Authors' Contributions

YJG designed the study. QP and KZ conducted the study. YW, ZD, and YJG collected and labeled the data. KZ, LH, ZD, QP, and LZ were responsible for coding and analyzing the results. QP and KZ cowrote the manuscript. XHW, KZ and QP critically reviewed and revised the manuscript. All authors discussed the results and commented on the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials.
[DOCX File , 16 KB - medinform_v9i5e14755_app1.docx]

References

1. Vos T, Flaxman A, Naghavi M, Lozano R, Michaud C, Ezzati M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 2012 Dec 15;380(9859):2163-2196 [FREE Full text] [doi: 10.1016/S0140-6736(12)61729-2] [Medline: 23245607]
2. Yang H, Liu H, Li Z, Zhang K, Wang J, Wang H, et al. Low back pain associated with lumbar disc herniation: role of moderately degenerative disc and annulus fibrous tears. Int J Clin Exp Med 2015 Feb 15;8(2):1634-1644 [FREE Full text] [Medline: 25932092]
3. Li W, Yang Y, Zhang K, Long E, He L, Zhang L, et al. Dense anatomical annotation of slit-lamp images improves the performance of deep learning for the diagnosis of ophthalmic disorders. Nat Biomed Eng 2020 Aug;4(8):767-777. [doi: 10.1038/s41551-020-0577-y] [Medline: 32572198]
4. Wang L, Zhang K, Liu X, Long E, Jiang J, An Y, et al. Comparative analysis of image classification methods for automatic diagnosis of ophthalmic images. Sci Rep 2017 Jan 31;7:41545 [FREE Full text] [doi: 10.1038/srep41545] [Medline: 28139688]
5. Talo M, Baloglu UB, Yıldırım Ö, Rajendra Acharya U. Application of deep transfer learning for automated brain abnormality classification using MR images. Cogn Syst Res 2019 May;54:176-188. [doi: 10.1016/j.cogsys.2018.12.007]
6. Lu J, Pedemonte S, Bizzo B, Doyle S, Andriole K, Michalski M, et al. DeepSPINE: Automated Lumbar Vertebral Segmentation, Disc-level Designation, and Spinal Stenosis Grading Using Deep Learning. 2018 Presented at: Machine Learning for Healthcare Conference; August 17-18, 2018; Palo Alto, California, USA.
7. Koh J, Chaudhary V, Dhillon G. Disc herniation diagnosis in MRI using a CAD framework and a two-level classifier. Int J Comput Assist Radiol Surg 2012 Nov;7(6):861-869. [doi: 10.1007/s11548-012-0674-9] [Medline: 22392057]
8. Alomari RS, Corso JJ, Chaudhary V, Dhillon G. Computer-aided diagnosis of lumbar disc pathology from clinical lower spine MRI. Int J Comput Assist Radiol Surg 2010 May;5(3):287-293. [doi: 10.1007/s11548-009-0396-9] [Medline: 20033498]
9. Hashia B, Mir AH. Texture features' based classification of MR images of normal and herniated intervertebral discs. Multimed Tools Appl 2018 Dec 11;79(21-22):15171-15190. [doi: 10.1007/s11042-018-7011-4]
10. Ruiz-España S, Arana E, Moratal D. Semiautomatic computer-aided classification of degenerative lumbar spine disease in magnetic resonance imaging. Comput Biol Med 2015 Jul;62:196-205. [doi: 10.1016/j.compbiomed.2015.04.028] [Medline: 25957744]
11. Preetha J, Selvarajan S. Computer aided diagnostic system for automatic cervical disc herniation classification. J Med Imaging Health Inform 2016 Nov 01;6(7):1589-1593. [doi: 10.1166/jmihi.2016.1855]
12. Alomari RS, Corso JJ, Chaudhary V, Dhillon G. Automatic diagnosis of lumbar disc herniation with shape and appearance features from MRI. 2010 Mar 09 Presented at: SPIE Medical Imaging 2010: Computer-Aided Diagnosis; February 13-18, 2010; San Diego, California, United States. [doi: 10.1117/12.842199]
13. Beulah A, Sharmila TS, Pramod VK. Disc bulge diagnostic model in axial lumbar MR images using Intervertebral disc Descriptor (IdD). Multimed Tools Appl 2018 Mar 23;77(20):27215-27230. [doi: 10.1007/s11042-018-5914-8]
14. 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]

15. Zhang K, Liu X, Liu F, He L, Zhang L, Yang Y, et al. An interpretable and expandable deep learning diagnostic system for multiple ocular diseases: Qualitative study. J Med Internet Res 2018 Nov 14;20(11):e11144 [FREE Full text] [doi: 10.2196/11144] [Medline: 30429111]
16. 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 Feb 02;542(7639):115-118. [doi: 10.1038/nature21056] [Medline: 28117445]
17. Abdel-Hamid O, Mohamed A, Jiang H, Deng L, Penn G, Yu D. Convolutional Neural Networks for Speech Recognition. IEEE/ACM Transactions on Audio Speech and Language Process 2014 Oct;22(10):1533-1545. [doi: 10.1109/taslp.2014.2339736]
18. Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE. A survey of deep neural network architectures and their applications. Neurocomputing 2017 Apr;234:11-26. [doi: 10.1016/j.neucom.2016.12.038]
19. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2017 Jun;39(6):1137-1149. [doi: 10.1109/TPAMI.2016.2577031] [Medline: 27295650]
20. Zhang K, Li X, He L, Guo C, Yang Y, Dong Z, et al. A human-in-the-loop deep learning paradigm for synergic visual evaluation in children. Neural Netw 2020 Feb;122:163-173. [doi: 10.1016/j.neunet.2019.10.003] [Medline: 31683144]
21. Yang J, Zhang K, Fan H, Huang Z, Xiang Y, Yang J, et al. Development and validation of deep learning algorithms for scoliosis screening using back images. Commun Biol 2019 Oct 25;2:390 [FREE Full text] [doi: 10.1038/s42003-019-0635-8] [Medline: 31667364]
22. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. 2014 Presented at: Internaltional Conference on computer vision and pattern recognition; June 24-27, 2014; Columbus, OH, USA p. 580-587. [doi: 10.1109/cvpr.2014.81]
23. Girshick R. Fast R-CNN. 2015 Presented at: IEEE International Conference on Computer Vision; December 13-16, 2015; Santiago, Chile p. 1440-1448. [doi: 10.1109/iccv.2015.169]
24. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, et al. Caffe: Convolutional Architecture for Fast Feature Embedding. 2014 Nov Presented at: 2014 ACM Multimedia Conference; November 2014; Orlando, Florida, USA p. 675-678. [doi: 10.1145/2647868.2654889]
25. Zhang X, Zhang K, Lin D, Zhu Y, Chen C, He L, et al. Artificial intelligence deciphers codes for color and odor perceptions based on large-scale chemoinformatic data. Gigascience 2020 Feb 01;9(2):giaa011 [FREE Full text] [doi: 10.1093/gigascience/giaa011] [Medline: 32101298]
26. Zhang K, Liu X, Jiang J, Li W, Wang S, Liu L, et al. Prediction of postoperative complications of pediatric cataract patients using data mining. J Transl Med 2019 Jan 03;17(1):2 [FREE Full text] [doi: 10.1186/s12967-018-1758-2] [Medline: 30602368]
27. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2015 Presented at: Internaltional Conference on Computer Vision and Pattern Recognition; June 8-10, 2015; Boston, Massachusetts, USA. [doi: 10.1109/cvpr.2016.90]
28. Bi L, Kim J, Ahn E, Feng D. Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. arXiv. Preprint posted online on March 17, 2017. [FREE Full text]
29. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers R. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. 2017 Presented at: Internaltional Conference on Computer Vision and Pattern Recognition; July 22-25, 2017; Honolulu, Hawaii, USA p. 3462-3471. [doi: 10.1109/cvpr.2017.369]
30. Jiang J, Liu X, Zhang K, Long E, Wang L, Li W, et al. Automatic diagnosis of imbalanced ophthalmic images using a cost-sensitive deep convolutional neural network. Biomed Eng Online 2017 Nov 21;16(1):132 [FREE Full text] [doi: 10.1186/s12938-017-0420-1] [Medline: 29157240]
31. Zhu X, He X, Wang P, He Q, Gao D, Cheng J, et al. A method of localization and segmentation of intervertebral discs in spine MRI based on Gabor filter bank. Biomed Eng Online 2016 Mar 22;15:32 [FREE Full text] [doi: 10.1186/s12938-016-0146-5] [Medline: 27000749]
32. Neubert A, Fripp J, Engstrom C, Schwarz D, Weber M, Crozier S. Statistical shape model reconstruction with sparse anomalous deformations: Application to intervertebral disc herniation. Comput Med Imaging Graph 2015 Dec;46 Pt 1:11-19. [doi: 10.1016/j.compmedimag.2015.05.002] [Medline: 26060085]
33. Cai L, Zhu Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Sci J 2015 May 22;14:2. [doi: 10.5334/dsj-2015-002]
34. Holovaty A, Kaplan-Moss J. The Definitive Guide to Django: Web Development Done Right. USA: Apress; 2009.
35. He L, Dong Z, Zhang L, Pan Q, Zhang K. Lumbar Disk Protrusion Detection System. URL: http://121.46.19.53/

Abbreviations

CNN: convolutional neural network
DICOM: Digital Imaging and Communications in Medicine
IVD: intervertebral disk
L: lumbar vertebra

MR: magnetic resonance
MRI: magnetic resonance imaging
R-CNN: region-based convolutional neural network
ROI: region of interest
S: sacral vertebra

Edited by G Eysenbach, Q Zeng; submitted 18.05.19; peer-reviewed by G Lim, S Kim, Z Li; comments to author 02.09.20; revised version received 27.10.20; accepted 15.04.21; published 21.05.21.

Please cite as:
Pan Q, Zhang K, He L, Dong Z, Zhang L, Wu X, Wu Y, Gao Y
Automatically Diagnosing Disk Bulge and Disk Herniation With Lumbar Magnetic Resonance Images by Using Deep Convolutional Neural Networks: Method Development Study
JMIR Med Inform 2021;9(5):e14755
URL: https://medinform.jmir.org/2021/5/e14755
doi:10.2196/14755
PMID:34018488

Original Paper

Improving Current Glycated Hemoglobin Prediction in Adults: Use of Machine Learning Algorithms With Electronic Health Records

Zakhriya Alhassan[1,2], PhD; Matthew Watson[1], MSc; David Budgen[1], PhD; Riyad Alshammari[3], PhD; Ali Alessa[4], PhD; Noura Al Moubayed[1], PhD

[1]Department of Computer Science, Durham University, Durham, United Kingdom

[2]College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia

[3]National Center for Artificial Intelligence, Saudi Data and Artificial Intelligence Authority, Riyadh, Saudi Arabia

[4]Department of Information Technology Programs, Institute of Public Administration, Riyadh, Saudi Arabia

Corresponding Author:
Noura Al Moubayed, PhD
Department of Computer Science
Durham University
Mountjoy Centre
Durham, DH1 3LE
United Kingdom
Phone: 44 1913 341724 ext 41749
Email: noura.al-moubayed@durham.ac.uk

Abstract

Background: Predicting the risk of glycated hemoglobin (HbA$_{1c}$) elevation can help identify patients with the potential for developing serious chronic health problems, such as diabetes. Early preventive interventions based upon advanced predictive models using electronic health records data for identifying such patients can ultimately help provide better health outcomes.

Objective: Our study investigated the performance of predictive models to forecast HbA$_{1c}$ elevation levels by employing several machine learning models. We also examined the use of patient electronic health record longitudinal data in the performance of the predictive models. Explainable methods were employed to interpret the decisions made by the black box models.

Methods: This study employed multiple logistic regression, random forest, support vector machine, and logistic regression models, as well as a deep learning model (multilayer perceptron) to classify patients with normal (<5.7%) and elevated (≥5.7%) levels of HbA$_{1c}$. We also integrated current visit data with historical (longitudinal) data from previous visits. Explainable machine learning methods were used to interrogate the models and provide an understanding of the reasons behind the decisions made by the models. All models were trained and tested using a large data set from Saudi Arabia with 18,844 unique patient records.

Results: The machine learning models achieved promising results for predicting current HbA$_{1c}$ elevation risk. When coupled with longitudinal data, the machine learning models outperformed the multiple logistic regression model used in the comparative study. The multilayer perceptron model achieved an accuracy of 83.22% for the area under receiver operating characteristic curve when used with historical data. All models showed a close level of agreement on the contribution of random blood sugar and age variables with and without longitudinal data.

Conclusions: This study shows that machine learning models can provide promising results for the task of predicting current HbA$_{1c}$ levels (≥5.7% or less). Using patients' longitudinal data improved the performance and affected the relative importance for the predictors used. The models showed results that are consistent with comparable studies.

(JMIR Med Inform 2021;9(5):e25237) doi:10.2196/25237

KEYWORDS

glycated hemoglobin HbA1c; prediction; machine learning; deep learning; neural network; multilayer perceptron; electronic health records; time series data; longitudinal data; diabetes

Introduction

Background

The level of glycated hemoglobin (HbA_{1c}) is used to measure the average glucose concentration in red blood cells [1,2]. Unlike other glucose blood tests, such as random blood sugar (RBS) and fasting blood sugar (FBS), HbA_{1c} provides a long-term measure of a patient's blood glucose levels [3]. The HbA_{1c} test can therefore provide physicians with a reliable means of monitoring a patient's hyperglycemia without requiring the patient to undertake overnight fasting prior to being tested.

A concentration of 6.5% for the HbA_{1c} in patient blood is considered as the cutoff point for the diagnosis of diabetes [4]. However, patients with a concentration of less than 6.5% are not completely excluded from a diabetes diagnosis, as the range of elevation levels ($5.7\% \leq HbA_{1c} < 6.5\%$) can indicate the future onset of diabetes. Therefore, HbA_{1c} can act as an early predictor for the potential development of type-2 diabetes mellitus (T2DM) [2]. Ackermann et al [3] suggested using the HbA_{1c} test as a measure for identifying those adults who are at a greater risk of developing T2DM in the future.

Research has shown that reducing HbA_{1c} levels can significantly reduce the possibility of developing serious complications. Hence, close monitoring of HbA_{1c} levels is recommended for all diabetic patients and those with the potential for developing diabetes [5]. It is also suggested that diabetic and nondiabetic patients with raised HbA_{1c} levels should be clinically checked and monitored as a preventive intervention to avoid developing T2DM [6].

Currently, the clinical data collected from patient visits consists of a set of readings for vital signs and lab tests, diagnoses, physicians' notes, and treatments that are stored in electronic health records (EHRs). These are collected on an irregular basis, according to clinical needs, and stored with an associated time stamp.

In recent years, machine learning models have shown powerful capabilities for analyzing and understanding complex data across a wide variety of applications. Our research question for this study was as follows: "Can HbA_{1c} prediction be improved by using machine learning with longitudinal data that are normally available in EHR systems?"

This paper reports an investigation into the performance of machine learning models to predict current HbA_{1c} levels as a binary classification problem using EHR data. Nondiabetic patients with an HbA_{1c} level of 5.7% or more are considered to have an elevated HbA_{1c}, while those with levels lower than this are considered normal. The models combine current visit data with extra features (independent variables) extracted from previous visits by patients. We used explainable methods to rank the features in order of their importance to the decision made by each of the models. To the best of our knowledge, this study is the first to employ machine learning models that use longitudinal data from EHR systems for the purpose of HbA_{1c} elevation risk prediction. This study is also the first to use explainable machine learning techniques to explain the classification decisions made by black box models, support vector machine (SVM), and multilayer perceptron (MLP), in predicting HbA_{1c} elevation risk (≥5.7%), in order to better understand the behavior of the model.

Related Work

EHR data have been intensively investigated for a variety of medical decision support tasks [7]. These tasks include the analysis of complex patterns and prediction of major medical events (for example, diagnostic imaging and gene interactions) [8,9]. Several studies have demonstrated the successful employment of EHR data with prediction models [10]. For instance, machine learning has been intensively used with EHR data in diagnosing diabetes and discovering its related patterns [11-15]. However, we are not aware of any studies that have explored machine learning models for the prediction of current elevated HbA_{1c} levels using EHR data from a nondiabetic population or the impact of patient longitudinal data on the effectiveness of such predictive machine learning models.

Several studies have investigated the association between HbA_{1c} levels and clinical variables using statistical models [16,17]. A study by Rose et al [18] discussed the correlation between RBS and HbA_{1c} levels. Stanley et al [19] used a linear regression model for imputation of missing HbA_{1c} data. Their model calculates HbA_{1c} levels for patient records with missing HbA_{1c} values as continuous and categorical values and uses 4 predictors extracted from an EHR system—RBS, FBS, age, and gender—as predictors to calculate the level of HbA_{1c} for a diabetic population. Simone et al [20] used linear regression models to predict HbA_{1c} levels after 6 years for nondiabetic patients using different populations.

A study by Wells et al [21] in 2018 was the first to focus on predicting current HbA_{1c} elevation levels for nondiabetic patients through use of an EHR data set. Multiple logistic regression (MLR) was employed to calculate the probability of a patient having an elevated HbA_{1c} level (≥5.7%). The data set was extracted from an EHR system used in the United States. The authors used 8 independent variables fitted to the model using restricted cubic splines with 3 knots to formulate the final equation. The performance of the MLR model was compared to that of the models used by Baan et al [22] and Griffin et al [23]. However, the models by Baan and Griffin aimed at predicting the onset of patients' diabetes rather than predicting HbA_{1c} levels for nondiabetic patients. In addition, the experimental data set used by Wells et al to train and test their model was imbalanced with 74% of the samples having normal HbA_{1c} levels (5.7%) and only 26% of the samples having elevated HbA_{1c} levels (≥5.7%).

We performed a differentiated replication of the study by Wells et al [21] using the more balanced King Abdullah International Medical Research Center (KAIMRC) data set [24]. Although the significant variables identified in our replication were in general agreement with those of the original study, there were some differences in the ranking of importance for these, suggesting that such models do need to be "tuned" to the characteristics of different populations.

Methods

Study Design

To study the impact of using advanced predictive models with EHR data to predict current HbA_{1c} levels, we employed the MLR, random forest (RF), SVM, and logistic regression (LR) models, as well as a deep learning model, MLP [25]. The problem was formulated into a binary classification problem whereby the target variable, HbA_{1c} level, was encoded as 1 when the level of HbA_{1c} was 5.7% or more and with 0 otherwise. The results obtained from using these models were compared to those obtained from employing the model used by Wells et al with the KAIMRC data set (detailed in the Data Set subsection).

The performance of the models was investigated using current visit data only and with additional longitudinal data from current and previous visits. The performance of each model was evaluated using measures commonly employed in clinical applications. For the SVM and MLP models, the relative importance of the features was also calculated using explainable machine learning techniques.

Explainable Methods for Black Box Models

Using black box machine learning models in health care can have adverse effects on the trust and confidence placed in their outcomes; the risk of misclassification is potentially too high for clinicians to confidently use black box models for high risk health care decisions, and not being able to interpret a model's decision exacerbates this problem [26]. Explainable methods for machine learning models allow interpretable outcomes that can expose the reasons behind the decision made by the model [27]. This transparency provides both health professionals and patients with the confidence and trust in the outcome of the models. The widely used Shapley Additive Explanations (SHAP) values [28] and local interpretable model-agnostic explanations (LIME) score [29] techniques have therefore been used to provide a degree of transparency to our deep learning model.

SHAP values are derived from Shapley values used in game theory and provide a method of calculating the contribution of each feature (variable) to the final prediction via the GradientSHAP approximation. This is achieved for each feature by comparing the prediction the model makes when the feature is present with the prediction obtained when the feature takes some baseline value [28]. Consequently, the SHAP values for a given input "explain" how each feature affects the output of the model when compared to the baseline (or "default") output of the model. We used SHAP values to interpret our black box models, so they could be efficiently calculated, and their use enabled a global view of the model to be constructed through the computation of SHAP values from across the whole data set.

SHAP values were computed using the feature's mean marginal contribution across different coalitions of all features. SHAP values themselves are computationally intensive to compute, and so approximation methods are commonly used when calculating the values.

To ensure that the SHAP values we calculated were not too greatly affected by the approximation method used, we also computed the LIME [29] scores for the models across the entire data set. LIME tries to estimate locally faithful linear explanations (ie, explanations that correspond to how the model behaves around the instance being explained) for any classifier. LIME achieves this by creating local linear classifiers that approximate the behavior of the original model in the vicinity of the data being explained. As linear models are inherently interpretable through their parameters, they can be used to generate explanations of the original model. Both SHAP and LIME have the advantage that they are model-agnostic techniques, and so we were able to apply both methods to both of our black box classification models (SVM and MLP).

Data Set

The data used in this study were taken from the KAIMRC data set. The data were collected from King Abdulaziz Medical City located in the central and western regions of Saudi Arabia, an area which has been ranked second in the Middle East and seventeenth in world in diabetes prevalence by the World Health Organization (WHO) [30]. According to the International Diabetes Federation, the diabetes prevalence rate in Saudi Arabia is 18.3%. Therefore, the availability of the data from this population provides considerable opportunities for research into the early prediction of diabetes.

The data set contains a full history of patient details, vital signs, and lab test readings for each patient visit for the period from 2016 to the end of 2018. As the aim of this study was to identify nondiabetic patients that are at a high risk of HbA_{1c} elevation, all patients previously diagnosed with hyperglycemia were excluded from the experimental data set. The remaining cohort formed our experimental data set and was categorized by using the American Diabetes Association's guidelines [31], in which patients with HbA_{1c} readings of more than 5.7% are considered as being in the prediabetic range, while those with less than 5.7% are considered to be in the normal range.

Most medical data sets are imbalanced [32-34]. These imbalances occur when the proportion of one class of patients in the data set is greater than its counterpart class [35,36]. However, unusually, our experimental data set was not imbalanced. Slightly over half of the patients in our experimental data set (9826/18,844, 52.14%) were found to have elevated levels of HbA_{1c} (≥5.7%) while 47.86% (9018/18,844) of patients had normal HbA_{1c} levels (<5.7%). This can be ascribed to the high incidence of diabetes in the region from which the data set was collected [37].

A detailed illustration of the patients' class distribution (HbA_{1c} levels) by age groups and gender is shown in Figure 1. This shows that as the age of patients increased, so did the proportion of patients who had elevated HbA_{1c} levels. The data set also exhibited a balanced gender distribution, with 49.40% (9308/18,844) of the patients being male and 50.60%

(9536/18,844) being female. However, the proportion of male patients with elevated levels of HbA_{1c} (≥5.7%) was greater than that of the female patients. Also, female patients with normal levels of HbA_{1c} (<5.7%) made more visits than did males. Table 1 shows the profile for the distribution of HbA_{1c} elevation levels organized by gender.

Figure 1. HbA_{1c} elevation levels distributed over age range and gender in the King Abdullah International Medical Research Center (KAIMRC) data set (before sampling). HbA_{1c}: glycated hemoglobin.

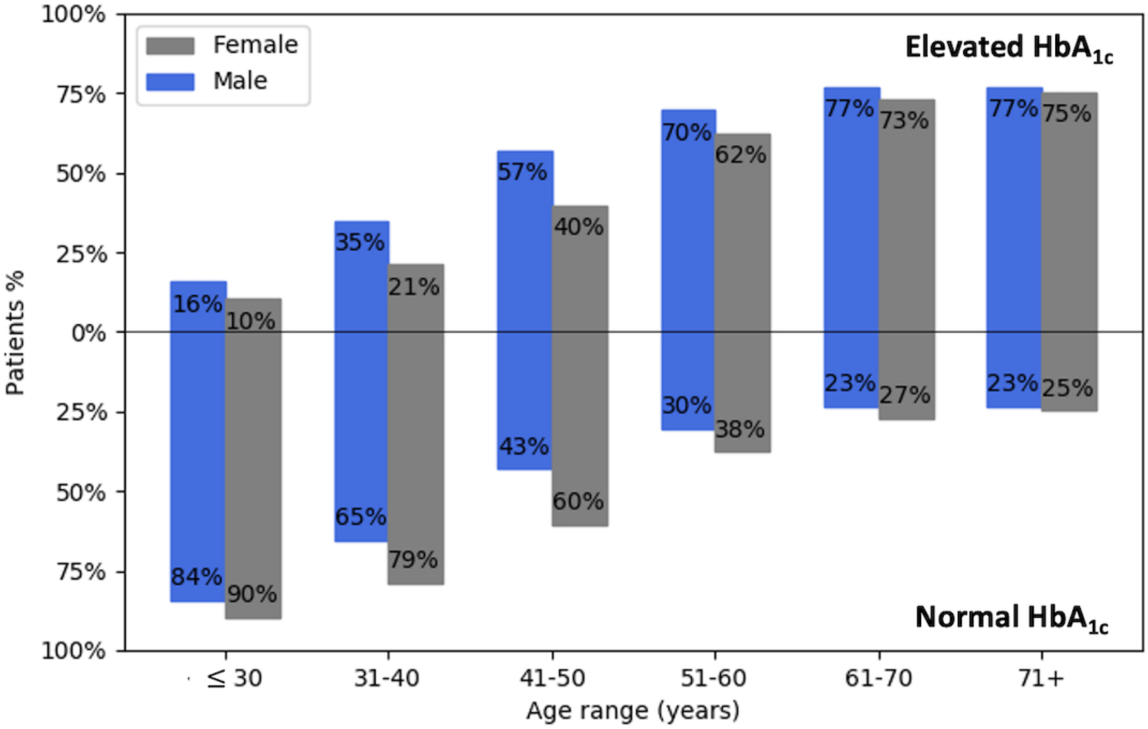


Table 1. Profile for the class distribution over gender.

Characteristics	HbA_{1c}[a] <5.7%, n/N (%)	HbA_{1c} ≥5.7%, n/N (%)
Number of patients (N=18,844)		
Total	9018/18,844 (47.86)	9826/18,844 (52.14)
Male	3764/9018 (41.74)	5544/9826 (56.42)
Female	5253/9018 (58.26)	4282/9826 (43.58)
Number of visits (N=157,600)		
Total	79,607/157,600 (50.51)	77,993/157,600 (49.49)
Male	31,620/79,607 (39.72)	41,591/77,993 (53.32)
Female	47,987/79,607 (60.28)	36,402/77,993 (46.68)

[a] HbA_{1c}: glycated hemoglobin.

Feature Selection and Data Sampling

Six main variables (features) were extracted from the KAIMRC EHR data set to be used in this study. These features, which were selected first for their theoretical association with hyperglycemia and second for their availability in the KAIMRC data set, were the following: age, BMI, estimated glomerular filtration rate (eGFR), RBS, total cholesterol, and non–high-density lipoprotein. The lab codes of the features used are available in Multimedia Appendix 1 Table S1. The descriptive statistics (using the data for the current visit only for unique patients), units, and *P* values for the selected features are presented in Table 2.

Table 2. Descriptive statistics of the selected features from the King Abdullah International Medical Research Center (KAIMRC) data set.

Feature	HbA$_{1c}$[a] 5.7%, mean (SD)	HbA$_{1c}$ 5.7%, mean (SD)	*P* value
Age (years)	43.94 (16.38)	58.92 (15.12)	<0.001
BMI (Kg/m^2)	29.11 (6.75)	30.90 (6.55)	<0.001
eGFR[b] (ml/min/1.73 m^2)	100.03 (29.22)	85.81 (28.239)	<0.001
RBS[c] (mmol/L)	5.45 (1.26)	7.88 (4.19)	<0.001
CHOL[d] mean (mmol/L)	4.65 (1.07)	4.42 (1.20)	<0.001
non-HDL[e] mean (mmol/L)	3.45 (1.01)	3.37 (1.115)	<0.001

[a] HbA$_{1c}$: glycated hemoglobin.

[b] eFGR: estimated glomerular filtration rate.

[c] RBS: random blood sugar.

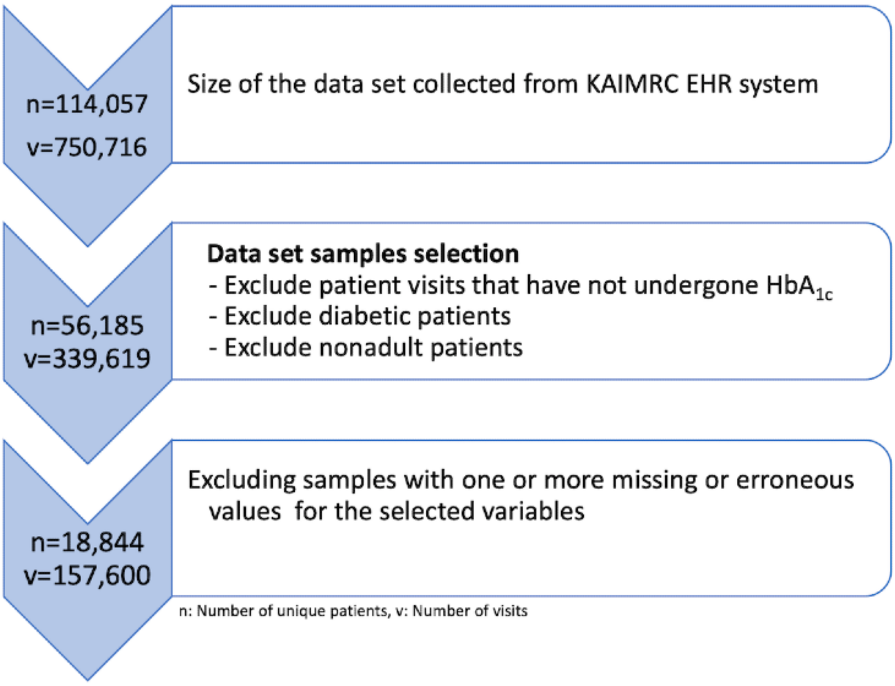
[d] CHOL: total cholesterol.

[e] non-HDL: non–high-density lipoprotein.

It is very common in clinical practice that physicians may require some lab tests and vital signs to be frequently recorded. In these cases, the average value of all readings taken on a given day (the basic time interval used for this study) was used. For inpatient visits, only data for the first day were considered, and, where there were missing values, the first available values from the visit were used.

For the purpose of this study, we aimed at predicting the HbA$_{1c}$ levels (≥5.7%) for current (last) patient visits only. Unlike the sampling approach used by Wells et al, which was based on independent hospital visits for patients (including for the same patients), the sampling approach used in this study included independent patients to ensure only unseen patients data were used for testing the models. Although we aimed to identify patients with elevated levels of HbA$_{1c}$ from a nondiabetic population, patients previously diagnosed with diabetes were excluded. We also excluded nonadult patients and those with erroneous or missing values [24]. Figure 2 shows the details of the tasks performed to refine the sample selection. This resulted in a reduction in the size of the experimental data set from 114,057 patients with 750,709 visits to 18,844 unique patients with 157,600 visits.

Figure 2. Details of the sampling approach performed on the KAIMRC data set. EHR: electronic health record; HbA$_{1c}$: glycated haemoglobin; KAIMRC: King Abdullah International Medical Research Center.


n=114,057
v=750,716
Size of the data set collected from KAIMRC EHR system
n=56,185
v=339,619
Data set samples selection
- Exclude patient visits that have not undergone HbA1c
- Exclude diabetic patients
- Exclude nonadult patients
n=18,844
v=157,600
Excluding samples with one or more missing or erroneous values for the selected variables
n: Number of unique patients, v: Number of visits

The inputs (input features space) for the models used in this study were continuous values. Values for age, eGFR, RBS and total cholesterol features were directly available in the KAIMRC data set. The values for the BMI and non–high-density lipoprotein variables were calculated from other available features using the formulae in Multimedia Appendix 2.

Input Preparation for the Models

The input structure for the deep learning model was organized as a matrix, based on current and previous time-stamped patient visits. It contained the current visit data concatenated with approximated values for the selected features from all previous visits, which we refer to as the “Approximated Time Series Data”.

Each patient visit was described by the selected features, represented as x_1, x_2 …, x_n. These features were formed as episodes based on the time-stamped values available in each visit (v_i).

Here, x_{ij} is the feature value at a patient visit ($0 < i \geq s$, $0 < j \geq n$); s is the number of time series steps (the length of the input sequence); and n is the number of features for each time step, which was set to 6 as explained earlier.

If the number of visits (longitudinal time series visits) for a patient was fewer than s, the input for this patient was padded out with the mean value of the available visits to compensate for the missing time series data (Multimedia Appendix 3 shows an example of the padding approach used). Where the number of longitudinal visits for a patient was more than s, the piecewise aggregation approximation (PAA) technique [38] was applied to the data for these visits to account for all data from patient visits.

PAA transforms the longitudinal time series data using s as a number of sliding windows (or segments) into a reduced number of time steps data (approximated) employing the mean value of the series falling within that window (segment) [39]. We tested the models with several values for the size of the sliding window (s), and 3 was shown to be the optimal value. The formula used to calculate the approximated time-series data was as follows:

Where represents the approximated value for x, r is the total number of visits for a patient, and s is the reduced number of time series steps (Multimedia Appendix 4 shows an example of the PAA technique used).

The approximated time series data forming the output of the PAA was then concatenated with the current visit data to form the final input for the deep learning model. As the MLR, RF, SVM, and LR models are not capable of handling multidimensional data (formed as matrices), the output of the PAA was reorganized for these into a single-dimensional input by vectorizing the matrix used in equation 1 as below:

$$\text{Input} = [x_{11}x_{12}x_{13}\ldots x_{sn}] \quad \textbf{(3)}$$

The last data preprocessing task before training the predictive models was data scaling. The experimental data set was scaled using the normalization technique that rescales the ranges of each of the features to be between 0 and 1 using minimum and maximum values of that feature.

Predictive Models and Experimental Setups

As a baseline comparison, we employed the MLR model used by Wells et al [21], and compared the results from this with those from 4 commonly used machine learning models.

The MLR model is used to create a mathematical equation that can best calculate the probability of a value by assigning weights (coefficients) to the independent variables (features) based on their importance [40]. In this study we employed the same approach used by Wells et al by which the continuous features were fitted into the MLR model using restricted cubic splines technique with 3 knots. When we used the longitudinal input, the variables that caused collinearity were excluded.

Random forest is an algorithm very commonly used for classification. It combines several decision trees that are generated during the training process. Each decision tree is trained using a random subset of the training data set. The final classification is then based on the majority voting results of all generated decision trees [41]. The quality function used in the employed RF model is the Gini importance, with a value of 100 for the number of tree parameters.

Logistic regression is commonly used to solve binary classification problems. It calculates the odds ratio of the variables and is similar to MLR but uses a binomial distribution of the dependent variable (ie, more than 1). Thus, it includes a logit function that handles different types of relationships between the dependent and independent variables [42,43].

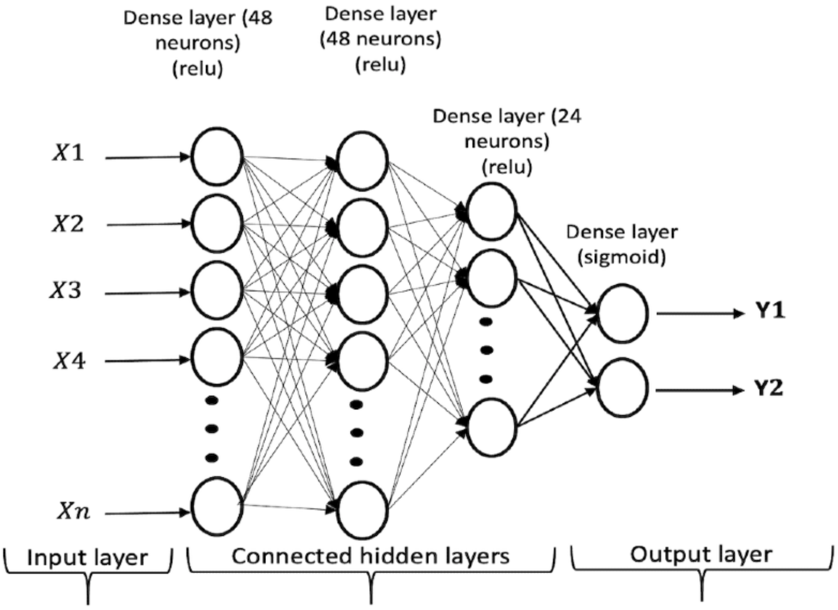
Support vector machine was introduced by Vapnik [44] in 1998. It can solve both classification and regression problems. It uses the training feature space to decide on the separation boundaries (hyperplane) that best divides the training data set into regions, 1 for each class. The very close points to the hyperplanes are the support vectors. SVMs also use kernels to help enhance class separation by mapping the training features into a higher dimensional space with an increased number of dimensions [44,45]. The kernel function used in the SVM model employed is a radial base function with a value of 1 for the cost parameter (C).

A multilayer perceptron, also known as a feed-forward neural network, is one of the most common deep learning approaches. It is mainly used to address supervised learning problems by learning the dependencies between the input layer (the features or variables) and output layer (the classification decision) using a fully connected hidden layer in between. The layers, including hidden ones, contain a number of neurons that are connected to the neurons of the next and previous layers via weights and nonlinear functions. MLP uses a backpropagation algorithm to update the weights and biases within the hidden layers to minimize the output error rate [25,46].

To optimize the MLP model, fine-tuning of the structure and hyperparameters was performed and involved the number of hidden layers and neurons, activation functions, optimizers, and loss functions. The optimized structure of the MLP model used in this study contained 3 hidden layers. The number of neurons in the hidden layers were 48, 48, and 24, respectively. The final layer (the output layer) contained 2 neurons for the final output of the model ($Y1$ for normal HbA_{1c} or $Y2$ for elevated HbA_{1c}).

A rectified linear unit activation function was used in the 3 hidden layers, while a sigmoid was used in the output layer. The detailed structure of the MLP model is shown in Figure 3. The model was trained using an Adam optimizer with mean squared error as the loss function.

Figure 3. The structure used for multilayer perceptron trained with the longitudinal data. relu: rectified linear unit.


Dense layer (48 neurons) (relu)
Dense layer (48 neurons) (relu)
Dense layer (24 neurons) (relu)
Dense layer (sigmoid)
X1
X2
X3
X4
Xn
Y1
Y2
Input layer
Connected hidden layers
Output layer

Evaluation of Model Performance

The models all employed the same data preprocessing, training, and testing techniques. The models were validated using the 10-fold cross-validation technique. The k-fold cross-validation is one of the most commonly used approximation approaches for validating the obtained results [47,48]. For the MLP model, 100 epochs were used to train each fold.

As our measure for evaluating and comparing the performance of the proposed models, we used the area under the receiver operating characteristic (AUC-ROC) curve, which is equal to the concordance statistic [49]. We also report values for a set of measures that are commonly used in clinical applications: balanced accuracy (that calculates the recall average for each class), overall accuracy, *F* score, precision, and precision-recall area under the curve (PR-AUC).

To determine the importance that the black box models (SVM and MLP) place upon each variable, we first computed the SHAP values and LIME scores for all samples in our data set and then calculated the average absolute SHAP value and LIME score for each predictor.

Results

Table 3 shows the performance metrics obtained using the MLR, RF, SVM, LR, and MLP models with and without the longitudinal data. The results show that the models achieved competitive performance using the reported measures. The LR and MLP models trained with and without the longitudinal data achieved better performance with regards to the AUC-ROC measure than did the MLR (statistical model employed by Wells et al) or the RF and SVM models (more details about AUC-ROC and PR-AUC curve plots are presented in Multimedia Appendix 5). The results also show that the SVM, LR, and MLP models trained with and without the longitudinal data achieved better performance than did the MLR and RF models using the balanced accuracy measure.

Table 3 also shows that all models, including the MLR, achieved better performance using all reported measures when they were trained with the features from patients' longitudinal data. The MLP with longitudinal data slightly outperformed all other models with respect to the reported measures.

Table 3. Classifiers performance for current glycated hemoglobin level prediction.

Model	AUC-ROC[a], % (SD)	Balanced accuracy, % (SD)	Accuracy, % (SD)	*F* score, % (SD)	Precision, % (SD)	PR-AUC[b], % (SD)
MLR[c]						
No[d]	81.38 (3.82)	72.74 (4.15)	73.59 (3.79)	74.91 (5.12)	73.20 (5.05)	82.14 (6.04)
Yes[e]	82.45 (4.09)	73.49 (4.19)	74.30 (4.02)	75.11 (6.00)	74.36 (5.26)	83.45 (6.29)
RF[f]						
No	80.82 (1.14)	72.57 (1.17)	72.64 (1.14)	73.97 (1.04)	73.42 (1.84)	82.03 (1.35)
Yes	82.38 (1.04)	73.86 (0.98)	73.91 (0.95)	75.07 (0.86)	74.81 (1.68)	84.06 (1.17)
SVM[g]						
No	81.05 (1.04)	73.69 (1.35)	73.88 (1.33)	75.76 (1.18)	73.42 (1.90)	80.56 (1.48)
Yes	82.04 (0.89)	74.25 (1.11)	74.40 (1.08)	76.08 (0.92)	74.20 (1.65)	83.16 (1.19)
LR[h]						
No	81.51 (1.26)	73.18 (1.10)	73.17 (1.08)	73.96 (1.03)	74.88 (1.69)	82.49 (1.46)
Yes	82.59 (1.04)	74.11 (1.15)	74.05 (1.13)	74.55 (0.98)	76.31 (1.72)	84.13 (1.04)
MLP[i]						
No	82.07 (1.06)	73.61 (1.04)	73.83 (1.03)	75.87 (1.10)	73.07 (1.62)	83.42 (1.19)
Yes	83.22 (0.92)	74.45 (1.18)	74.55 (1.18)	75.99 (1.95)	74.78 (2.07)	84.85 (0.78)

[a] AUC-ROC: area under the receiver operating characteristic.
[b] PR-AUC: precision-recall area under the curve.
[c] MLR: multiple logistic regression.
[d] Without longitudinal data.
[e] With longitudinal data.
[f] RF: random forest.
[g] SVM: support vector machine.
[h] LR: logistic regression.
[i] MLP: multilayer perceptron.

Figure 4 summarizes the 10-fold performance achieved for the set of measures where the models were trained without longitudinal data, and Figure 5 shows the performance where they were trained with the longitudinal data. Both figures show a more consistent prediction trend for RF, LR, SVM, and MLP with and without longitudinal data, as the measures for these models show a small variation between the folds. As shown in Figure 4 and Figure 5, the SD values for MLR with and without longitudinal data are larger than those for the other models. This indicates that the machine learning models used can not only enhance the performance, but can also improve the classification confidence for HbA_{1c} prediction.

Figure 4. Box plot showing the detailed 10-fold performance of all models trained without longitudinal data. AUR-ROC: area under the receiver operating characteristic; LR: logistic regression; MLP: multilayer perceptron; MLR: multiple logistic regression; PR-AUC: precision-recall area under the curve; RF: random forest; SVM: support vector machine.

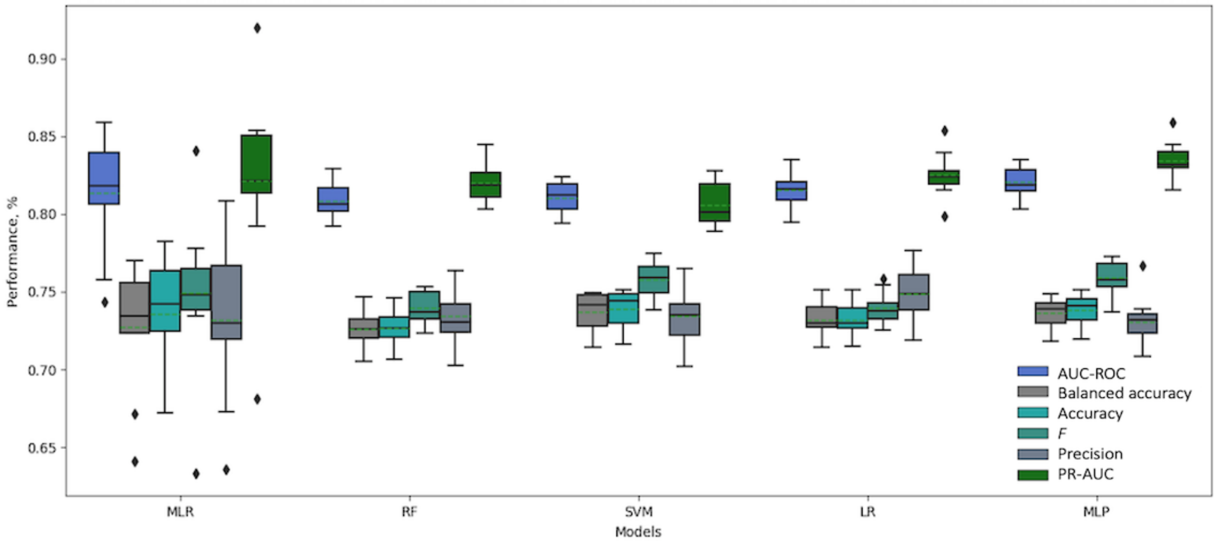


Figure 5. Boxplot showing the detailed 10-fold performance of all models trained with longitudinal data. AUR-ROC: area under the receiver operating characteristic; LR: logistic regression; MLP: multilayer perceptron; MLR: multiple logistic regression; PR-AUC: precision-recall area under the curve; RF: random forest; SVM: support vector machine.

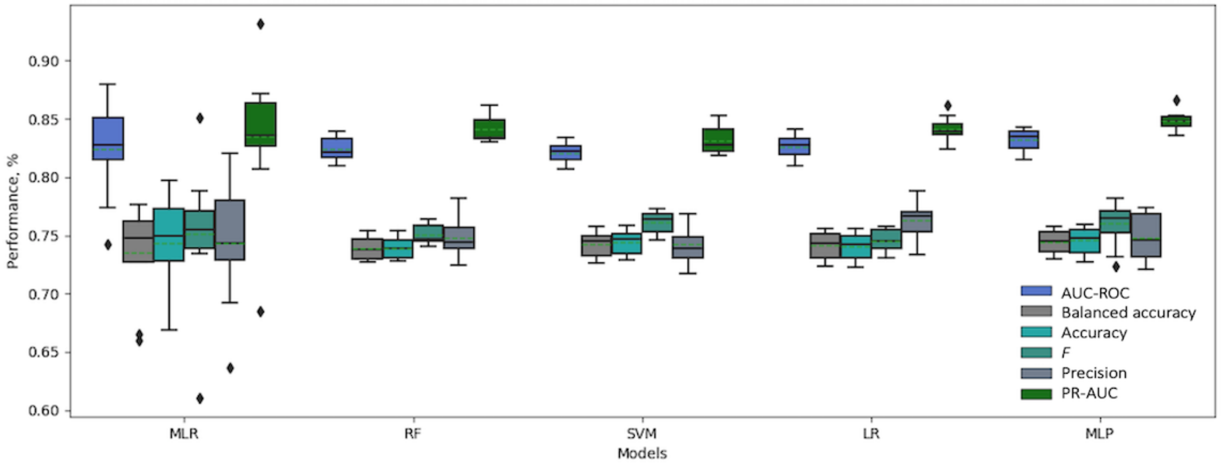


Table 4 shows the ranked order of importance of the set of predictors used for training the models. Further details on the actual importance values for each model are provided in Multimedia Appendix 6 (refer to Multimedia Appendix 7 for more details of the MLR and LR calculator). Calculating the importance of the predictors for the MLR models using vectorized longitudinal data was not possible due to the collinearity caused by having multiple variables for BMI. The order of importance results obtained using the SHAP method for both the SVM and MLP were identical to those obtained using LIME and provided greater confidence in the explainable methods used (see Multimedia Appendix 6).

Table 4. Order of importance of predictors for the models.

Model	Importance rank					
	1st	2nd	3rd	4th	5th	6th
MLR[a]						
No[b]	Age	RBS[c]	BMI	CHOL[d]	Non-HDL[e]	eGFR[f]
RF[g]						
No	Age	RBS	BMI	eGFR	CHOL	Non-HDL[h]
Yes[h]	RBS	Age	CHOL	eGFR	Non-HDL	BMI
LR[i]						
No	RBS	Age	Non-HDL	CHOL	BMI	eGFR
Yes	RBS	Age	Non-HDL	eGFR	CHOL	BMI
SVM[j] **(SHAP**[k] **& LIME**[l]**)**						
No	Age	RBS	BMI	Non-HDL	CHOL	eGFR
Yes	RBS	Age	CHOL	Non-HDL	BMI	eGFR
MLP[m] **(SHAP & LIME)**						
No	RBS	Age	Non-HDL	CHOL	BMI	eGFR
Yes	RBS	Age	eGFR	CHOL	Non-HDL	BMI

[a] MLR: multiple logistic regression.

[b] Without longitudinal data.

[c] RBS: random blood sugar.

[d] CHOL: total cholesterol.

[e] non-HDL: non–high-density lipoprotein.

[f] eGFR: estimated glomerular filtration rate.

[g] RF: random forest.

[h] With longitudinal data.

[i] LR: logistic regression.

[j] SVM: support vector machine.

[k] SHAP: Shapley Additive Explanations.

[l] LIME: local interpretable model-agnostic explanations.

[m] MLP: multilayer perceptron.

Table 4 and the figures in Multimedia Appendix 6 show that all of the models were heavily and interchangeably reliant on age and RBS when making classification decisions. The RF and SVM models, when trained with longitudinal data, ranked RBS over age. Figure 6 and Figure 7 highlight the importance that our best performing model, MLP, placed upon the features in our data set using SHAP and LIME, respectively. Both figures show that the RBS contributed the most to the MLP's final prediction, while the patient's BMI contributed the least.

Figure 6. Relative importance of predictors obtained from the multilayer perceptron trained with longitudinal data using SHAP. CHOL: total cholesterol; eGFR: estimated glomerular filtration rate; non-HDL: non–high-density lipoprotein; RBS: random blood sugar; SHAP: Shapley Additive Explanations.

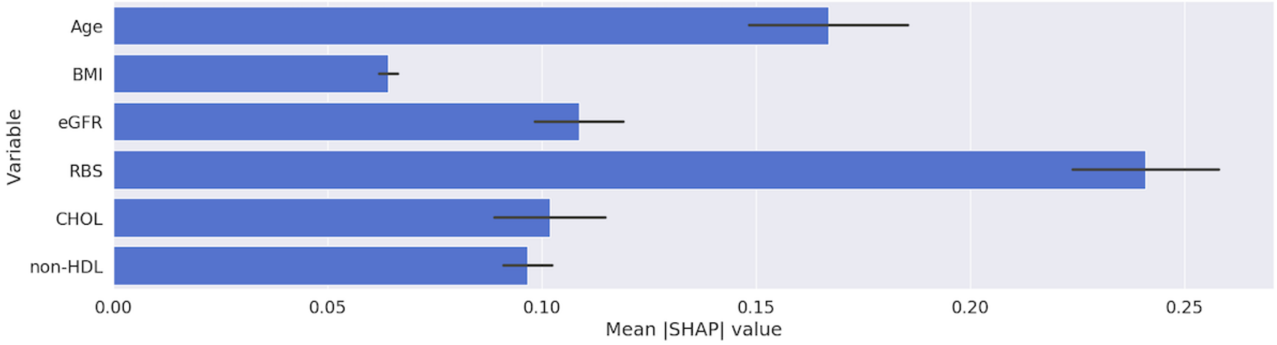
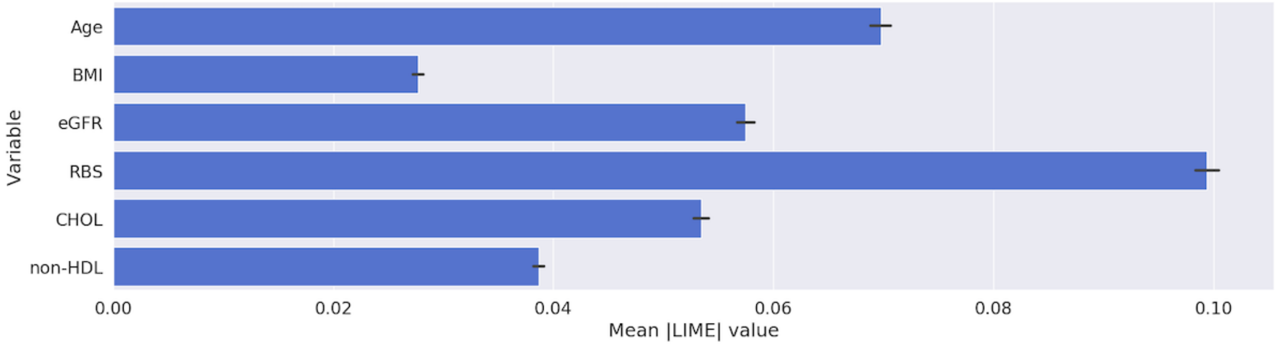


Figure 7. Relative importance of predictors obtained from multilayer perceptron trained with longitudinal data using LIME. CHOL: total cholesterol; eGFR: estimated glomerular filtration rate; LIME: local interpretable model-agnostic explanations; non-HDL: non–high-density lipoprotein; RBS: random blood sugar.



For all models trained with longitudinal data, BMI was ranked lower than when the models were trained without longitudinal data. However, the importance value produced for the BMI variable from the models was still not insignificant (see the figures in Multimedia Appendix 7). This indicates that models are able to find subtle relationships in the longitudinal data that are more relevant to the prediction than is BMI, rendering it less important.

When MLP and LR models trained on the longitudinal data were used, the eGFR variable was ranked higher than total cholesterol and BMI, in contrast to when these were trained on the current visit only. None of the other models trained with the current visit only, except for RF, considered it important. Again, we ascribe this to the information that the model learns from the variations of eGFR values between a patient's visits (longitudinal EHR data).

SHAP values are calculated on the sample level. Figures 8 and 9 illustrate the SHAP values for 2 randomly selected sample patients from our data set. These figures highlight how different inputs have different SHAP values. The patient in Figure 8 (for whom our model correctly predicted elevated HbA_{1c} levels of ≥5.7%) had a higher RBS value than did the patient in Figure 9 (for whom our model correctly predicted normal HbA_{1c} levels of <5.7%). This explains why our MLP model placed much more importance on the RBS value of the patient in Figure 6.

Figure 8. An example showing the SHAP values for a randomly selected sample with elevated glycated hemoglobin levels (≥5.7%). CHOL: total cholesterol; eGFR: estimated glomerular filtration rate; non-HDL: non–high-density lipoprotein; RBS: random blood sugar; SHAP: Shapley Additive Explanations.

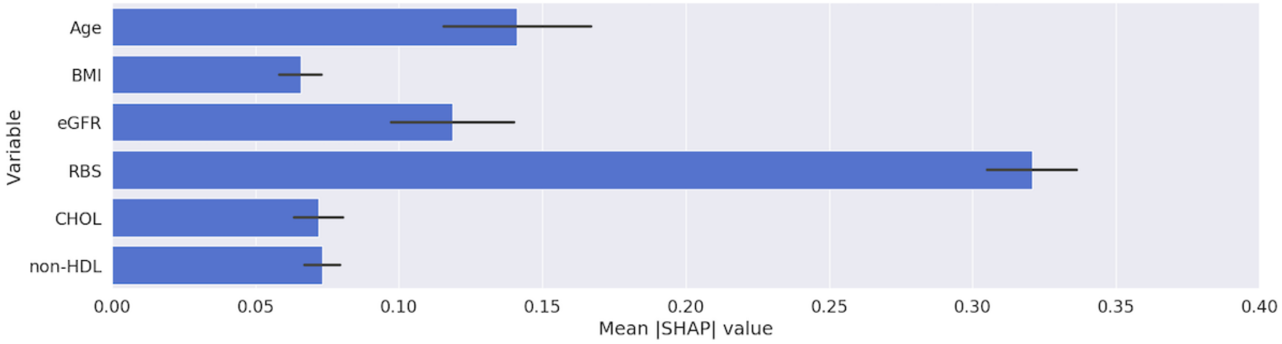
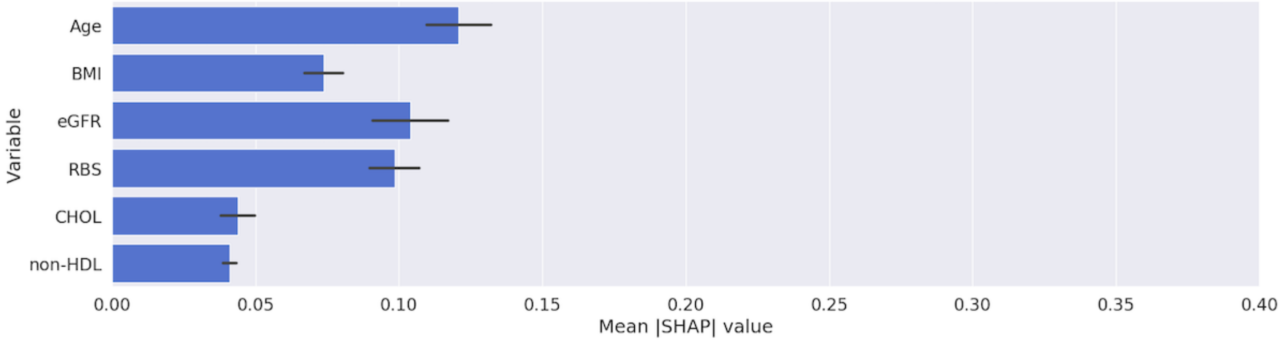
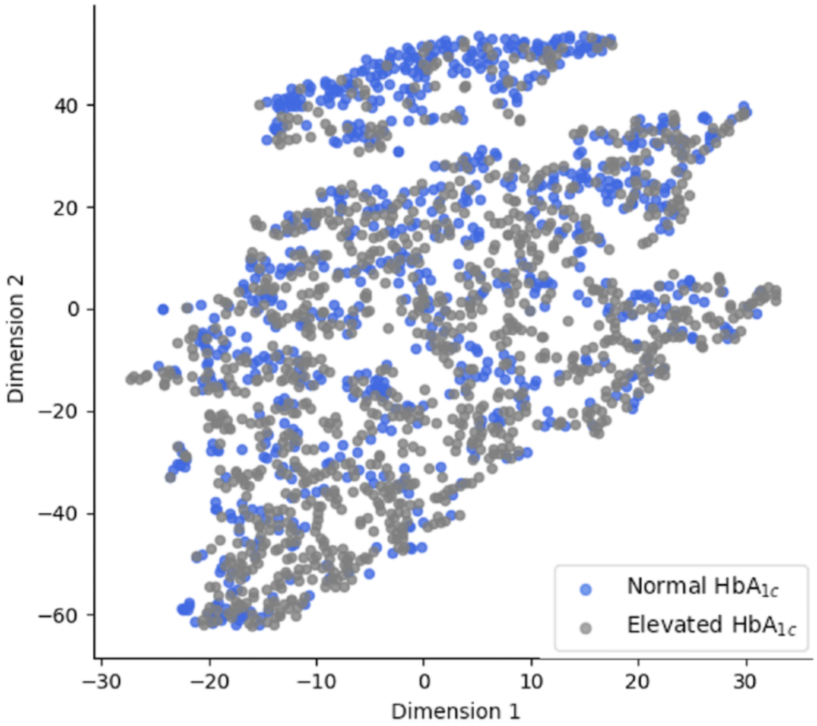


Figure 9. An example showing the SHAP values for randomly selected sample with normal glycated hemoglobin levels (<5.7%). CHOL: total cholesterol; eGFR: estimated glomerular filtration rate; non-HDL: non–high-density lipoprotein; RBS: random blood sugar; SHAP: Shapley Additive Explanations.



The task of predicting HbA_{1c} elevation risk can be challenging. Figure 10 provides a visualization of the data points for the 2 classes (prediabetic with ≥5.7%; normal with <5.7%) after mapping of the data points (for the test data) into 2 dimensions with t-distributed stochastic neighbor embedding was performed [50]. The overlap in the data points visualized in the figure demonstrates the challenge of separating the patients with and without elevated levels of HbA_{1c} (≥5.7%) in the KAIMRC data set. We avoided intensive feature engineering techniques in the sampling approach used. However, the approaches adopted were able to achieve promising results with an accuracy of 83.22% for the AUC-ROC using MLP with historical data.

Figure 10. Two-dimensional visualization using t-distributed stochastic neighbor embedding for a randomly selected subset of the data. HbA_{1c}: glycated hemoglobin.



In summary, all models showed promising results for predicting the current HbA_{1c} elevation levels (≥5.7%) with EHR data. The results emphasize that the HbA_{1c} predictive models can exhibit more learnability when they are trained with the longitudinal patient data observations typically available from EHR systems.

Discussion

Strengths and Limitations

EHR systems were adopted for the purpose of improving health care outcomes and were not originally intended for research purposes [19]. Patient data stored in EHR systems can be obtained at irregular intervals, as lab instructions are carried out with different frequencies based on the physician's decisions and a patient’s visit patterns. It is very common that medical data extracted from EHR systems suffer from problems such as irregularity, incompleteness, and noisy and imbalanced data [13]. These can be challenging obstacles for any technology used for predictive analytics.

In our study, the sampling approach used did not affect the balanced nature of the data set used. As shown in Figure 2, there were 56,185 unique patients present before removal of the records with 1 or more missing values. The number of unique patients with elevated HbA_{1c} levels (≥5.7%) before removal of the incomplete records was 27,354, resulting in a retention of 48.68% (27,354/56,185). The number of unique patients with normal HbA_{1c} levels was 28,831, resulting in a retention of 51.32% (28,831/56,185). We would argue that the absence or the presence of the HbA_{1c} readings is not random, as the sample was collected from the population of Saudi Arabia and thus the likelihood of a patient taking an HbA_{1c} test is large because of the prevalence of diabetes in this country [51]. This may affect the reproducibility of this work using different populations from different countries especially those with lower rates of diabetes.

It is hoped that these outcomes will encourage further investigation into the predictability of current HbA_{1c} levels (≥5.7%) using more of the readings normally provided in EHR data. For example, other important readings such as FBS and triglycerides have shown clinical correlations with diabetes [52]. In addition, our data set contained only 3 years of patient data, which limits the number of patient visits recorded. Figure 11 shows the number of visits made by patients from 2016 to 2018, while Figure 12 details the number of visits made by patients (after removal of the outliers) over HbA_{1c} levels. Both figures show that the majority of the patients have made relatively few visits: 52% (8713/16818) of the patients made 4 visits or fewer over the 3 years (1.3 visit per year). This also justifies the size of the sliding window (*s*=3) as the optimal input size for the models used. However, we hypothesize that the longitudinal behavior of the features used can be enriched by including more values obtained over longer periods. Therefore, incorporating more features and their longitudinal behavior over longer periods into the models used in this study would likely improve the prediction performance of our chosen models.

Figure 11. Histogram showing the trend in the number of visits made by patients.

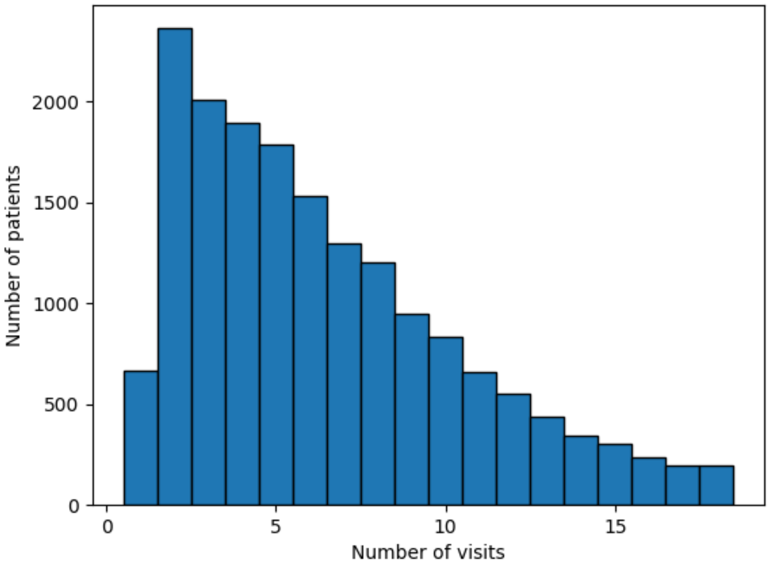
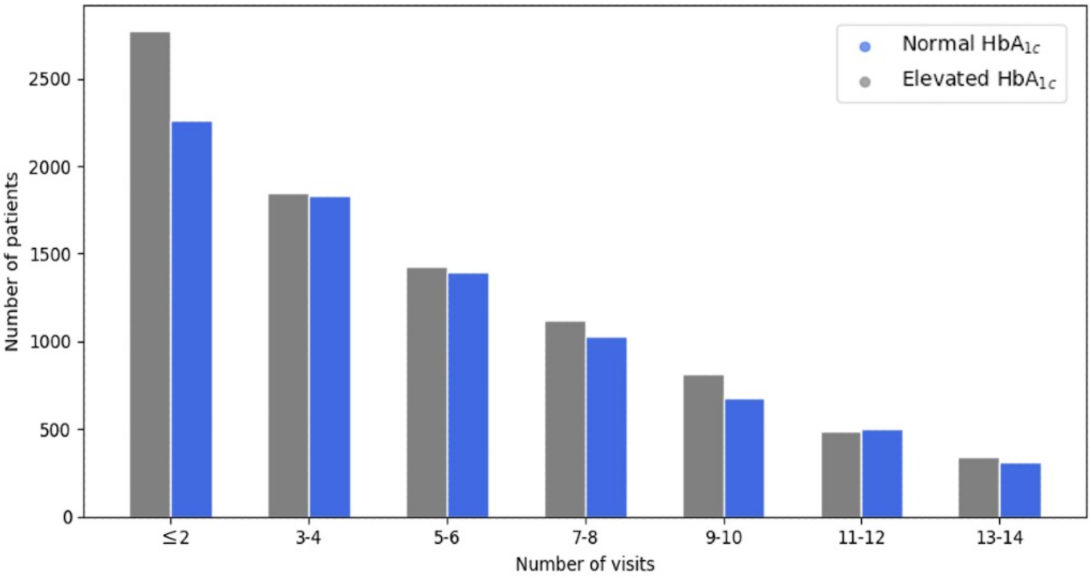


Figure 12. The details for the number of visits made over number of patients. HbA_{1c}: glycated hemoglobin.



Variations in the data or model produce slightly different attribution values. However, due to the critical nature of many health care applications, it is always important to verify that the models make “sensible” predictions. Without the use of SHAP/LIME, this would be hard to verify for any nonlinear model. Although it is possible to see that the models have high performance, we would be unable to verify that a model is not making spurious correlations. Furthermore, through the use of SHAP, we can verify that MLPs trained on the longitudinal data are learning to use the extra information contained in the longitudinal data (as indicated by the higher importance of eGFR), allowing us to pinpoint the reason these models gain higher performance.

To investigate the effect of temporal dependencies in the data, this study investigated the use of other deep learning models along with the MLP, including long short-term memory (LSTM) and bidirectional LSTM [25,53] for HbA_{1c} prediction. Table 5 reports the results of using these models. The MLP model achieved similar performance to the LSTM and bidirectional LSTM models according to all reported measures. This suggests that directly modeling the temporal dynamics in the data is not very helpful. This could be due to the short lengths of the time series or a too-weak temporal dependency.

Table 5. LSTM and BiLSTM Classifiers performance trained with longitudinal data for current HbA_{1c} levels prediction.

Model	AUC-ROC[a], % (SD)	Balanced Accuracy, % (SD)	Accuracy, % (SD)	*F* score, % (SD)	Precision, % (SD)	PR-AUC[b], % (SD)
LSTM[c]	83.26% (0.91)	74.17% (1.05)	74.59% (1.23)	75.64% (1.50)	74.59% (3.26)	81.88% (0.95)
BiLSTM[d]	83.16% (0.87)	74.21% (1.24)	74.30% (1.15)	75.46% (1.39)	75.19% (2.36)	84.75% (0.75)

[a] AUC-ROC: area under the receiver operating characteristic.

[b] PR-AUC: precision-recall area under the curve.

[c] LSTM: long short-term memory.

[d] BiLSTM: bidirectional LSTM.

Generalizing our findings using other data sets is challenging because of the accessibility and privacy restrictions that apply to medical data sets. For this reason, and because of the lack of similar studies that have used machine learning for HbA_{1c} prediction with EHR data, comparing the performance achieved by the models outlined in this study with those developed by other researchers will require the availability of alternative anonymized data sets.

Conclusions

We believe that this study is the first to investigate the performance of machine learning models used with EHR data for predicting current HbA_{1c} elevation risk (≥5.7%) for nondiabetic patients. It is also the first to investigate employing the longitudinal data that are normally stored on EHR systems to enhance the prediction of HbA_{1c} elevation levels. Our findings show that the MLP model achieves better results when a patient's longitudinal data are combined with current visit data, and the use of longitudinal data also affects the relative importance for the predictors used.

As this work formed a continuation of previous work [24], we avoided changing the sampling approach used. However, studying the impact of applying different sampling approaches could be valuable to explore in future work as would the use of a larger data set with more variables and the recording of longitudinal behavior over longer periods.

Acknowledgments

We would like to acknowledge the contribution the KAIMRC for providing the data set under the approved projects: Diabetes Early Warning System (research protocol no. SP14/042), Finding the Common Related Diseases with Diabetes using Data Mining Association Techniques (research protocol no. SP15/064,) and extension project (no. RYD-17-417780-187503) to collect the newest data set. The authors would also like to thank Cievert Ltd and the European Regional Development Fund for sponsoring this work.

Authors' Contributions

ZA was responsible for implementing and building predictive models. ZA, MW, DB, and NAM were responsible for the design of the study and for writing the manuscript. ZA, MW, DB, and NAM were responsible for designing and validating the models. MW and ZA were responsible for analyzing the explainability of the machine learning model. ZA, AA, and RA were responsible for extracting and describing the data set. All authors participated in reviewing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Lab test and diagnostic codes.

[PDF File (Adobe PDF File), 93 KB - medinform_v9i5e25237_app1.pdf]

Multimedia Appendix 2

Formulae for the calculated variables.

[PDF File (Adobe PDF File), 77 KB - medinform_v9i5e25237_app2.pdf]

Multimedia Appendix 3

An example of the padding approach used.

[PDF File (Adobe PDF File), 169 KB - medinform_v9i5e25237_app3.pdf]

Multimedia Appendix 4

An example of the PAA technique.

[PDF File (Adobe PDF File), 240 KB - medinform_v9i5e25237_app4.pdf]

Multimedia Appendix 5
AUC-ROC and PR-AUC curves for the models (with 10 folds) trained with longitudinal data.
[PDF File (Adobe PDF File), 1011 KB - medinform_v9i5e25237_app5.pdf]

Multimedia Appendix 6
Variable relative importance charts for the models.
[PDF File (Adobe PDF File), 578 KB - medinform_v9i5e25237_app6.pdf]

Multimedia Appendix 7
Multiple logistic regression (MLR) and logistic regression (LR) details.
[PDF File (Adobe PDF File), 157 KB - medinform_v9i5e25237_app7.pdf]

References

1. Larsen ML, Hørder M, Mogensen EF. Effect of long-term monitoring of glycosylated hemoglobin levels in insulin-dependent diabetes mellitus. New England Journal of Medicine 1990 Oct 11;323(15):1021-1025. [doi: 10.1056/NEJM199010113231503] [Medline: 2215560]
2. Pradhan AD, Rifai N, Buring JE, Ridker PM. Hemoglobin A1c predicts diabetes but not cardiovascular disease in nondiabetic women. The American Journal of Medicine 2007 Aug;120(8):720-727. [doi: 10.1016/j.amjmed.2007.03.022] [Medline: PMC2585540]
3. Ackermann RT, Cheng YJ, Williamson DF, Gregg EW. Identifying adults at high risk for diabetes and cardiovascular disease using hemoglobin A1c: National Health and Nutrition Examination Survey 2005-2006. American Journal of Preventive Medicine 2011 Jan;40(1):11-17. [doi: 10.1016/j.amepre.2010.09.022] [Medline: 21146762]
4. World Health Organization. Use of glycated haemoglobin (HbA1c) in diagnosis of diabetes mellitus: abbreviated report of a WHO consultation. World Health Organization 2011:a. [Medline: 26158184]
5. Khaw K, Wareham N, Bingham S, Luben R, Welch A, Day N. Association of hemoglobin A1c with cardiovascular disease and mortality in adults: the European prospective investigation into cancer in Norfolk. Ann Intern Med 2004 Sep 21;141(6):413. [doi: 10.7326/0003-4819-141-6-200409210-00006] [Medline: 15381514]
6. American Diabetes Association. Classification and diagnosis of diabetes: standards of medical care in diabetes—2018. Dia Care 2017 Dec 08;41(Supplement 1):S13-S27. [doi: 10.2337/dc18-s002]
7. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, et al. Electronic health records: new opportunities for clinical research. Journal of internal medicine 2013 Oct 18;274(6):547-560. [doi: 10.1111/joim.12119] [Medline: 23952476]
8. McKinney BA, Reif DM, Ritchie MD, Moore JH. Machine learning for detecting gene-gene interactions: a review. Appl Bioinformatics 2006;5(2):77-88 [FREE Full text] [doi: 10.2165/00822942-200605020-00002] [Medline: 16722772]
9. Goldenberg SL, Nir G, Salcudean SE. A new era: artificial intelligence and machine learning in prostate cancer. Nature Reviews Urology 2019 May 15;16(7):391-403. [doi: 10.1038/s41585-019-0193-3] [Medline: 31092914]
10. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. Summit Transl Bioinform 2010:1. [Medline: 21347133]
11. Perveen S, Shahbaz M, Keshavjee K, Guergachi A. Prognostic modeling and prevention of diabetes using machine learning technique. Scientific reports 2019 Sep 24;9(1):1. [doi: 10.1038/s41598-019-49563-6] [Medline: 31551457]
12. Esteban S, Rodríguez Tablado M, Peper FE, Mahumud YS, Ricci RI, Kopitowski KS, et al. Development and validation of various phenotyping algorithms for Diabetes Mellitus using data from electronic health records. Computer Methods and Programs in Biomedicine 2017 Dec;152:53-70. [doi: 10.1016/j.cmpb.2017.09.009] [Medline: 29054261]
13. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Scientific reports 2016 May 17;6(1):1-10. [doi: 10.1038/srep26094] [Medline: 27185194]
14. Hippisley-Cox J, Coupland C, Robson J, Sheikh A, Brindle P. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. BMJ 2009 Mar 17;338(mar17 2):b880-b880. [doi: 10.1136/bmj.b880] [Medline: 19297312]
15. Alhassan Z, McGough A, Alshammari R, Daghstani T, Budgen D, Al MN. Type-2 diabetes mellitus diagnosis from time series clinical data using deep learning models. 2018 Presented at: International Conference on Artificial Neural Networks; 2018 Oct 4-7; Greece. [doi: 10.1007/978-3-030-01424-7_46]
16. McCarter RJ, Hempe JM, Chalew SA. Mean blood glucose and biological variation have greater influence on HbA1c levels than glucose instability: an analysis of data from the Diabetes Control and Complications Trial. Diabetes Care 2006 Jan 27;29(2):352-355. [doi: 10.2337/diacare.29.02.06.dc05-1594] [Medline: 16443886]
17. Nathan DM, Kuenen J, Borg R, Zheng H, Schoenfeld D, Heine RJ. Translating the A1C assay into estimated average glucose values. Diabetes Care 2008 Jun 07;31(8):1473-1478. [doi: 10.2337/dc08-0545] [Medline: 18540046]

18. Rose E, Ketchell D. Clinical inquiries. Does daily monitoring of blood glucose predict hemoglobin A1c levels? J Fam Pract 2003:1. [Medline: 12791231]
19. Xu S, Schroeder EB, Shetterly S, Goodrich GK, O'Connor PJ, Steiner JF, et al. Accuracy of hemoglobin A1c imputation using fasting plasma glucose in diabetes research using electronic health records data. Stat., optim. inf. comput 2014 Jun 01;2(2):93-104. [doi: 10.19139/68]
20. Rauh SP, Heymans MW, Koopman ADM, Nijpels G, Stehouwer CD, Thorand B, et al. Predicting glycated hemoglobin levels in the non-diabetic general population: Development and validation of the DIRECT-DETECT prediction model - a DIRECT study. PLoS ONE 2017 Feb 10;12(2):e0171816. [doi: 10.1371/journal.pone.0171816] [Medline: 28187151]
21. Wells BJ, Lenoir KM, Diaz-Garelli J, Futrell W, Lockerman E, Pantalone KM, et al. Predicting current glycated hemoglobin values in adults: development of an algorithm from the electronic health record. JMIR Med Inform 2018 Oct 22;6(4):e10780. [doi: 10.2196/10780] [Medline: 30348631]
22. Baan CA, Ruige JB, Stolk RP, Witteman JC, Dekker JM, Heine RJ, et al. Performance of a predictive model to identify undiagnosed diabetes in a health care setting. Diabetes Care 1999 Feb 01;22(2):213-219. [doi: 10.2337/diacare.22.2.213] [Medline: 10333936]
23. Griffin SJ, Little PS, Hales CN, Kinmonth AL, Wareham NJ. Diabetes risk score: towards earlier detection of Type 2 diabetes in general practice. Diabetes/metabolism research and reviews 2000 May;16(3):164-171. [doi: 10.1002/1520-7560(200005/06)16:3<164::aid-dmrr103>3.0.co;2-r] [Medline: 10867715]
24. Alhassan Z, Budgen D, Alshammari R, Al Moubayed N. Predicting current glycated hemoglobin levels in adults from electronic health records: validation of multiple logistic regression algorithm. JMIR Med Inform 2020 Jul 3;8(7):e18963. [doi: 10.2196/18963] [Medline: 32618575]
25. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 27;521(7553):436-444. [doi: 10.1038/nature14539]
26. Ahmad M, Eckert C, Teredesai A. Interpretable machine learning in healthcare. 2018 Presented at: Proceedings of the ACM international conference on bioinformatics, computational biology, and health informatics; 2018 Aug 29-Sept 1; Washington DC. [doi: 10.1145/3233547.3233667]
27. Lipton ZC. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. ACM 2018 Jun;16(3):31-57. [doi: 10.1145/3236386.3241340]
28. Lundberg S, Lee SI. A unified approach to interpreting model predictions. 2017 Presented at: Advances in neural information processing systems; 2017 Dec 4-9; Long Beach.
29. Ribeiro M, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. 2016 Presented at: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016 Aug 13-16; San Francisco. [doi: 10.1145/2939672.2939778]
30. Abdulaziz Al Dawish M, Alwin Robert A, Braham R, Abdallah Al Hayek A, Al Saeed A, Ahmed Ahmed R, et al. Diabetes mellitus in Saudi Arabia: a review of the recent literature. Current diabetes reviews 2016 Oct 26;12(4):359-368. [doi: 10.2174/1573399811666150724095130] [Medline: 26206092]
31. Understanding A1C. American Diabetes Association. URL: https://www.diabetes.org/a1c [accessed 2020-11-07]
32. Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter 2004 Jun;6(1):20-29. [doi: 10.1145/1007730.1007735]
33. Zhang L, Yang H, Jiang Z. Imbalanced biomedical data classification using self-adaptive multilayer ELM combined with dynamic GAN. BioMedical Engineering OnLine volume 2018 Dec 4;17(1):1. [doi: 10.1186/s12938-018-0604-3] [Medline: 30514298]
34. Rahman MM, Davis DN. Addressing the class imbalance problem in medical datasets. International Journal of Machine Learning and Computing 2013:224-228. [doi: 10.7763/ijmlc.2013.v3.307]
35. Longadge R, Dongre S, Malik L. Class imbalance problem in data mining review. IJCSN 2013;2(1):1-7.
36. Alhassan Z, Budgen D, Alshammari R, Daghstani T, McGough A, Al MN. Stacked denoising autoencoders for mortality risk prediction using imbalanced clinical data. 2018 Presented at: International Conference on Machine Learning and Applications (ICMLA); 2018 Dec 17; Orlando. [doi: 10.1109/icmla.2018.00087]
37. Alqurashi KA, Aljabri KS, Bokhari SA. Prevalence of diabetes mellitus in a Saudi community. Annals of Saudi Medicine 2011 Jan;31(1):19-23. [doi: 10.4103/0256-4947.75773] [Medline: 21245594]
38. Keogh E, Chakrabarti K, Pazzani M, Mehrotra S. Locally adaptive dimensionality reduction for indexing large time series databases. 2001 Presented at: The 2001 ACM SIGMOD International Conference on Management of Data; 2001 May 21-25; Santa Barbara. [doi: 10.1145/375663.375680]
39. Zhao J, Papapetrou P, Asker L, Boström H. Learning from heterogeneous temporal data in electronic health records. Journal of Biomedical Informatics 2017 Jan;65:105-119. [doi: 10.1016/j.jbi.2016.11.006] [Medline: 27919732]
40. McDonald J. Handbook of Biological Statistics. Baltimore, MD: Sparky House Publishing; 2009.
41. Breiman L. Random forests. Machine learning 2001;45(1):5-32.
42. Rawlings J, Pantula S, Dickey D. Applied Regression Analysis. New York: Springer; 2001:a.
43. Sperandei S. Understanding logistic regression analysis. Biochemia Medica 2014:12-18. [doi: 10.11613/bm.2014.003] [Medline: 24627710]
44. Vapnik V. The Nature of Statistical Learning Theory. New York: Springer; 2013.

45. Noble WS. What is a support vector machine? Nature Biotechnol 2006 Dec;24(12):1565-1567. [doi: 10.1038/nbt1206-1565]
46. Gardner M, Dorling S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric Environment 1998 Aug;32(14-15):2627-2636. [doi: 10.1016/s1352-2310(97)00447-0]
47. Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep Learning. Cambridge, MA: MIT Press; 2016.
48. Bobadilla J, Ortega F, Hernando A, Gutiérrez A. Recommender systems survey. Knowledge-Based Systems 2013 Jul;46:109-132. [doi: 10.1016/j.knosys.2013.03.012]
49. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. BMC medical research methodology 2012 Jun 20;12(1):109-132. [doi: 10.1186/1471-2288-12-82] [Medline: 22716998]
50. Maaten L, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. (Nov) 2008;9:2579-2605.
51. Al-Zahrani J, Aldiab A, Aldossari K, Al-Ghamdi S, Batais M, Javad S. Prevalence of prediabetes, diabetes and its predictors among females in Alkharj, Saudi Arabia: a cross-sectional study. Annals of Global Health 2019;85(1):A. [doi: 10.5334/aogh.2467] [Medline: 31348623]
52. Naqvi S, Naveed S, Ali Z, Ahmad S, Khan R, Raj H. Correlation between glycated hemoglobin and triglyceride level in type 2 diabetes mellitus. Cureus 2017;9(6):1. [doi: 10.7759/cureus.1347] [Medline: 28713663]
53. Schuster M, Paliwal K. Bidirectional recurrent neural networks. IEEE Trans. Signal Process 1997;45(11):2673-2681. [doi: 10.1109/78.650093]

Abbreviations

AUR-ROC: area under the receiver operating characteristic
eGFR: estimated glomerular filtration rate
EHR: electronic health records
FBS: fasting blood sugar
HbA$_{1c}$: glycated hemoglobin
KAIMRC: King Abdullah International Medical Research Center
LIME: local interpretable model-agnostic explanations
LR: logistic regression.
LSTM: long short-term memory
MLP: multilayer perceptron
MLR: multiple logistic regression
PAA: piecewise aggregation approximation
PR-AUC: precision-recall area under the curve
RBS: random blood sugar
RF: random forest
SHAP: Shapley Additive Explanations
SVM: support vector machine
T2DM: type-2 diabetes mellitus
WHO: World Health Organization

Edited by C Lovis; submitted 23.10.20; peer-reviewed by S Veeranki, F Agakov, C Doogan; comments to author 13.11.20; revised version received 05.01.21; accepted 22.04.21; published 24.05.21.

Please cite as:
Alhassan Z, Watson M, Budgen D, Alshammari R, Alessa A, Al Moubayed N
Improving Current Glycated Hemoglobin Prediction in Adults: Use of Machine Learning Algorithms With Electronic Health Records
JMIR Med Inform 2021;9(5):e25237
URL: https://medinform.jmir.org/2021/5/e25237
doi:10.2196/25237
PMID:34028357

Original Paper

Effective Training Data Extraction Method to Improve Influenza Outbreak Prediction from Online News Articles: Deep Learning Model Study

Beakcheol Jang[1], PhD; Inhwan Kim[1], BSc; Jong Wook Kim[2], PhD

[1]Graduate School of Information, Yonsei University, Seoul, Republic of Korea

[2]Department of Computer Science, Sangmyung Univerisity, Seoul, Republic of Korea

Corresponding Author:
Jong Wook Kim, PhD
Department of Computer Science
Sangmyung Univerisity
20, Hongjimun 2-gil, Jongno-gu
Seoul, 03016
Republic of Korea
Phone: 82 027817590
Fax: 82 0222870072
Email: jkim@smu.ac.kr

Abstract

Background: Each year, influenza affects 3 to 5 million people and causes 290,000 to 650,000 fatalities worldwide. To reduce the fatalities caused by influenza, several countries have established influenza surveillance systems to collect early warning data. However, proper and timely warnings are hindered by a 1- to 2-week delay between the actual disease outbreaks and the publication of surveillance data. To address the issue, novel methods for influenza surveillance and prediction using real-time internet data (such as search queries, microblogging, and news) have been proposed. Some of the currently popular approaches extract online data and use machine learning to predict influenza occurrences in a classification mode. However, many of these methods extract training data subjectively, and it is difficult to capture the latent characteristics of the data correctly. There is a critical need to devise new approaches that focus on extracting training data by reflecting the latent characteristics of the data.

Objective: In this paper, we propose an effective method to extract training data in a manner that reflects the hidden features and improves the performance by filtering and selecting only the keywords related to influenza before the prediction.

Methods: Although word embedding provides a distributed representation of words by encoding the hidden relationships between various tokens, we enhanced the word embeddings by selecting keywords related to the influenza outbreak and sorting the extracted keywords using the Pearson correlation coefficient in order to solely keep the tokens with high correlation with the actual influenza outbreak. The keyword extraction process was followed by a predictive model based on long short-term memory that predicts the influenza outbreak. To assess the performance of the proposed predictive model, we used and compared a variety of word embedding techniques.

Results: Word embedding without our proposed sorting process showed 0.8705 prediction accuracy when 50.2 keywords were selected on average. Conversely, word embedding using our proposed sorting process showed 0.8868 prediction accuracy and an improvement in prediction accuracy of 12.6%, although smaller amounts of training data were selected, with only 20.6 keywords on average.

Conclusions: The sorting stage empowers the embedding process, which improves the feature extraction process because it acts as a knowledge base for the prediction component. The model outperformed other current approaches that use flat extraction before prediction.

(JMIR Med Inform 2021;9(5):e23305) doi:10.2196/23305

KEYWORDS

influenza; training data extraction; keyword; sorting; word embedding; Pearson correlation coefficient; long short-term memory; surveillance; infodemiology; infoveillance; model

Introduction

Influenza is a highly contagious disease that affects 3 to 5 million people and kills 290,000 to 650,000 worldwide each year [1]. To track and counter its effects, various countries have established influenza surveillance systems, such as the European Influenza Surveillance Scheme in Europe and the Centers for Disease Control and Prevention (CDC) in the United States. These mechanisms provide clinical data, such as physician visits with influenza-like illness (ILI). However, a proper extraction of actionable insights is hindered by a delay of approximately 2 weeks for such information to become available. To solve this problem, studies in the field of infodemiology [2,3] have been trying to gain novel and effective insights into diseases from internet-based data. Hence, various recent studies in infodemiology have attempted to deter this time delay to predict impending outbreaks by monitoring influenza in real time using cloud-sourced data, such as online news articles and social network services [4-9].

Key studies have been conducted on influenza prediction systems based on search queries, including Google Flu Trends [10,11], in which Google provided surveillance and prediction services for influenza using search queries [2,10,12-16]. Twitter has recently received significant attention as a potential source of data for the prediction of influenza outbreaks. The number of studies that leverage tweets to predict influenza has multiplied and they have achieved moderately accurate prediction accuracy [17-23]. The advantage of predicting a fast-spreading outbreak via social network data (such as Twitter) is the speed at which people can share the news, hence providing a prompt opportunity to use an analytical system to predict a serious outbreak. However, various obstacles—such as privacy issues for search query data—hinder the real-time prediction because of the failure to capture the inherent features of the data [24]. In addition, the tweets are created by amateur users and are prone to noise due to poor writing standards, typographical errors, use of jargon expressions, and meaningless content [19,25].

Previous studies have used these web data to surveil influenza outbreaks and improved predictive performance, but the problem exists that which data are used depends on the subjective choice of the experimenter [2,10,18,26,27]. Owing to these drawbacks, the performance of any machine learning approach that leverages such data depends on a meticulous extraction of data and the extraction of key latent features. Because training data are extracted from the internet based on keywords, it is important to select influenza-related keywords that perfectly reflect the latent characteristics of the data [10,18,26]. In previous studies, the keywords were selected by calculating the correlations between each word and influenza-related tokens [10], directly filtering all words that referred to influenza [19,25], or extracting all words that were subjectively related to influenza [27]. Calculating the correlations for all words is the most effective approach to selecting keywords that properly capture the hidden features of the data. However, this approach requires a lot of time because of the sheer number of correlation coefficients that must be calculated. On the other hand, the selection of keywords by screening the words that directly refer to influenza or are subjectively defined to be related to influenza fails to capture the ingrained features, even if the method is relatively fast.

To solve these problems, we proposed a method that combines word embedding [28-32] with cosine similarity to capture only the word vectors that are highly correlated with influenza using the distributed vectors. Filtering is followed by a sorting process that ranks these keywords according to their relationship with the actual influenza outbreak. To assess the effect of the sorting process on embeddings, we applied a long short-term memory (LSTM) [33] predictive model that predicts the impending influenza outbreak.

Word embedding is a natural language processing–based feature extraction technique that consists of establishing a distributed representation of words. Importantly, the features that are generated from word embedding can capture the context between tokens. However, in the context of influenza, using the features obtained through word embedding alone results in a large vector space that includes unnecessary tokens and deteriorates the prediction performance. To reduce the number of tokens to be considered in the prediction stage, the cosine similarity function empowers the word embedding by selecting influenza-related features according to their similarity.

After filtering the features of the tokens that are related to influenza keywords, it is also important to determine the optimal amount of training data to be used for the predictive model to improve its performance. To preferentially use keywords that are highly related to influenza outbreaks among keywords selected by word embedding and cosine similarity, these keywords are sorted using the Pearson correlation coefficient (PCC) [34] between the actual influenza outbreak keywords and the extracted features of the training data. The ultimate purpose of the sorting stage is to ensure that during the training, only the features that are highly correlated with the true features are input to the predictive model. The sorting reduces the error and facilitates the optimization process during the LSTM model training. The model is trained with the fine-grained features, and the sorting process improves the performance of the LSTM predictive model considerably. To assess the effect of the embedding process, various embedding approaches are evaluated.

We compared the model’s performance when the keywords used were sorted versus when they were unsorted. For the evaluation of the performance, we recorded the root-mean-square error (RMSE). FastText continuous bag-of-words (CBOW) outperformed other embedding schemes with a PCC of 0.8986 and an RMSE of 0.0090 with sorted keywords.

Methods

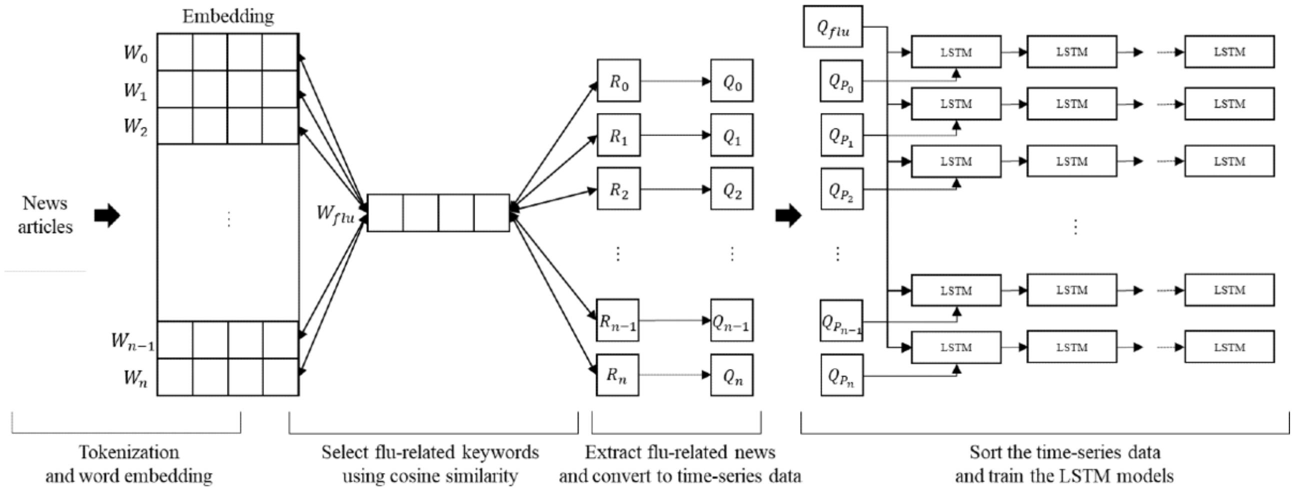
Online News Articles

Online news articles offer a rich opportunity to predict epidemic diseases such as influenza. However, news articles extracted based solely on the presence of the “influenza” token do not capture the hidden insights from the news. The main reason for this is the presence of noisy tokens, such as advertising content that has no association with influenza. To reflect the characteristics of the data, before keyword selection, we used an effective embedding stage to capture the latent relationship between words. Furthermore, to preferentially use the keywords most relevant to the influenza outbreak among the selected keywords, we sorted them according to the PCC based on the actual influenza outbreak and the proportion of news articles containing the keyword. Moreover, the classification model was trained on the extracted keywords.

Main Components of the Overall Methodology

In this section, we cover the overall methodology, which includes 4 main parts: (1) tokenization and word embedding, (2) selection of flu-related keywords via cosine similarity, (3) extraction of flu-related news and its conversion into time-series data, and (4) training and classification. Figure 1 depicts the following 4 components of the model.

Figure 1. System architecture. LSTM: long short-term memory.


Embedding
W_0
W_1
W_2
News articles
W_{flu}
W_{n-1}
W_n
R_0
R_1
R_2
R_{n-1}
R_n
Q_0
Q_1
Q_2
Q_{n-1}
Q_n
Q_{flu}
Q_{P_0}
Q_{P_1}
Q_{P_2}
$Q_{P_{n-1}}$
Q_{P_n}
LSTM
Tokenization and word embedding
Select flu-related keywords using cosine similarity
Extract flu-related news and convert to time-series data
Sort the time-series data and train the LSTM models

Tokenization and Word Embedding

Various tokens that are present in news articles do not have a semantic or syntactic relationship to the classification of the articles. Words such as “at” and “in” or adverbs such as “many” and “very” are filler words that must be removed before the embedding process. Hence, these stop words were stripped. To use only nouns as influenza-related words, tokenization was performed using the Mecab class provided by the morpheme analyzer KoNLPy [35]. The tokenized articles were fed to an embedding module that established a distributed representation of input tokens.

As shown in Figure 1, given the input article made of tokens , the objective of the embedding process is to learn a distributed feature representation of each token in the form of a distributed matrix, , where *n* represents the number of tokens and *d* represents the embedding size. The embedding matrix is structured such that the cosine similarity between the features that represent related tokens is higher. The generated vectors have the same dimension, and thus facilitate the training process.

Here, *W* learns a hidden vector that produces a context vector *W'* that considers other words when representing a given word. Given the input word, W_i, the corresponding word vector in *W* (which is denoted as v_{wi}) generates a corresponding context vector in *W'* (denoted as). The embedding output layer uses a softmax function to estimate the probability, , of generating the output word W_o from W_i via the context vector as follows:

The vector assigned to each word uses the distance between the vectors to capture the relationship between words. Using the cosine similarity between the obtained embedding vectors, it is feasible to express the similarity between words. For instance, the results of the embedding are such that the cosine similarity between the vectors for “influenza” and “sneeze” is closer to 1 and is very close to the similarity between “malaria” and “fever.” Key hyperparameters are set during the training process. The embedding size *d* is the length of a dense vector that represents each word, and the window size is the number of words to be checked simultaneously to learn semantic relationships. The min count represents the minimum number of words to ponder during the training, and any word whose number of appearances is less than this count will be disregarded. In our implementation, we set the embedding size to 300, the window size to 5, and the min count to 100. There are various embedding approaches; in this study, we compared them to evaluate their performance in influenza detection. We compared Word2Vec skip-gram, Word2Vec CBOW, GloVe, FastText CBOW, and FastText skip-gram.

Selection of Flu-Related Keywords

The main objective of our model was to filter the influenza-related tokens to be considered for prediction. For this, we measured the cosine similarity to establish the closeness of each token with the word “influenza.” The cosine function was applied to the embeddings obtained in the previous step. Cosine similarity is a method of measuring the similarity between 2 vectors using the cosine between the 2 vectors. It has a value between –1 and 1. The formula to measure the similarity using vector *W* of a specific word and vector W_{flu} of influenza is as follows:

The above formula means that the inner product of vector *W* of a specific word and vector W_{flu} of influenza is divided by the length of the 2 vectors. We selected *n* influenza-related keywords in the order of high cosine similarity.

Extraction of the Flu-Related News and Its Conversion Into Time-Series Data

Following the selection of influenza-related keywords, we extracted influenza-related news articles containing the keywords selected by word embedding and the word "influenza" simultaneously to ensure that the news articled reflected the characteristics of the data. In other words, news articles extracted through this process were a subset of the news articles that contained only the word "influenza." The following step involved the conversion of news articles that contained only the word "influenza" and news articles that contained both the word "influenza" and the keywords selected by word embedding into time-series data to use as a training set. The *n* related keywords selected by the word embeddings were converted into time-series data by the following process:

In the above equation, *D(t)* represents the number of news articles in the *t*-th week, and $D(W_{flu}\ AND\ R_k)$ is the number of news articles that contain both the word "influenza" and the related keyword R_k. Therefore, Q(k, t) refers to the proportion of news articles containing both "influenza" and R_k news articles from the *t*-th week. The time-series data Q_k are an array of Q(k, t) corresponding to each week .

Sorting the Time-Series Data

Another key objective of the model was to capture a weekly match between influenza trends in news articles and the actual occurrences of influenza. Hence, the sorting of the obtained time-series data was critical to progressive prediction and trend capturing. Time-series data extracted using the keyword selected by word embedding were in the order that was highly related to the word "influenza." Therefore, we sorted the keywords and the time-series data based on the PCCs between the actual influenza outbreak and extracted time-series data to preferentially use the time-series data that were most relevant to the influenza outbreak. For example, since "headache" is a word associated with influenza symptoms that tends to appear alongside "influenza" in many news articles, the generated embeddings for these 2 tokens are likely to be close to encode a high association between "influenza" and "headache." However, because "headache" is a symptom of various diseases, it can be difficult to determine if the "headache" in the text refers to "influenza" outbreak. Therefore, for effective training of influenza prediction, we applied a sorting process that preferentially uses highly relevant tokens to influenza outbreaks. After this step, we trained the (n+1) predictive model by adding the sorted time-series data sequentially to the time-series data extracted using only the word "influenza" (Q_{flu}). This was performed to check the change in performance according to the additional training data and find the optimal number of training data. In other words, the input dimension of the *k*-th predictive model was k-1, and were used as training data.

Training of the Predictive LSTM Model

We built an LSTM model [33] to predict the weekly ILI-related cases. LSTM networks have recently been used for various prediction studies and performed well compared with vanilla recurrent neural networks (RNNs). LSTM networks use a gating mechanism that helps them overcome the vanishing gradient problem faced by RNNs. LSTM networks perform efficiently with time-series data, as they can choose which past information to forget or use while encoding a given time step. Bidirectional LSTM [36], recently studied in the field of natural language processing, showed better performance than unidirectional LSTM on average in time-series prediction such as influenza prediction [37]. However, in order to evaluate the proposed keyword selection process and the performance according to the type of word embeddings, we trained a prediction model using LSTM, which was mainly used in existing influenza studies [6,38,39].

During the training, we calculated the RMSE loss function, which is the square root of the difference between the predicted number of ILI cases and the actual numbers reported by the CDC. The model was optimized using the Adam optimizer [40], the time step was fixed to 5, and the layer size was set to 64.

Results

Embedding Models

To identify the most suitable word embeddings for the selection of influenza-related keywords, we selected 100 keywords that were highly related to influenza using 5 word-embedding models: Word2Vec CBOW, Word2Vec skip-gram, GloVe, FastText CBOW, and FastText skip-gram. The PCC [34] was used to sort the extracted keywords so that only the highly correlated ones were input to the LSTM model for training. The predictive accuracy of each model was evaluated using the PCC and the RMSE [41].

Experimental Setup

We trained each word embedding model to evaluate its performance. As per the recent trend, many studies skip the embedding stage by using pretrained vectors. Although pretrained vectors are obtained from a large data set, they contain many tokens and have exhibited good performance in various recent studies. However, it is difficult to obtain efficient pretrained embeddings for languages other than English. Therefore, we collected approximately 2 million news articles over 2 years from September 11, 2017, to September 15, 2019, and the size of the collected data was approximately 761 MB, containing about 140,000 words as shown in Table 1. Table 2 shows the hyperparameters used when training word embeddings and the LSTM model. Epoch means the number of training repetitions; dimension of word embeddings means the dimension of the vector representing the word, and in the case of LSTM models it means the layer size. The window size of word embeddings means the number of surrounding words to be used for training, and min count means the minimum number of occurrences of words to be used for learning. The LSTM model's time step means how many weeks of data to use for prediction.

Table 1. Summary of news data for word embeddings.

Parameter	Value
Time period	September 11, 2017, to September 15, 2019
Total articles	2,093,120
Total bytes	761,233,009
Total terms	142,651

Table 2. Hyperparameters for word embeddings and long short-term memory model training.

Hyperparameter	Word embeddings	Long short-term memory model
Epoch	10	200
Dimension	300	64
Window size	5	–
Min count	100	–
Time step	–	5 weeks

Experimental Results

Figures 2 to 6 show the accuracy of the predictive model for 100 keywords selected from each word embedding. The black dotted line in each figure depicts the condition when no keyword was selected and only "influenza" was used, and all time-series data related to the word "influenza" were used as input. Moreover, for each embedding schema, the figures show the PCC and the RMSE of the predictive model using the time-series data of only the word "influenza." In the figures, "sorted" means that the keywords selected by the word embeddings were sorted based on the PCC—that is, the keywords were sorted in the order of their correlation with the influenza outbreak. "Unsorted" means that the keywords were not sorted. We expected that both sorted and unsorted approaches would show an accuracy increase to a certain level and then decrease with a further increase in the number of keywords. The sorted version achieved better accuracy than the unsorted method.

Figure 2 shows the accuracy of the LSTM model using PCC and RMSE when adding 1 to 100 time-series training data for the selected keyword using Word2Vec CBOW. As the number of keywords increased, both sorted and unsorted approaches showed an accuracy increase to a certain level and then decreased with a further increase in the number of keywords. The sorted version achieved better accuracy than the unsorted method. In the case of the sorted method, the maximum value achieved by PCC was 0.8951 with 22 keywords used, and the minimum RMSE value was 0.0082 when the same number of keywords was used. In the case of the unsorted method, the maximum PCC was 0.8784 with 59 keywords, and the minimum RMSE value was 0.0095 with 19 keywords. The sorted method showed better accuracy with fewer keywords. When using keywords that were highly related to influenza outbreaks, as the number of keywords increased, the accuracy decreased significantly. However, the decrease in accuracy was a natural result of using less relevant keywords. It was judged that the training data added in the sorted order had a more positive effect on accuracy improvement.

Figure 2. Pearson correlation coefficient (PCC) (A) and root-mean-square error (RMSE) (B) of long short-term memory models using Word2Vec continuous bag-of-words.

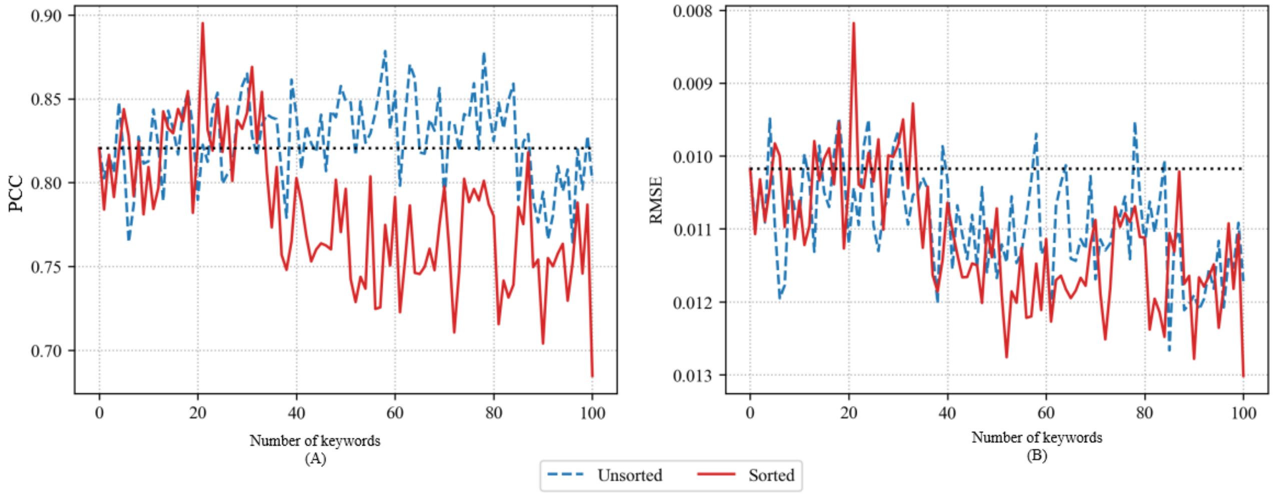


Figure 3 shows the accuracy of the LSTM model using PCC and RMSE when adding 1 to 100 time-series training data for the selected keyword using Word2Vec skip-gram. Both the sorted and unsorted methods of Word2Vec skip-gram showed repeated increases and decreases in accuracy as keywords were added. This means that the keywords selected using Word2Vec skip-gram were somewhat less related to the influenza outbreak than were the keywords selected using Word2Vec CBOW. However, in the case of the sorted method, although the repeated increase and decrease was large, it tended to increase to a certain level and then decrease with a further increase in the number of keywords. For the sorted keywords, the maximum PCC was 0.8942 with 8 keywords, and the minimum RMSE was 0.008 with the same number of keywords. In the case of the unsorted method, the maximum PCC was 0.8942 with 8 keywords, and the minimum RMSE was 0.0089 with 9 keywords.

Figure 3. Pearson correlation coefficient (PCC) (A) and root-mean-square error (RMSE) (B) of long short-term memory models using Word2Vec skip-gram.

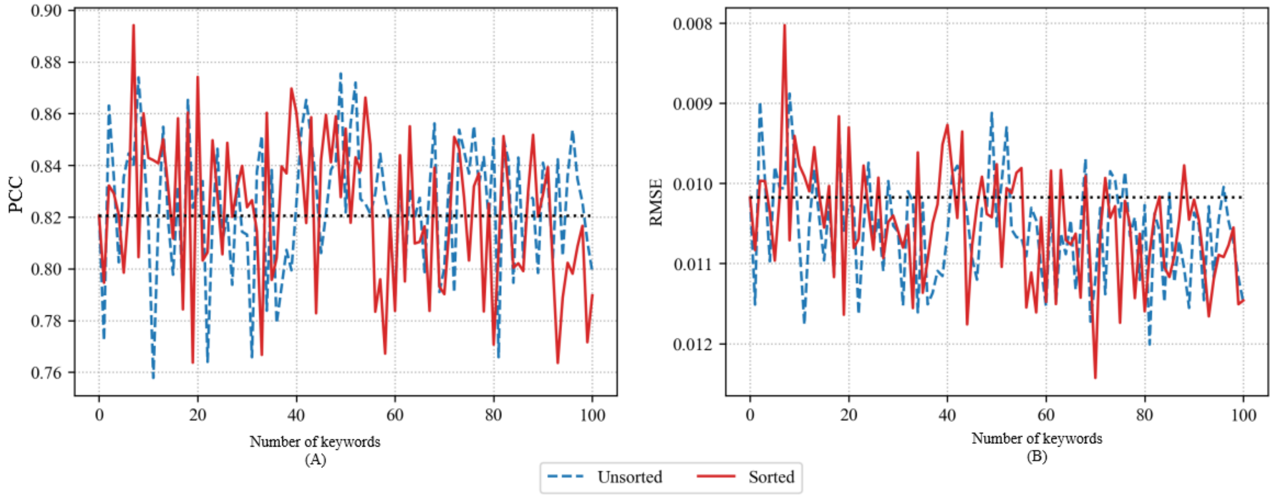
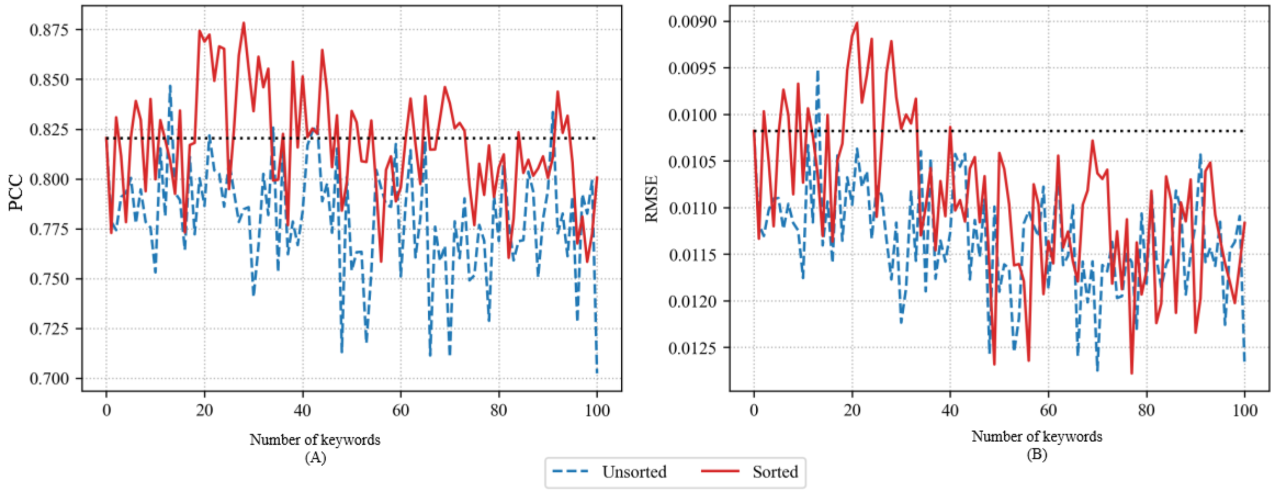


Figure 4 shows the accuracy of the LSTM model using PCC and RMSE when adding 1 to 100 keywords using GloVe. The accuracy of the predictive model using GloVe was similar to that of the predictive model using Word2Vec CBOW. Both the unsorted and sorted methods temporarily exhibited a boost in accuracy as per the increase in the number of keywords. However, the accuracy gently decreased as the number of keywords increased further. Generally, the sorted method achieved higher accuracy. However, as shown in the figure, when the number of added keywords was very large, the accuracy of the unsorted and sorted methods was similar. In the case of the sorted method, the maximum PCC was 0.8783 with 29 keywords, and the minimum RMSE was 0.009 with 22 keywords. In the case of the unsorted method, the maximum PCC was 0.8467 with 14 keywords, and the minimum RMSE was 0.0095 with the same number of keywords.

Figure 4. Pearson correlation coefficient (PCC) (A) and root-mean-square error (RMSE) (B) of long short-term memory models using GloVe.



The accuracy of the LSTM model using PCC and RMSE when adding 1 to 100 time-series training data for the selected keywords using FastText CBOW is depicted in Figure 5. Similar to the accuracy of the predictive model using the previous word embeddings, the sorted method outperformed the unsorted method. The sorted method achieved a maximum PCC of 0.8986 with 34 keywords and a minimum RMSE of 0.009 with the same number of keywords. The unsorted method achieved a maximum PCC of 0.8467 with 42 keywords and a minimum RMSE of 0.0095 with 11 keywords.

Figure 5. Pearson correlation coefficient (PCC) (A) and root-mean-square error (RMSE) (B) of long short-term memory models using FastText continuous bag-of-words.

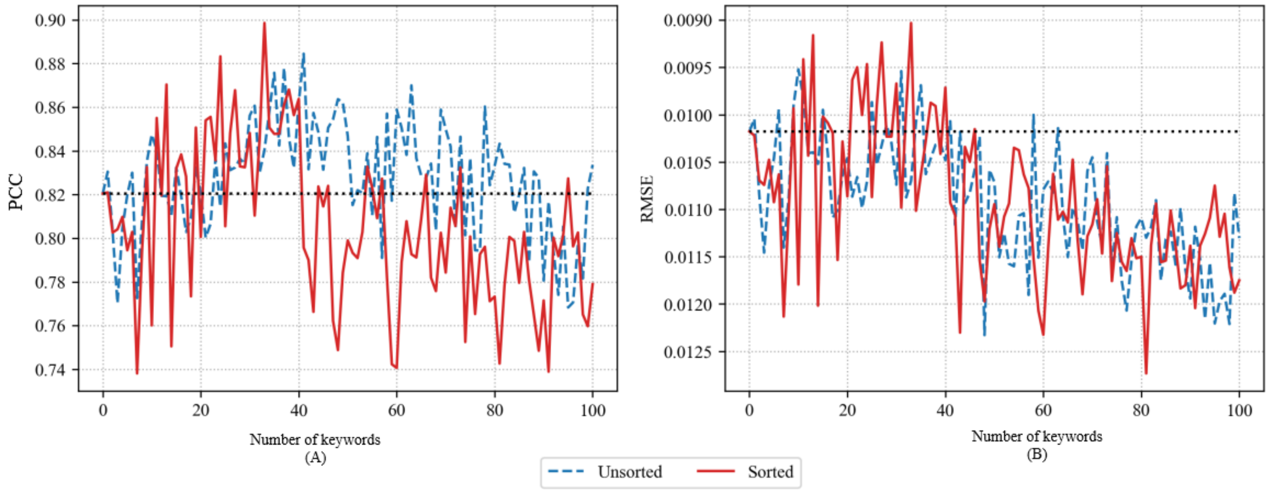
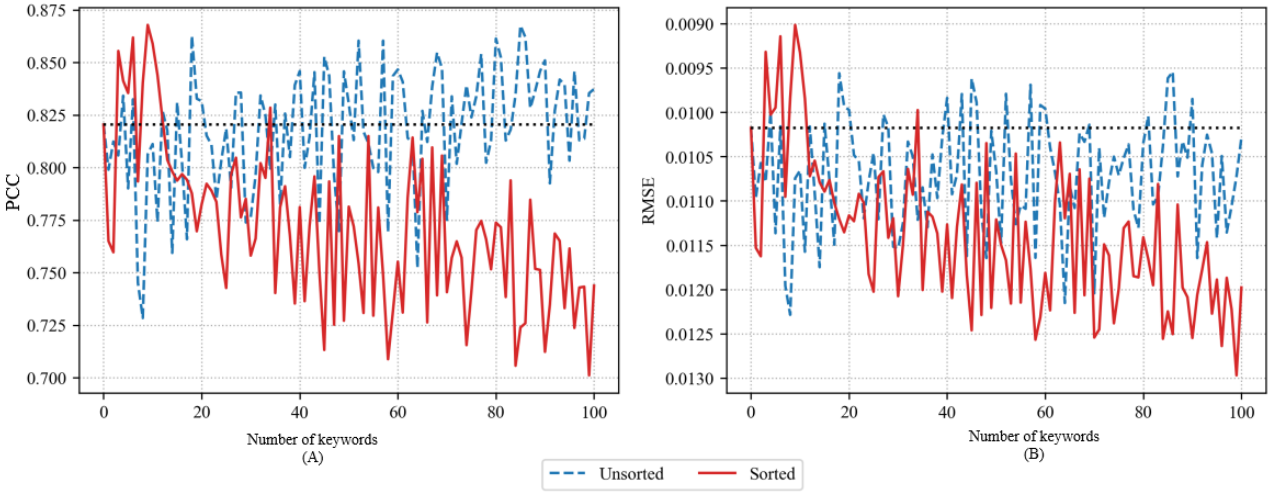


Figure 6 depicts the accuracy of the LSTM model using PCC and RMSE when adding 1 to 100 time-series training data for the selected keywords using FastText skip-gram. The general accuracy of unsorted and sorted methods was lower than that of other word embeddings covered thus far. This means that the time-series data for keywords selected using the FastText skip-gram were negatively correlated with actual influenza outbreaks. In the case of the sorted method, the maximum PCC was 0.8679 with 10 keywords, and the minimum RMSE was 0.009 with the same number of keywords. However, the model that used more keywords than the model with maximum accuracy showed a sharp decline in accuracy. The accuracy was lower than that of the model that used only “influenza” as a keyword. In the case of the unsorted method, the maximum PCC was 0.8676 with 86 keywords, and the minimum RMSE was 0.0095 with 87 keywords. However, similar to the sorted method, the accuracy increased sharply and decreased significantly.

Figure 6. Pearson correlation coefficient (PCC) (A) and root-mean-square error (RMSE) (B) of long short-term memory models using FastText skip-gram.



Analysis

In this study, we aimed to obtain the optimal word embedding when the PCC-based sorting was applied after keyword selection. We compared the best accuracy of the LSTM models trained using each type of word embedding against the number of selected keywords using PCC and RMSE. We considered 2 cases: whether PCC-based sorting was applied or not. Table 3 shows the highest accuracy of the LSTM predictive model using different word embedding techniques and the number of keywords used at each time. We found that the sorted method used fewer keywords but performed better on average. This means that using the data highly related to influenza outbreaks through the sorted method effectively selected training data and improved the average accuracy of the predictive model. Moreover, we found that among the word embedding techniques, FastText CBOW had the highest performance in terms of PCC and Word2Vec skip-gram had the highest performance in terms of RMSE. The process of training by using the context words is the same except that FastText produces a word vector using subword information while Word2Vec considers vectors for complete words. Therefore, there is a slight difference in the performance of Word2Vec and FastText, but it can be confirmed that they are very similar. GloVe, which utilizes the statistical data of the entire document, showed lower performance than the other embedding techniques.

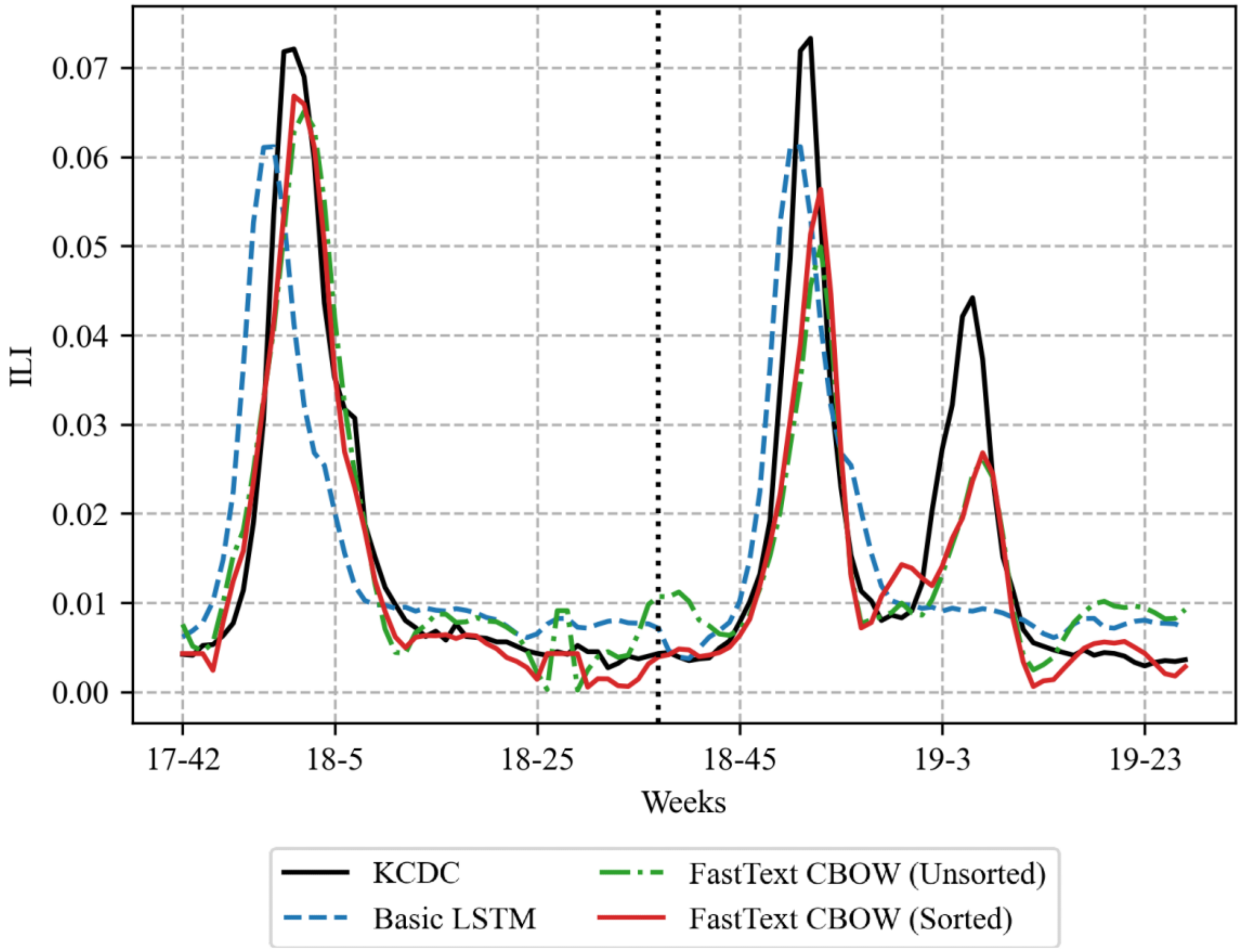
Table 3. Pearson correlation coefficient (PCC) and root-mean-square error (RMSE) for influenza prediction models using different word embedding techniques.

Prediction model	PCC (number of keywords)		RMSE (number of keywords)	
	Unsorted	Sorted	Unsorted	Sorted
Word2Vec CBOW[a]	0.8784 (59)	0.8951 (22)	0.0095 (19)	0.0082 (22)
Word2Vec skip-gram	0.8755 (50)	0.8942 (8)	0.0089 (9)	0.0080 (8)
GloVe	0.8467 (14)	0.8783 (29)	0.0095 (14)	0.0090 (22)
FastText CBOW	0.8845 (42)	0.8986 (34)	0.0095 (11)	0.0090 (34)
FastText skip-gram	0.8676 (86)	0.8679 (10)	0.0095 (87)	0.0090 (10)
Mean	0.8705 (50)	0.8868 (21)	0.0094 (28)	0.0086 (19)

[a] CBOW: continuous bag-of-words.

Figure 7 shows the prediction results of the model using only the time-series data of “influenza” (basic LSTM) and the unsorted and sorted methods using FastText CBOW, respectively, which showed the highest PCCs (Table 3). In Figure 7, the left side of the black dotted line drawn vertically at weeks 18-37 is the prediction result using the training data set, and the right side is the prediction result using the test data set. The predictive model using Korea Centers for Disease Control and Prevention ILI data and time-series data of only “influenza” hardly predicted the influenza peak at weeks 19-5 in the test data set. However, the predictive model trained on time-series data of additional keywords selected by FastText CBOW substantially improved the prediction accuracy compared with the model that used only the word “influenza.” In addition, the method that sorted the keywords selected by FastText CBOW based on PCC and added time-series data outperformed the unsorted method. Both unsorted and sorted methods using FastText CBOW predicted the influenza peaks at weeks 18-1 included in the training data set. However, neither method accurately predicted the influenza peaks at weeks 18-52 and 19-5 in the test data set. This is because the proportion of news articles containing the word “influenza” at the second (18-52) and the third (19-5) peak decreased compared with the first (18-1) peak, which affected the performance of all predictive models.

Figure 7. Comparison of actual influenza outbreaks and influenza prediction results from prediction models. CBOW: continuous bag-of-words; ILI: influenza-like illness; KCDC: Korea Centers for Disease Control and Prevention; LSTM: long short-term memory.



Discussion

Related Work

The accurate and timely prediction of influenza outbreaks has recently gained significant research attention. Many studies rely on legacy statistical approaches. High-performing methods use machine learning with internet-sourced and social network–sourced cloud data.

Eysenbach [2] found a close correlation between epidemiological data on flu and the number of clicks on Google's keyword-triggered links, which is based on the fact that many people use the internet to find health information. The PCC for the number of clicks in the current week and influenza cases in the following week was 0.91, which was a better predictor for influenza than ILIs reported by sentinel physicians. Eysenbach [2] also defined "information epidemiology" or "infodemiology" as a set of research methods such as tracking health information trends on the internet and distributing people's health information. Infodemiology data have the advantage that they can be collected and analyzed in real time.

Ginsberg et al [10] proposed a linear regression model using the search query from the Google search engine and the ILI data provided by the CDC in the United States to predict influenza. The rationale behind the study was that the search frequency of any influenza-related search query was correlated with the occurrence of influenza. The study established a list of candidate query groups to be used in the regression model by calculating the correlation between time-series forms of all search queries and the ILI value from the CDC. Hence, the top 100 of these correlated search queries were selected for training the model. The performance of the model improved depending on the number of highly correlated queries. The accuracy improved with 100 queries but did not improve with 45 queries.

Achrekar et al [19] proposed the framework of social network–enabled flu trends, which monitored flu trends. The study developed a model based on autoregression with an exogenous input that used tweets to predict influenza warnings and ILI occurrences. Tweets with the keywords "flu," "H1N1," and "swine flu" were defined as influenza-related tweets. Support vector machines (SVMs) [42] were used to exclude meaningless tweets. The study concluded that Twitter data were highly correlated with ILI rates.

Li and Cardie [25] developed a model that predicted influenza using Twitter and a probabilistic graphical Bayesian approach based on a Markov network. The approach divided influenza progression into 4 phases: nonepidemic, rising epidemic, stationary epidemic, and declining epidemic. Tweets containing the keywords "flu," "H5N1," "H5N9," "swine flu," and "bird flu" were defined as influenza-related tweets, and SVMs were used to remove the unnecessary tweets.

Zhang et al [27] implemented FluOutlook, an online system for predicting influenza outbreaks in 7 countries using statistical regression analysis and Global Epidemic and Mobility models [43,44]. The model was based on Influweb [45]—a voluntary participation information collection system—and Twitter. FluOutlook collected tweets containing 40-50 defined keywords and assigned a priority flag based on the correlation between the time-series data corresponding to each keyword and actual flu occurrences. The limited number of keywords helped mitigate the effect of noise included in the collected raw tweets.

These recent influenza prediction studies have used search queries and microblogging, such as Twitter, for real-time prediction. However, search queries provided by search engines (such as Google) cannot be used for real-time prediction because it is difficult and imprecise to infer the exact search trends. Moreover, as already asserted, Twitter and other social platforms are prone to noise. On the other hand, web-based news data exhibit less vulnerability to noise and have recently been adopted in several prediction studies [46-48]. The strength of these news data is due to real-time online accessibility and rigorous professional editing.

A crucial aspect to consider during the extraction of training data from the internet is the selection of keywords. Various studies calculated correlations for all words or used keywords that directly indicated influenza or were subjectively selected. Calculating the correlation coefficient for every token has been argued to be the best approach. However, it requires a lot of computing resources and training time. The direct or subjective selection of influenza-related keywords cannot be generalized to various data sets because it is challenging to extract the inherent features of the data set. Therefore, a method for selecting related keywords by reflecting the latent characteristics of the data during the selection of keywords improves the model considerably.

Various studies have also focused on word embedding as a feature extraction method that can capture the semantic and contextual aspects from texts by establishing a distributed representation of each token.

Mikolov et al [29,30] proposed Word2Vec—a model that uses a shallow neural network to assign a distributed vector to each word by calculating the co-occurrence probability. Using the distributional hypothesis [49], the probabilities are calculated such that words with close meaning or words that are likely to appear together in a certain context window are close in the vector space. The model consists of 2 distinct learning paradigms: skip-gram and CBOW. To build the distributed vector, skip-gram learns the probability of occurrence of context words from the target word, while CBOW learns the probability of occurrence of the target words from context words. Word2Vec uses local information (context window) between words in the context by disregarding the global information. Hence, Pennington et al [31] proposed GloVe, which assigns a vector to each word by using the proportion of the target word appearing along with other words throughout the document.

Another key limitation of Word2Vec is that it ignores the internal morphology of words and fails to capture proper vectors for rare words. To address this limitation, Joulin et al [32] proposed FastText, which considers the subwords of each word. Rather than feeding the individual words to the neural network, FastText breaks them into n-grams and uses skip-grams to learn the distributed representation of each of these subwords. The final representation of a distinct word is the sum of these n-grams.

Limitations and Future Work

When predicting influenza from news articles, we used word embedding to find words related to influenza and sorted them based on their association with actual influenza outbreaks, effectively extracting training data and improving the accuracy of predictions. However, our research has the following limitations, and future studies are needed. First, we need to check whether our approach works well for novel data sets other than news articles. Recently, influenza prediction has been studied using various data [38,50-53]. Therefore, it is necessary to study whether our approach can improve performance when applied to different data sets used in the recent state-of-the-art studies. In this study, we focused on improving the representation of the training data rather than on the learning scheme. Hence, we used the standard, unmodified LSTM model, which is widely used in existing influenza prediction studies [6,38,39]. However, research is being conducted to change the standard LSTM model in state-of-the-art influenza prediction [54,55] or to apply a prediction model that shows better performance in other fields [56,57]. Therefore, it is necessary to study whether our approach can lead to improvement in performance when applied to predictive models other than the standard LSTM model. Third, we used word embedding to extract keyword candidates for training data extraction, but we need to see if our sorting process can improve performance even when other keyword extraction methods are used.

Conclusions

In this paper, we proposed an effective training data extraction method to improve influenza prediction from news articles. The input data selected by the extraction method encoded the relationship between the words with influenza-related keywords. Subsequently, these data were filtered as per their relationship with the actual influenza outbreak. This process was ensured by sorting the selected keywords based on PCCs between the actual influenza outbreak and the proportion of news articles containing the keywords. The predictive model that was trained on the extracted data using only the word "influenza" did not reflect the characteristics of the collected data; hence, it showed unsatisfactory performance. However, because the predictive models trained on the data extracted through the proposed method reflected the characteristics of the data, it was confirmed that the performance was greatly improved. We also compared the performance of the predictive models with 5 popular word embedding techniques. The experimental results proved that with the proposed method, FastText CBOW outperformed other embedding techniques with unsorted and sorted keywords.

Acknowledgments

This work was supported by the National Research Foundation of Korea Grant funded by the Korea Government under Grant NRF2019R1F1A1058058.

Conflicts of Interest

None declared.

References

1. Influenza (Seasonal). World Health Organization. 2018 Nov 06. URL: https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal) [accessed 2020-05-10]
2. Eysenbach G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. : American Medical Informatics Association; 2006 Presented at: AMIA annual symposium proceedings; November; Washington, DC p. 244.
3. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res 2009 Mar 27;11(1):e11 [FREE Full text] [doi: 10.2196/jmir.1157] [Medline: 19329408]
4. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS. Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. PLoS Comput Biol 2015 Oct;11(10):e1004513 [FREE Full text] [doi: 10.1371/journal.pcbi.1004513] [Medline: 26513245]
5. Hirose H, Wang L. Prediction of Infectious Disease Spread Using Twitter: A Case of Influenza. : IEEE Computer Society; 2012 Presented at: 2012 Fifth International Symposium on Parallel Architectures, Algorithms and Programming; December; NW Washington, DC p. 100-105. [doi: 10.1109/paap.2012.23]
6. Liu L, Han M, Zhou Y, Wang Y. LSTM Recurrent Neural Networks for Influenza Trends Prediction. 2018 Presented at: International Symposium on Bioinformatics Research and Applications; June; Beijing, China p. 259-264. [doi: 10.1007/978-3-319-94968-0_25]
7. Signorini A, Segre AM, Polgreen PM. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PLoS One 2011 May 04;6(5):e19467 [FREE Full text] [doi: 10.1371/journal.pone.0019467] [Medline: 21573238]
8. Paul M, Dredze M, Broniatowski D, Generous N. Worldwide influenza surveillance through twitter. 2015 Presented at: AAAI workshop: WWW and public health intelligence; January; Palo Alto, California.
9. Achrekar H, Gandhe A, Lazarus R, Yu S, Liu B. Predicting flu trends using twitter data. 2011 Presented at: IEEE conference on computer communications workshops (INFOCOM WKSHPS); March; Shanghai, China p. 702-707. [doi: 10.1109/infcomw.2011.5928903]
10. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009 Feb 19;457(7232):1012-1014. [doi: 10.1038/nature07634] [Medline: 19020500]
11. Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. PLoS One 2011;6(8):e23610 [FREE Full text] [doi: 10.1371/journal.pone.0023610] [Medline: 21886802]
12. Yuan Q, Nsoesie EO, Lv B, Peng G, Chunara R, Brownstein JS. Monitoring influenza epidemics in china with search query from baidu. PLoS One 2013;8(5):e64323 [FREE Full text] [doi: 10.1371/journal.pone.0064323] [Medline: 23750192]
13. Lampos V, Miller AC, Crossan S, Stefansen C. Advances in nowcasting influenza-like illness rates using search query logs. Sci Rep 2015 Aug 03;5:12760 [FREE Full text] [doi: 10.1038/srep12760] [Medline: 26234783]
14. Moss R, Zarebski A, Dawson P, McCaw JM. Forecasting influenza outbreak dynamics in Melbourne from Internet search query surveillance data. Influenza Other Respir Viruses 2016 Jul;10(4):314-323 [FREE Full text] [doi: 10.1111/irv.12376] [Medline: 26859411]
15. Zimmer C, Leuba SI, Yaesoubi R, Cohen T. Use of daily Internet search query data improves real-time projections of influenza epidemics. J R Soc Interface 2018 Oct 10;15(147):20180220 [FREE Full text] [doi: 10.1098/rsif.2018.0220] [Medline: 30305417]
16. Xu Q, Gel YR, Ramirez Ramirez LL, Nezafati K, Zhang Q, Tsui K. Forecasting influenza in Hong Kong with Google search queries and statistical model fusion. PLoS One 2017;12(5):e0176690 [FREE Full text] [doi: 10.1371/journal.pone.0176690] [Medline: 28464015]
17. Aramaki E, Maskawa S, Morita M. Twitter catches the flu: detecting influenza epidemics using Twitter. : Association for Computational Linguistics; 2011 Presented at: Proceedings of the 2011 Conference on empirical methods in natural language processing; July; Edinburgh, Scotland, UK p. 1568-1576.

18. Culotta A. Towards detecting influenza epidemics by analyzing Twitter messages. 2010 Presented at: Proceedings of the First Workshop on Social Media Analytics; July; Washington D.C. District of Columbia p. 115-122. [doi: 10.1145/1964858.1964874]
19. Achrekar H, Gandhe A, Lazarus R, Yu S, Liu B. Twitter Improves Seasonal Influenza Prediction. 2012 Presented at: Healthinf; Feburary; Vilamoura, Algarve, Portugal p. 70. [doi: 10.5220/0003780600610070]
20. Paul MJ, Dredze M, Broniatowski D. Twitter improves influenza forecasting. PLoS Curr 2014 Oct 28;6:ecurrents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117 [FREE Full text] [doi: 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117] [Medline: 25642377]
21. Wang F, Wang H, Xu K, Raymond R, Chon J, Fuller S, et al. Regional Level Influenza Study with Geo-Tagged Twitter Data. J Med Syst 2016 Aug;40(8):189. [doi: 10.1007/s10916-016-0545-y] [Medline: 27372953]
22. Allen C, Tsou MH, Aslam A, Nagel A, Gawron J. Applying GIS and Machine Learning Methods to Twitter Data for Multiscale Surveillance of Influenza. PLoS One 2016;11(7):e0157734 [FREE Full text] [doi: 10.1371/journal.pone.0157734] [Medline: 27455108]
23. Grover S, Aujla GS. Prediction model for influenza epidemic based on Twitter data. Int J Adv Res Comput Commun Eng 2014;3(7):7541-7545.
24. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. Science 2014 Mar 14;343(6176):1203-1205. [doi: 10.1126/science.1248506] [Medline: 24626916]
25. Li J, Cardie C. Early stage influenza detection from twitter. arXiv Prepr arXiv 2013:1309.7340.
26. Kim EK, Seok JH, Oh JS, Lee HW, Kim KH. Use of hangeul twitter to track and predict human influenza infection. PLoS One 2013;8(7):e69305 [FREE Full text] [doi: 10.1371/journal.pone.0069305] [Medline: 23894447]
27. Zhang Q, Gioannini C, Paolotti D, Perra N, Perrotta D, Quaggiotto M, et al. Social data mining and seasonal influenza forecasts: the fluoutlook platform. 2015 Presented at: Joint European Conference on Machine Learning and Knowledge Discovery in Databases; September 7-11; Porto, Portugal p. 237-240. [doi: 10.1007/978-3-319-23461-8_21]
28. Hinton G. Distributed representations. Technical Report CMU-CS-84-157 1984:1-31.
29. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: Proceedings of the International Conference on Learning Representations (ICLR 2013). 2020 Mar 31 Presented at: International Conference on Learning Representations (ICLR 2013); May 2-4, 2013; Scottsdale, Arizona p. 71-77. [doi: 10.3126/jiee.v3i1.34327]
30. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. arXiv :1310.4546 Preprint posted online October 16, 2013. [doi: 10.5040/9781474284974.00399]
31. Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. 2014 Presented at: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October; Doha, Qatar p. 1532-1543. [doi: 10.3115/v1/d14-1162]
32. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers.: Association for Computational Linguistics; 2017 Presented at: EACL 2017,; April 3-7, 2017; Valencia, Spain. [doi: 10.18653/v1/e17-2068]
33. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
34. Benesty J, Chen J, Huang Y, Cohen I. Pearson correlation coefficient. In: Noise Reduction in Speech Processing. New York City: Springer; 2009:1-4.
35. Park EL, Cho S. KoNLPy: Korean natural language processing in Python. 2014 Presented at: Annual Conference on Human and Language Technology; October; Chuncheon, Korea p. 133-136.
36. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv :1508.01991 Preprint posted online August 9, 2015.
37. Siami-Namini S, Tavakoli N, Namin A. A Comparative Analysis of Forecasting Financial Time Series Using ARIMA, LSTM, and BiLSTM. arXiv :1911.09512 Preprint posted online November 21, 2019.
38. Venna SR, Tavanaei A, Gottumukkala RN, Raghavan VV, Maida AS, Nichols S. A Novel Data-Driven Model for Real-Time Influenza Forecasting. IEEE Access 2019;7:7691-7701. [doi: 10.1109/access.2018.2888585]
39. Volkova S, Ayton E, Porterfield K, Corley CD. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. PLoS One 2017;12(12):e0188941 [FREE Full text] [doi: 10.1371/journal.pone.0188941] [Medline: 29244814]
40. Kingma D, Ba J. Adam: A method for stochastic optimization. 2015 Presented at: 3rd International Conference for Learning Representations; 2015; San Diego, California.
41. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geosci Model Dev 2014 Jun 30;7(3):1247-1250. [doi: 10.5194/gmd-7-1247-2014]
42. Lilleberg J, Zhu Y, Zhang Y. Support vector machines and Word2vec for text classification with semantic features. 2015 Presented at: 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing; July; Beijing, China p. 136-140. [doi: 10.1109/icci-cc.2015.7259377]

43. Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, Vespignani A. Multiscale mobility networks and the spatial spreading of infectious diseases. Proc Natl Acad Sci U S A 2009 Dec 22;106(51):21484-21489 [FREE Full text] [doi: 10.1073/pnas.0906910106] [Medline: 20018697]
44. Balcan D, Gonçalves B, Hu H, Ramasco JJ, Colizza V, Vespignani A. Modeling the spatial spread of infectious diseases: the GLobal Epidemic and Mobility computational model. J Comput Sci 2010 Aug 01;1(3):132-145 [FREE Full text] [doi: 10.1016/j.jocs.2010.07.002] [Medline: 21415939]
45. Influweb. 2020. URL: https://www.influweb.it/ [accessed 2020-05-07]
46. Shynkevich Y, McGinnity T, Coleman SA, Belatreche A. Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning. Decision Support Systems 2016 May;85:74-83. [doi: 10.1016/j.dss.2016.03.001]
47. McGough SF, Brownstein JS, Hawkins JB, Santillana M. Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data. PLoS Negl Trop Dis 2017 Jan;11(1):e0005295 [FREE Full text] [doi: 10.1371/journal.pntd.0005295] [Medline: 28085877]
48. Liu D, Clemente L, Poirier C, Ding X, Chinazzi M, Davis J, et al. Correction: Real-Time Forecasting of the COVID-19 Outbreak in Chinese Provinces: Machine Learning Approach Using Novel Digital Data and Estimates From Mechanistic Models. J Med Internet Res 2020 Sep 22;22(9):e23996 [FREE Full text] [doi: 10.2196/23996] [Medline: 32960774]
49. Sahlgren M. The distributional hypothesis. Ital J Disabil Stud 2008;20:33-53.
50. Zimmer C, Leuba SI, Yaesoubi R, Cohen T. Use of daily Internet search query data improves real-time projections of influenza epidemics. J R Soc Interface 2018 Oct 10;15(147):20180220 [FREE Full text] [doi: 10.1098/rsif.2018.0220] [Medline: 30305417]
51. Schneider PP, van Gool CJ, Spreeuwenberg P, Hooiveld M, Donker GA, Barnett DJ, et al. Using web search queries to monitor influenza-like illness: an exploratory retrospective analysis, Netherlands, 2017/18 influenza season. Euro Surveill 2020 May;25(21):1900221 [FREE Full text] [doi: 10.2807/1560-7917.ES.2020.25.21.1900221] [Medline: 32489174]
52. Xue H, Bai Y, Hu H, Liang H. Regional level influenza study based on Twitter and machine learning method. PLoS One 2019;14(4):e0215600 [FREE Full text] [doi: 10.1371/journal.pone.0215600] [Medline: 31013324]
53. Molaei S, Khansari M, Veisi H, Salehi M. Predicting the spread of influenza epidemics by analyzing twitter messages. Health Technol 2019 Mar 21;9(4):517-532. [doi: 10.1007/s12553-019-00309-4]
54. Zhu X, Fu B, Yang Y, Ma Y, Hao J, Chen S, et al. Attention-based recurrent neural network for influenza epidemic prediction. BMC Bioinformatics 2019 Nov 25;20(Suppl 18):575 [FREE Full text] [doi: 10.1186/s12859-019-3131-8] [Medline: 31760945]
55. Zhang J, Nawata K. Multi-step prediction for influenza outbreak by an adjusted long short-term memory. Epidemiol Infect 2018 May;146(7):809-816 [FREE Full text] [doi: 10.1017/S0950268818000705] [Medline: 29606177]
56. Kondo K, Ishikawa A, Kimura M. Sequence to sequence with attention for influenza prevalence prediction using Google Trends. 2019 Presented at: Proceedings of the 2019 3rd International Conference on Computational Biology and Bioinformatics; October; New York, United States p. 1-7. [doi: 10.1145/3365966.3365967]
57. Wu N, Green B, Ben X, O?Banion S. Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case. arXiv :2001.08317 Preprint posted online January 23, 2020.

Abbreviations

CBOW: continuous bag-of-words
CDC: Centers for Disease Control and Prevention
ILI: influenza-like illness
LSTM: long short-term memory
PCC: Pearson correlation coefficient
RMSE: root-mean-square error
RNN: recurrent neural network
SVM: support vector machine

Edited by G Eysenbach; submitted 07.08.20; peer-reviewed by K Edemacu, G Harerimana; comments to author 23.08.20; revised version received 13.10.20; accepted 01.04.21; published 25.05.21.

Please cite as:
Jang B, Kim I, Kim JW
Effective Training Data Extraction Method to Improve Influenza Outbreak Prediction from Online News Articles: Deep Learning Model Study
JMIR Med Inform 2021;9(5):e23305
URL: https://medinform.jmir.org/2021/5/e23305
doi:10.2196/23305
PMID:34032577

Original Paper

Leveraging Genetic Reports and Electronic Health Records for the Prediction of Primary Cancers: Algorithm Development and Validation Study

Nansu Zong[1], PhD; Victoria Ngo[2], PhD; Daniel J Stone[1], BSc; Andrew Wen[1], MSc; Yiqing Zhao[1], PhD; Yue Yu[1], PhD; Sijia Liu[1], PhD; Ming Huang[1], PhD; Chen Wang[1], PhD; Guoqian Jiang[1], MD, PhD

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, United States

[2]University of California Davis Health, Sacramento, CA, United States

Corresponding Author:
Guoqian Jiang, MD, PhD
Department of Health Sciences Research
Mayo Clinic
200 First Street
Rochester, MN
United States
Phone: 1 480 301 8000
Email: Jiang.Guoqian@mayo.edu

Abstract

Background: Precision oncology has the potential to leverage clinical and genomic data in advancing disease prevention, diagnosis, and treatment. A key research area focuses on the early detection of primary cancers and potential prediction of cancers of unknown primary in order to facilitate optimal treatment decisions.

Objective: This study presents a methodology to harmonize phenotypic and genetic data features to classify primary cancer types and predict cancers of unknown primaries.

Methods: We extracted genetic data elements from oncology genetic reports of 1011 patients with cancer and their corresponding phenotypical data from Mayo Clinic's electronic health records. We modeled both genetic and electronic health record data with HL7 Fast Healthcare Interoperability Resources. The semantic web Resource Description Framework was employed to generate the network-based data representation (ie, patient-phenotypic-genetic network). Based on the Resource Description Framework data graph, Node2vec graph-embedding algorithm was applied to generate features. Multiple machine learning and deep learning backbone models were compared for cancer prediction performance.

Results: With 6 machine learning tasks designed in the experiment, we demonstrated the proposed method achieved favorable results in classifying primary cancer types (area under the receiver operating characteristic curve [AUROC] 96.56% for all 9 cancer predictions on average based on the cross-validation) and predicting unknown primaries (AUROC 80.77% for all 8 cancer predictions on average for real-patient validation). To demonstrate the interpretability, 17 phenotypic and genetic features that contributed the most to the prediction of each cancer were identified and validated based on a literature review.

Conclusions: Accurate prediction of cancer types can be achieved with existing electronic health record data with satisfactory precision. The integration of genetic reports improves prediction, illustrating the translational values of incorporating genetic tests early at the diagnosis stage for patients with cancer.

(JMIR Med Inform 2021;9(5):e23586) doi:10.2196/23586

KEYWORDS

genetic reports; electronic health records; predicting primary cancers; Fast Healthcare Interoperability Resources; FHIR; Resource Description Framework; RDF

Introduction

Cancer is the second leading cause of death worldwide [1]. The health burden of cancer in the United States is substantial [2,3], with approximately 1.8 million new diagnoses and an estimated 600,000 deaths in 2020 alone [4]. Despite the advances in characterizing oncogenic mutations in the past few decades, overcoming the consequences of cellular self-renewal and neoplastic transformation remains a challenge in cancer therapy research [5]. Therefore, continued discoveries in causes, treatment, and management are needed to further the knowledge and understanding of this collection of related diseases [6].

Modern gene technology has provided an opportunity to identify certain gene mutations associated with increased cancer risk. Approximately 5% to 10% of all cancer diagnoses are linked to cancer predisposition syndromes [7-9]. Major syndromes of cancer disposition affecting adults include breast, ovarian, prostate, gastric, and pancreatic cancer [7]. Precision medicine initiatives call for the leveraging of clinical and genomic data to not only screen for cancers but also to help monitor cancer progression and guide therapy options [10]. Clinicians can facilitate early screening critical for risk assessment and surveillance [8]. If cancer is detected at an early stage, survival rates tend to be significantly higher than those for cancers diagnosed at an advanced stage [11-13]. Nash et al [11] cite figures as drastic as 90% survival for early ovarian cancer detection compared to only 5% survival with advanced stage detection, as an example. The utilization of genetic tests in diagnosing primary cancer also becomes critical when the symptoms and the physical exams suggest unspecified cancer known as cancer of unknown primary [14]. Cancer of unknown primary accounts for 3% to 5% of all tumors [15]. The prediction of the primary cancer of cancer of unknown primary can significantly increase our current knowledge of metastasis and benefit the treatment of patients with cancer of unknown primary.

The implementation and adoption of health information technology have given frontline clinicians access to a large repository of longitudinal clinical data collected during health care encounters [16,17]. Medical insight and clinical decision making rely heavily upon access to these data from electronic health records. Artificial intelligence techniques, such as machine learning methods, are promising for finding patterns and discovering associations in health care data to help predict diseases [18]. Improved predictions can be made by integrating diverse types of digital data in patients' charts, which include diagnosis codes, clinical notes, laboratory test results, and treatment data [19].

As demand grows for genetic testing from patients and as genomic data continue to be incorporated into electronic health records, there is a need to study how genetic reports, along with electronic health record data, can be leveraged to predict cancers. Conventional computational methods for predictive models are based on features extracted from diverse data sources, known as bag of features [20]. The features in these models are treated independently, and the potential connections and patterns among the features cannot be fully explored to serve the prediction. A network-based data model can be used to represent the association between data models with edges, and the potential patterns are embedded in the topological structure of the network. Predictions from network-based data representations have achieved promising results in diverse biomedical areas, such as drug-target prediction [21] and patient clustering [22]. Representing correlations among phenotypic and genetic data elements through network-based data modeling shows great potential in cancer prediction.

The objective of this study was to harmonize phenotypic and genetic features for accurate and explainable cancer prediction, specifically: (1) developing a network-based framework with standard health care data exchange frameworks, the HL7 Fast Healthcare Interoperability Resources (FHIR) [23] and the Resource Description Framework (RDF) for graph-based data representations, (2) employing a state-of-the-art graph embedding algorithm, Node2vec [24], to obtain features for machine learning and deep learning models, and (3) implementing the proposed method with a collection of genetic reports of patients with cancer and the corresponding phenotypic data from Mayo Clinic's electronic health record systems and comprehensive experiments.

Methods

Preliminary

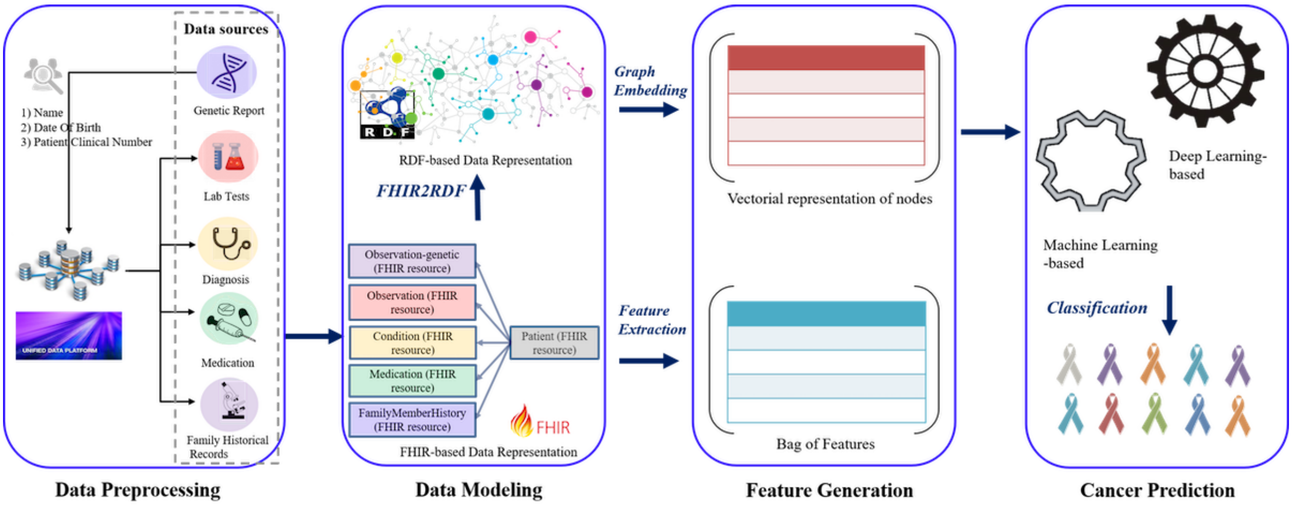
FHIR is a standardized data framework designed for data exchange between different medical centers to enable information to be captured as it is generated, significantly simplifying population and real-time updates of predefined data models [23,25]. The FHIR specification defines a set of granular clinical concepts and resources to provide standard data infrastructure to support implementations [23]. FHIR-based data models are built upon combinations of these resources and a set of attributes with value types. The common attributes (eg, *identifier*) and unique attributes (eg, *bodySite*) in a resource are used to facilitate data modeling. Common data types (eg, *String* and *CodeableConcepts*) are used to constrain the attribute based on an adaptation of clinically related ontologies, such as SNOMED CT [26], LOINC [27], and International Statistical Classification of Diseases ninth (ICD-9) and tenth revisions (ICD-10) [28].

RDF is a general metadata or data model that defines concepts and web-resources based on a variety of syntax notations and data serialization formats [29]. Inherited from the classical conceptual modeling approaches, RDF utilizes the expressions to form triples, subject-predicate-object, to model data elements (eg, web resources). Specifically, in this study, the subject denotes the clinical data elements (eg, patients), and the predicate denotes a relationship between 2 data elements.

Framework

We proposed a network-based framework (Figure 1) that represented cancer data using the FHIR standard and RDF to facilitate the cancer prediction process. Five types of data sources extracted from the electronic health record—genetic information, lab tests, diagnosis, medication, and family historical records—were represented with FHIR resources and converted to the RDF-based representation. A graph-embedding algorithm, Node2vec, was used to provide a vectorial representation of nodes in the resulting network along with bag of features to form the features for the classification models.

Figure 1. A network-based framework for cancer prediction based on Fast Healthcare Interoperability Resources and Resource Description Framework.



Data Preprocessing

Genetic data were extracted from 1011 aggregated anonymized genetic test results (Foundation Medicine Inc), including microsatellite instability and tumor mutational burden. Medical record data elements related to laboratory results, diagnoses, medications, and family histories were extracted from approximately 515,000 billing encounters (666,000 electronic health record encounters) retrieved from a Mayo Clinic clinical data warehouse of [30]. We integrated genetic and electronic health record data by mapping patient information based on 3 data elements: patient clinic number, names (first and last name), and date of birth. Lab tests, diagnosis, medication, and family historical records were searched based on the mapped patients. We used natural language processing to normalize the names and values. For *diagnosis* and *medication*, all diseases and medications were represented with standardized names encoded by ICD-9 [31] and RxNorm [32] codes. For lab tests, we represented all the tests with standard names encoded by LOINC [27]. For *family historical records*, each record was processed by a pipeline (NLP2FHIR [33]), where the medical concepts were identified and normalized using cTAKES [34], MedXN [35], and MedTime [36]. We encoded the diseases from family historical records using ICD-9 codes. To build the data set utilized for the cancer prediction, all the records within the billing circle related to the target cancers were removed. The top 10 elements in each data source can be found in Table 1.

Table 1. Distribution of the top 10 elements in each data source.

Code and verbatim description		Record, n (%)
Genes		
TP53	tumor protein p53	553 (54.70)
KRAS	KRAS proto-oncogene, GTPase	292 (28.88)
MLL2	lysine methyltransferase 2D[a]	173 (17.11)
LRP1B	LDL receptor related protein 1B	171 (16.91)
MLL3	lysine methyltransferase 2C[a]	150 (14.84)
APC	APC regulator of WNT signaling pathway	141 (13.95)
ARID1B	AT-rich interaction domain 1B	137 (13.55)
FAT1	FAT atypical cadherin 1	134 (13.25)
PRKDC	protein kinase, DNA-activated, catalytic subunit	128 (12.66)
ARID1A	AT-rich interaction domain 1A	126 (12.46)
Diagnosis[b]		
Z02.9	Work Status Exam (RTW)	204 (25.66)
I10	Hypertension (HTN) Chronic	142 (17.86)
401.9	HYPERTENSION NOS	138 (17.36)
272.4	HYPERLIPIDEMIA NEC/NOS	116 (14.59)
R91.8	Mass Lung	113 (14.21)
V68.9	ADMINISTRTVE ENCOUNT NOS	106 (13.33)
Z00.00	Maintenance Health (HM)	101 (12.70)
E78.5	Dyslipidemia NOS	93 (11.70)
V72.83	PREOP EXAMINATION NEC	79 (9.94)
V70.0	ROUTINE MEDICAL EXAM	79 (9.94)
Lab tests[c]		
777-3	Platelets [#/volume] in Blood by Automated count	991 (99.40)
2160-0	Creatinine [Mass/volume] in Serum or Plasma	988 (99.10)
965763	Hematocrit [Volume Fraction] of Blood by Automated count	985 (98.80)
718-7	Hemoglobin [Mass/volume] in Blood	985 (98.80)
788-0	Erythrocyte distribution width [Ratio] by Automated count	985 (98.80)
789-8	Erythrocytes [#/volume] in Blood by Automated count	985 (98.80)
1749545	Leukocytes [#/volume] in Blood by Automated count	985 (98.80)
787-2	MCV [Entitic volume] by Automated count	985 (98.80)
337180	Potassium [Moles/volume] in Serum or Plasma	975 (97.80)
383903	Sodium [Moles/volume] in Serum or Plasma	973 (97.60)
Family historical records[b]		
V47.2	Other cardiorespiratory problems	205 (29.54)
429.9	Heart disease, unspecified	205 (29.54)
429.89	Other ill-defined heart diseases	205 (29.54)
162.9	Malignant neoplasm of bronchus and lung, unspecified	133 (19.16)
162.8	Malignant neoplasm of other parts of bronchus or lung	130 (18.73)
272.4	Other and unspecified hyperlipidemia	124 (17.87)
434.91	Cerebral artery occlusion, unspecified with cerebral infarction	104 (14.99)

Code and verbatim description		Record, n (%)
799.9	Other unknown and unspecified cause of morbidity and mortality	84 (12.10)
311	Depressive disorder, not elsewhere classified	72 (10.37)
447.9	Unspecified disorders of arteries and arterioles	63 (9.08)
Medication[d]		
5956	Iohexol	399 (72.41)
1359867	Sodium Chloride 9 MG/ML Prefilled Syringe	374 (67.88)
1807638	20 ML Sodium Chloride 9 MG/ML Injection	304 (55.17)
1807639	1000 ML Sodium Chloride 9 MG/ML Injection	298 (54.08)
1740467	2 ML Ondansetron 2 MG/ML Injection	251 (45.55)
4337	Fentanyl	224 (40.65)
314659	heparin sodium, porcine	207 (37.57)
847630	Calcium Chloride 0.0014 MEQ/ML / Potassium Chloride 0.004 MEQ/ML / Sodium Chloride 0.103 MEQ/ML / Sodium Lactate 0.028 MEQ/ML Injectable Solution	202 (36.66)
198440	Acetaminophen 500 MG Oral Tablet	188 (34.12)
1808234	10 ML Propofol 10 MG/ML Injection	163 (29.58)
Cancers[b]		
162.9	Malignant neoplasm of bronchus and lung, unspecified	231 (22.85)
153.9	Malignant neoplasm of colon, unspecified site	124 (12.27)
155	Malignant neoplasm of liver, primary	118 (12.67)
157.9	Malignant neoplasm of pancreas, part unspecified	116 (11.47)
183	Malignant neoplasm of ovary	85 (8.41)
185	Malignant neoplasm of prostate	80 (7.91)
171.9	Malignant neoplasm of connective and other soft tissue, site unspecified	68 (6.73)
193	Malignant neoplasm of thyroid gland	55 (5.44)
174.9	Malignant neoplasm of breast (female), unspecified	53 (5.24)
—[e]	—	—

[a] Current standard gene symbols: *MLL2* is now *KMT2D*; *MLL3* is now *KMT2C.*

[b] International Statistical Classification of Diseases (ninth revision) code and description.

[c] LOINC code and description.

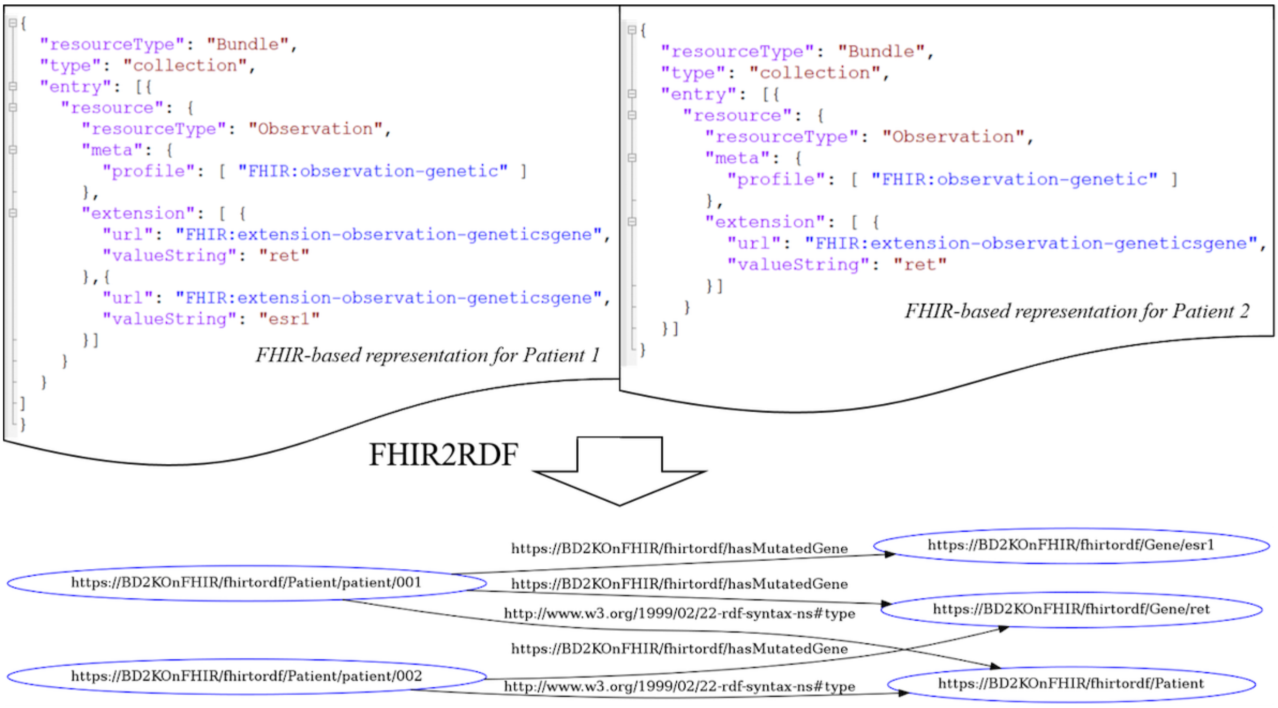
[d] RxNorm code and description.

[e] A tenth item is not included.

Data Preprocessing and Data Modeling Based on FHIR and RDF

We adapted FHIR-based data models from our previous work [37] employing FHIR resources to represent data elements of genetic reports and structured electronic health record data for phenome-wide association studies. Specifically, we represented *genetic* entries with the existing profile *Observation-genetics*, extended from the resource *Observation*. The *lab test*, *diagnosis*, and *medication* entries were represented with the resources *Observation*, *Condition*, and *Medication*, respectively, and were identified by encounters (eg, billing and electronic health record encounters) and service date. The *family historical records* entities were represented with the resource *FamilyMemberHistory* as diseases and were encoded with the attributed condition. All the resources were associated with the resource *Patient*. We further converted the JavaScript object notation–formatted FHIR data to RDF format based on the conversion rules, where (1) all the string-type values were considered as the entities in the RDF graph, and (2) all the values of the resources were considered as the object of the data-type property—named after the resource for the subject resource *Patient*. We illustrated an example of data representation based on FHIR and RDF in Figure 2.

Figure 2. An example of data representation based on Fast Healthcare Interoperability Resources (FHIR) and Resource Description Framework (RDF): 2 JavaScript object notation–formatted FHIR representations for patients 1 and 2 are merged and converted into 1 RDF graph.



```
{
  "resourceType": "Bundle",
  "type": "collection",
  "entry": [{
    "resource": {
      "resourceType": "Observation",
      "meta": {
        "profile": [ "FHIR:observation-genetic" ]
      },
      "extension": [ {
        "url": "FHIR:extension-observation-geneticsgene",
        "valueString": "ret"
      },{
        "url": "FHIR:extension-observation-geneticsgene",
        "valueString": "esr1"
      }]
    }
  }
]
}
```

```
{
  "resourceType": "Bundle",
  "type": "collection",
  "entry": [{
    "resource": {
      "resourceType": "Observation",
      "meta": {
        "profile": [ "FHIR:observation-genetic" ]
      },
      "extension": [ {
        "url": "FHIR:extension-observation-geneticsgene",
        "valueString": "ret"
      }]
    }
  }]
}
```

FHIR-based representation for Patient 1
FHIR-based representation for Patient 2
FHIR2RDF
https://BD2KOnFHIR/fhirtordf/Patient/patient/001
https://BD2KOnFHIR/fhirtordf/Patient/patient/002
https://BD2KOnFHIR/fhirtordf/hasMutatedGene
https://BD2KOnFHIR/fhirtordf/hasMutatedGene
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
https://BD2KOnFHIR/fhirtordf/hasMutatedGene
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
https://BD2KOnFHIR/fhirtordf/Gene/esr1
https://BD2KOnFHIR/fhirtordf/Gene/ret
https://BD2KOnFHIR/fhirtordf/Patient

Feature Generation and Cancer Prediction

Bag of Features

Bag of features is analogous to the bag-of-words representation and characterizes a sample with an orderless collection of features [38]. In this study, we used bag of features based on the attribute values from the FHIR model. Specifically, categorical values of mutated genes, lab test results, disease diagnoses, medications for treatment, and historical family disease diagnoses were collected as the features from *Observation-genetics*, *Observation*, *Condition*, *Medication*, and *FamilyMemberHistory*, respectively. Additionally, patient demographic features, such as age and gender, were also used.

Topological Features

In order to train a model with the features generated from the input RDF data, we adapted a methodology [21] that considered RDF graph as a network, $G(V,E)$ with a set of vertices V and a set of edges E, where V has 7 types of vertices (ie, genetics, lab tests, diagnosis, medication, family historical records, demographics, and patients) and E represents associations between the 6 types of vertices (ie, genetics, lab tests, diagnosis, medication, family historical records, demographics) and patients. We used the graph embedding method to learn the features of the patients, where a patient could be represented by a vector embedded within the topological structure of the patient in the network G. Node2vec [30] is a state-of-art graph embedding method that vectorizes the vertices of a network based on the topology of the network by maximizing the probability of observing the neighborhood $N(u)$ of each node u in G:

where

and $f(\cdot)$ was the feature representation of a node. In addition, we also generated a $|V|\times|V|$ adjacency matrix from G, where each cell of the matrix was set to 1 if there was a connection between nodes, otherwise the cell was set to 0.

We modeled cancer prediction as a multiple-label classification problem, where a given patient was represented with k-dimensional features, and a model categorized the patient into precisely 1 of 9 cancer types: colon cancer (ICD-9: 153.9), pancreas cancer (ICD-9: 157.9), ovary cancer (ICD-9: 183), prostate cancer (ICD-9: 185), connective and other soft tissue cancer (ICD-9: 171.9), thyroid gland cancer (ICD-9: 193), breast cancer (ICD-9: 174.9), liver cancer (ICD-9: 155), and bronchus and lung cancer (ICD-9: 162.9).

Experiment Design

Overview

There were 2 main drivers of this study: (1) from a methodological perspective—how could generated features be coordinated with classification methods in a favorable manner to achieve satisfactory prediction?—and (2) from a data perspective—which data sources, especially genetic data, are preferable in prediction? Our experiment was thus conducted as a sequence of 6 distinct tasks.

Task 1: Comparison of Combinations of Features and Popular Classification Methods

A comparison of 3 feature generation methods—bag of features, Node2vec, and bag of features+Node2vec (ie, a linear combination of bag of features and Node2vec)—was conducted. Seven classification methods—random forest [39], naive Bayes [40], logistic regression [41], support vector machine [42], deep neural network [43], convolutional neural network [44], and graph convolutional networks [45]—were used.

Task 2: Comparison of Combinations of Data Sources

There were 5 types of data sources used in this study. We took all possible combinations of the data sources into consideration and studied how the features generated from these sources affected the results.

Task 3: Comparison of Predictions for Each Cancer

To understand how the prediction varied in different cancers predictions, we conducted 9 prediction tasks for all the cancers to study.

Task 4: Analysis of Feature Contribution for Each Cancer Prediction

To interpret the model and understand which features were important to each cancer, we studied the features that contributed most to the prediction of cancer.

Task 5: Time Effect of Cancer Prediction

To understand how the prediction could be made precisely prior to a certain amount of time of the diagnosis, we studied the prediction based on data collected at different duration, ranging from 0 to 24 months, in advance.

Task 6: Prediction of Cancer of Unknown Primary Patients

We identified the 43 primary cancers from 81 patients with cancer of unknown primary based on the diagnosis records to understand how the proposed method performed for real cancer predictions. Please note, no patients with pancreas cancer of unknown primary were identified, and therefore, pancreatic cancer was not considered in this task.

Feature Selection and Classification

Two methods were used to generate features: bag of features and Node2vec. For bag of features, all genes, diseases, drugs in genetics, diagnosis, medication, and family historical records were considered as features. For the lab tests, the values were converted into categorical values (Null, Normal, or Abnormal) based on the normal range defined in the unified data platform. To avoid overfitting, the features were reduced to d={10,20,30,40,50,60,70,80,90,100} based on information gain [46]. For Node2vec, the parameter ranges for the grid search were specified as the number of walks γ={10,40}, return P={0.5,1.0,2.0}, in-out q={0.5,1.0,2.0}, dimension d={10,20,30,40,50,60,70,80,90,100}, window size w={5,10}, and walk length t={40,80}.

Four popular machine learning models and 3 deep learning models were used for classification. For machine learning methods, the following settings were used: L2 regularization for logistic regression, type C-SVC and linear kernel for support vector machine, 500 trees for random forest, and default settings for naive Bayes. For deep learning methods, the following structure were used: 5 dense layers with dimensions {256, 256,128, 64, 10} (4 rectified linear unit [ReLU] activation functions with 0.5 dropout rate and 1 softmax activation function) for deep neural network, 3 convolution layers with filters {256, 256, 256} (3 ReLU activation functions and maxpooling layers with 0.5 dropout rate) followed with 4 dense layers with dimensions {256,128, 64, 10} (3 ReLU activation functions with 0.5 dropout rate and 1 softmax activation function) for convolutional neural network, and 2 graph convolutional layers with channels {64, 10}(1 ReLU activation function with 0.5 dropout rate and 1 softmax activation function) for graph convolutional networks.

Node2vec was obtained from the Node2vec library [47]. The logistic regression classifier was obtained from the LIBLINEAR library [48]; naive Bayes, random forest, and information gain algorithms were obtained from Weka library [49], support vector machine was obtained from LIBSVM [50]. Deep neural network and graph convolutional networks were constructed based on Keras library [51]. Graph convolutional networks algorithms were obtained from Spektral library [52].

Validation and Evaluation Metrics

We used conventional 10-fold cross-validation for the evaluation, where 10 independent iterations of training and testing were conducted, and a random partition of the original samples into 10 equal-size subsamples was performed. To assess the quality of classification, we used area under the receiver operating characteristic curve (AUROC) [53]. In addition, the area under the precision-recall curve (AUPRC) [53] was used as a supplementary metric characterizing the results for imbalanced classes [54,55]. AUROC and AUPRC scores were calculated using the Java Receiver Operating Characteristic library [56] and Weka evaluation package [57].

Results

Combinations of Features and Popular Classification Methods

Table 2 shows the best performance result was achieved by using bag of features+Node2vec and random forest (AUROC 96.19%) (AUPRC: Table S1, Multimedia Appendix 1). Generally, using bag of features+Node2vec outperformed using bag of features (+1.27 %) and Node2vec (+1.41%). Although we observed that machine learning–based methods outperformed deep learning–based methods, in general, the best deep learning–based approach (AUROC 95.12%) was second to the best machine learning–based approach by only 1 percentage-point difference (outperforming the remaining machine learning–based approaches). As our implementation of deep learning models is based on simple architectures, the deep learning models with more complex architectures have the potential to facilitate feature generation and may directly contribute to improvements in cancer prediction.

Table 2. Prediction performance (area under the receiver operatic characteristic curve) for combinations of features and classification methods.

Classifiers	Feature generation algorithm		
	Bag of features	Node2vec	Bag of features+Node2vec
	AUROC[a] (%)	AUROC (%)	AUROC (%)
Random forest	94.82	91.89	96.19
Naive Bayes	92.30	92.91	94.76
Logistic regression	86.68	85.25	89.39
Support vector machine	84.62	83.92	86.72
Convolutional neural network	64.14	63.36	57.68
Deep neural network	92.56	92.87	95.12
Graph convolutional networks	79.67	83.62	83.83

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

Combinations of Data Sources

Table 3 shows better results were achieved by the model DML+G (diagnosis, medication, lab test, and genetic information; AUROC 96.56%). Steady improvement is obtained when more features are used (AUPRC: Table S2, Multimedia Appendix 1). For example, increasing average AUROCs (75.49%, 82.65%, 87.98%, and 91.74%) are achieved by adding 1 to 5 features successively without using genetic information. Table 3 also presents the importance of the features, where lab test is the most important feature (91.00%), followed by diagnosis (73.12%), medication (72.83%), and family historical records (65.01%). We also demonstrated the value of genetic information for cancer prediction—an average improvement of 10.52% was reached. Interestingly, such improvement is weakened when more feature types are used (+15.76% for using 1 feature type, +10.45% for 2 feature types, +6.92% for 3 feature types, and +4.45% for 4 feature types). Table 3 also indicates the potential of using diverse types of features alternatively when genetic information is not available.

Table 3. Prediction performance for combinations of data sourcing with bag of features+Node2vec and random forest algorithms.

Feature types	AUROC[a] (%)	
	Base feature set	With genetic information
1 feature type		
G[b]	73.12	90.89
D[c]	65.01	88.37
H[d]	91.00	95.80
L[e]	72.83	89.94
M[f]	73.21	90.92
2 feature types		
DH	91.55	96.09
DL	77.09	90.88
DM	91.30	95.92
HL	71.53	89.02
MH	91.22	95.75
ML	91.98	96.01
3 feature types		
DHL	76.76	91.28
DMH	91.76	96.56
DML	91.43	95.76
MHL	91.74	96.19
4 feature types		
DMHL	73.12	90.89

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

[b] G: genetic information.

[c] D: diagnosis.

[d] H: family historical records.

[e] L: lab test.

[f] M: medication.

Predictions for Each Cancer

Table 4 shows that the proposed method achieved high AUROC values across all 9 cancer types (AUPRC: Table S3, Multimedia Appendix 1), especially for thyroid gland (AUROC 99.80%), prostate (99.76%), breast (98.53%), ovary (98.29%), connective and other soft tissue (96.05%), and liver (95.41%). Genetic information improved the predictions in general (*P*<.001) based on a Wilcoxon signed-rank test [58], specifically for thyroid gland cancer (*P*=.03), ovary cancer (*P*=.03), connective and other soft tissue cancer (*P*=.03), liver cancer (*P*=.03), and colon cancer (*P*=.03).

Table 4. Prediction performance for 9 cancer types.

Cancer (ICD-9[a] code)	AUROC[b] (%)	
	DML[c]	DML+G[d]
Malignant neoplasm of thyroid gland (193)	99.55	99.80
Malignant neoplasm of prostate (185)	98.43	99.76
Malignant neoplasm of breast (female), unspecified (174.9)	96.80	98.53
Malignant neoplasm of ovary (183)	95.73	98.29
Malignant neoplasm of connective and other soft tissue, site unspecified (171.9)	82.39	96.05
Malignant neoplasm of liver, primary (155)	91.39	95.41
Malignant neoplasm of pancreas, part unspecified (157.9)	91.07	95.41
Malignant neoplasm of bronchus and lung, unspecified (162.9)	90.61	93.24
Malignant neoplasm of colon, unspecified site (153.9)	79.88	92.56

[a] ICD-9: International Statistical Classification of Diseases, ninth revision.

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

[c] DML: diagnosis, medication, and lab test.

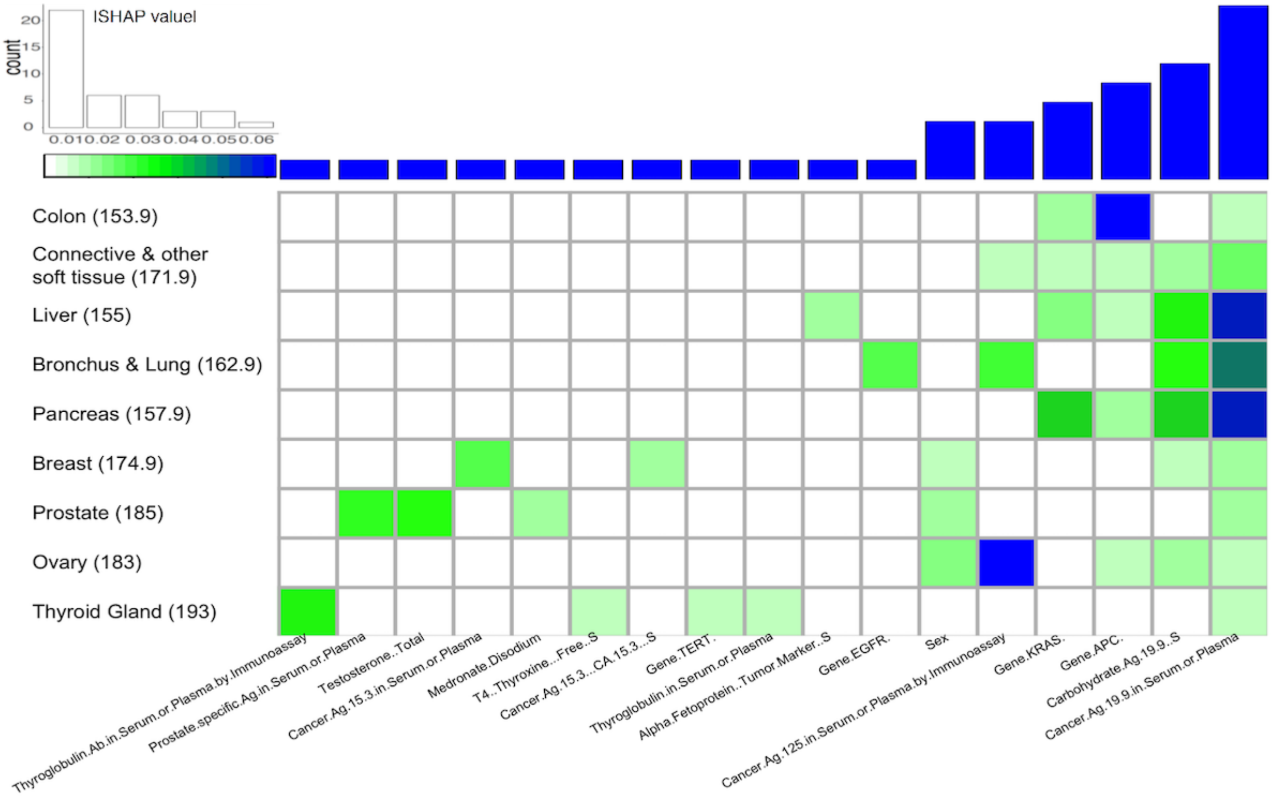
[d] DML+G: diagnosis, medication, and lab test, and genetic information.

Feature Contributions for Each Cancer Prediction

Our analysis examines the feature contribution based on SHAP values [59] for the cancer prediction and selects the top 5 features interpretable for each cancer (Figure 3). Frequent common features are lab tests (11/17); cancer antigen 19-9 in serum or plasma (2.03%), carbohydrate antigen 19-9, S (1.76%), and cancer antigen 125 in serum or plasma by immunoassay (2.59%) are the most common features across all the cancer types. These lab tests are considered to be predictive biomarkers for prognosis and chemotherapeutic effect for carcinomas [60-63]. Two genes—KRAS proto-oncogene, GTPase homolog (*KRAS*) (1.46%) and adenoma polyposis coli regulator of WNT signaling pathway (*APC*) (1.60%) contribute the most cancer predictions. *KRAS* is the most commonly mutated oncogene in human cancers. The sustained expression and signaling of *KRAS* results in the progress of many cancers thus make it the high-priority target in clinical therapeutic implications [64]. *APC* participates in a cytoplasmic complex and its mutation triggers negatively regulating canonical WNT signaling. *APC* counteracts proliferation, facilitates apoptosis, and suppresses tumor progression, thus APC-deficient tumors drive colorectal and gastric cancers [65,66].

Lab tests testosterone (2.49%) and prostate-specific antigen in serum or plasma (2.29%) were found to be the major contributors to prostate cancer prediction. Evidence supports the androgen hypothesis, where prostate cancer development and progression are related to androgens. These findings drive the studies to explore the correlation between testosterone and prostate cancer development and progression [67,68]. For thyroid gland cancer prediction, thyroglobulin antibody in serum or plasma by immunoassay (2.69%), thyroglobulin in serum or plasma (0.58%), T4 (thyroxine) (0.62%), and gene telomerase reverse transcriptase (*TERT*) (SHAP value 0.59%) were found to be the major contributors. Associations between autoimmune thyroiditis and thyroid cancer have been documented [69] in studies where thyroid autoimmunity was assessed by measuring thyroglobulin antibody and thyroid peroxidase antibody [70,71]. Thyroglobulin in serum also plays a key role in the surveillance of differentiated patients with thyroid cancer [72]. *TERT* promoter mutations have been found to be strongly associated with different pathological types of thyroid cancers and are considered as the biomarker to the preoperative diagnosis and prognosis of thyroid cancers [73]. Cancer antigen 15-3 in serum or plasma (1.57%) and cancer antigen 15-3 (CA 15-3) S (0.98%) lab tests are the major contributors to breast cancer prediction. Cancer antigen 15-3 is a protein made by a variety of cells, particularly breast cancer cells, and the cancer antigen 15-3 test is A biomarker test used to monitor breast cancer [74]. In addition, the cancer markers alpha-fetoprotein, tumor marker, S (0.78%) and epidermal growth factor receptor (EGFR) (1.56) were found to be the main contributors for the prediction of cancers of the liver [75] and bronchus and lung [76]. In our study, sex appears to be the major contributor to prediction of cancers of the breast (0.76%), prostate (0.92%), and ovary (1.16%).

Figure 3. Top 5 features contributing to cancer prediction.



Time Effect of Cancer Prediction

Table 5 shows predictions based on different resources with different combinations of time-dependent (diagnosis, medication, and lab test) features (AUPRC: Table S4, Multimedia Appendix 1). Among the 7 models, diagnosis and lab test were the best (average AUROC 90.31 %). In general, the performance of prediction decreases as more time increases prior to the formal diagnosis. For example, the average performance was reduced from 92.37% to 77.18% from 0 months to 24 months in advance, with an average decrease of 3.04%. Table 5 also demonstrates the performance of the model (ie, diagnosis, medication, lab test, and genetic information) based on genetic information (AUROC 91.38% at 24 months in advance, an improvement of +11.38% over diagnosis, medication, and lab test). The difference between the two increase as time increases (eg, 1.06 for 0 months to 11.38% for 24 months), which suggests the importance of genetic testing at early stages.

Table 5. Prediction performance (AUROC) 0 months to 24 months in advance.

Months	Feature type							
	DML+G[a]	Diagnosis, medication, and lab test	Diagnosis and lab test	Diagnosis and medication	Medication and lab test	Diagnosis	Medication	Lab test
	AUROC[b] (%)	AUROC (%)	AUROC (%)	AUROC (%)	AUROC (%)	AUROC (%)	AUROC (%)	AUROC (%)
0	99.43	98.36	98.41	97.89	88.67	97.90	70.39	87.93
1	98.08	95.62	95.51	94.31	86.83	94.53	71.01	86.67
3	96.52	93.16	93.22	90.74	84.85	91.20	69.36	84.18
6	95.21	89.69	89.91	85.53	83.09	85.26	68.12	83.38
12	93.17	84.39	84.60	78.20	80.56	78.21	66.76	79.99
24	91.38	80.01	80.20	71.81	77.73	71.71	66.22	78.35

[a] DML+G: diagnosis, medication, and lab test, and genetic information.

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

Prediction of Patients With Cancer of Unknown Primary

In spite of the challenge in identifying patients with cancer of unknown primary in the clinical setting, hybrid features—the diagnosis, medication, lab test, and genetic information model—outperformed the diagnosis, medication, and lab test model (AUPRC: Table S5, Multimedia Appendix 1), and bag of features+Node2vec outperform the bag of features and Node2vec in most cases. Table 6 shows promising prediction results for 4 cancers, especially for breast (AUROC 92.31%), connective and other soft tissue (AUROC 92.31%). Cancers of the liver and lung have the largest number of patients (24/43) and also achieved satisfactory predictions (AUROCs 88.21% and 85.51%). We also note that the proposed method performed suboptimally in predicting cancer of the colon (AUROC 52.56%). Prediction of the prostate, thyroid gland, and colon cancers had better results for the bag of features+Node2vec model with diagnosis, medication, and lab test features and for the bag of features or Node2vec model with diagnosis, medication, lab test, and genetic information features (Table S6, Multimedia Appendix 1), suggesting a more flexible strategy of model adaptation for the prediction of cancer of unknown primary in practice.

Table 6. AUROC (%) of prediction for 9 cancer types.

Cancer (ICD-9[a] code)	AUROC[b] (%)		Patients, n
	DML[c]	DML+G[d]	
Malignant neoplasm of breast (female), unspecified (174.9)	83.97	92.31	4
Malignant neoplasm of connective and other soft tissue, site unspecified (171.9)	53.21	92.31	4
Malignant neoplasm of liver, primary (155)	84.10	88.21	13
Malignant neoplasm of bronchus and lung, unspecified (162.9)	74.43	85.51	11
Malignant neoplasm of ovary (183)	65.85	80.49	2
Malignant neoplasm of prostate (185)	91.67	79.17	3
Malignant neoplasm of thyroid gland (193)	90.24	75.61	2
Malignant neoplasm of colon, unspecified site (153.9)	64.74	52.56	4

[a] ICD-9: International Statistical Classification of Diseases, ninth revision.

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

[c] DML: diagnosis, medication, and lab test.

[d] DML+G: diagnosis, medication, and lab test, and genetic information.

Discussion

It is recognized that both genetic and nongenetic factors may lead to the development of cancers, and they are, therefore, considered to be risk factors in the plethora of cancer prediction models based on statistical analysis; this leads to performance (eg, AUROC) ranging from 60% to 90% [77]. For example, the variables of high DNA load of high-risk human papillomavirus, age, marital status, smoking status, and age at sexual debut are the critical factors to achieve the AUROC 90% in the prediction of cervical intraepithelial neoplasia grade 2 or worse [78]. DNA methylation-based markers-based method achieves AUROC 93% in the detection of preinvasive neoplasia and cervical cancer [79]. Computational methods (eg, machine learning and deep learning) have been adapted to provide solutions for cancer prediction challenges in a controlled environment (eg, UCI machine repository [80]). For example, linear support vector machines achieved AUROC 96.7% [81] and k-nearest neighbors classifier achieved an accuracy of 99.28% [82] for breast cancer prediction.

Public genetic expression databases (eg, The Cancer Genome Atlas) are frequently used to train diverse deep learning models. A convolutional neural network–based model achieved accuracies of 93.9% to 95.0% in the prediction of 34 cancer types [83]. For lung, stomach, and breast cancer, AUROCs 99.5%, 97.1%, and 95.0%, respectively, were achieved by a stacked sparse auto-encoder–based classification model [84]. Prostate cancer prediction achieved an AUROC of 95.5% with a genetic algorithm–optimized artificial neural network [85]. Accuracies of 95.3% for breast cancer, 57.9% for leukemia, and 84.9% for colon cancer were achieved by sample expansion based 1D convolutional neural network [86].

Electronic health records are utilized in cancer prediction. DeepPatient has proposed a novel unsupervised feature learning method based on autoencoders for disease prediction [87]. The overall AUROC was 77.3%, where AUROCs of 88.7% for cancer of rectum and anus, 88.6% for cancer of the liver and intrahepatic bile duct, 85.9% for cancer of the prostate were predicted with a time interval of 12 months. Multiple studies have utilized electronic health record data to predict specific cancers, where AUROCs of 88.1% for lung cancer [88], 64.8% for breast cancer [89], 85% for pancreatic cancer [90] were achieved, and 85.7% precision and 60.0% recall were achieved for colorectal cancer [91]. Our method achieved AUROC 96.56% in general and outperformed the state-of-the-art methods for most cancer types. Specifically, prostate cancer (99.8%), breast cancer (AUROC 98.5%), liver cancer (95.4%), and pancreas cancer (95.4%) predictions results were better for our method.

In this study, we designed and developed a network-based framework leveraging the FHIR resources and RDF for cancer prediction. Our contributions can be summarized as exploration of utilizing FHIR and RDF technology to provide a network-based representation for the prediction of patient health status, demonstrating the value of integrating the phenotypic and genetic features data sources to improve the accuracy and interpretability in cancer prediction models. To enable the standard representation of data, a FHIR-based representation was used as the core to support the network population and feature generation. It is one of the most popular clinical data standards and is widely used among modern electronic health record vendors and data providers, enabling the plug and play functionality of the proposed method to be used across the different institutions, and it provides the specification and tools to seamlessly convert to RDF format and support the efficient data communication based on the popular data exchanging formats, such as XML or JavaScript object notation.

This study demonstrated a solution for the prediction of unknown cancer in clinical practice. Despite the value of this work, there are several limitations that should be addressed.

First, the genetic alterations in the genetic reports provided in Foundation Medicine are all somatic mutations in tumors and are collected from somatic tissues. Thus, we could not differentiate the germline and somatic mutations in our model. The bias introduced to the system caused by a failure in capturing this difference weakens the findings of our study.

Second, as most genetic tests are based on specimens collected from the biopsy or surgery, the best-performing (diagnosis, medication, lab test, and genetic information) model introduced in Task 5 might not be adaptable as some medical organizations have limited access to genetic information available for study. We, therefore, consider that it is more practical to learn a large amount of phenotypical information for cancer prediction with the full utilization of existing generic information. On the other hand, as the costs of genetic testing are reduced, we believe that the genetic information will be increasingly used in prediction models for different tasks, which makes the proposed method a good reference as a pilot study.

Third, within 81 patients who have been documented as having cancer of unknown primary (from genetic reports), we could identify specific cancer types for 43 patients based on the review of patients' diagnostic report for task 6. We understand that the limited data set used might affect result analysis, which is a limitation of this experiment. We also noticed that the proposed method performs differently in task 6, especially with some notable failures. Such failures indicate the patterns of the value distribution for the features learned in the training data are not the same as the patterns in the cancer of unknown primary. The cancer of unknown primary source is not considered a single type of cancer and is known to spread at the early stage without causing phenotypical symptoms at the origin site [92]. As such, the proposed model is affected in Task 5 accordingly.

Fourth, our experiment demonstrates the performance of the proposed method based on data collected over a varying timeline. Data were used in isolation to train classification models, ignoring the continuous changing of the measurable values of phenotypes (eg, lab tests) during cancer progression. The introduction of deep learning models, such as recurring neural networks [93] and long short-term memory [94], which are capable of processing time-series data may potentially improve predictions.

Fifth, cancers related to the same genetic alteration (eg, both colorectal and gastric cancers are related to the *APC* gene) inspire us to explore the potential of considering dependent phenotypes of the genetic alteration. With the utilization of phenotype and genotype dependence based on the ontology structure, a more sophisticated method can be designed to empower the prediction. In the future, we plan to reach out to other institutions to apply our method both with and without genetic information on diverse electronic health record systems. We consider it is necessary to adopt other medical data standards, such as Observational Health Data Sciences and Informatics Common Data Model [95], to cover the diversity. We are aware that there are some challenging issues in genetic data modeling with relational databases, such as how to anonymize and aggregate genomic data. We believe that the research community will develop solutions for handling these challenging issues. We will incorporate such developments into our framework as part of future work to better support these requirements. The data process and cancer prediction tools of this study are publicly available [96].

Acknowledgments

This work was supported by funding from Genentech Research Fund in Individualized Medicine, National Institute of Health, National Institute of General Medical Sciences (K99GM135488), Big Data to Knowledge (U01HG009450), and FHIRCat (R56EB028101).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables.
[XLSX File (Microsoft Excel File), 16 KB - medinform_v9i5e23586_app1.xlsx]

References

1. Cancer. World Health Organization. 2018. URL: https://www.who.int/news-room/fact-sheets/detail/cancer [accessed 2021-05-11]
2. Islami F, Miller KD, Jemal A. Cancer burden in the United States—a review. Ann Cancer Epidemiol 2018;1:1-1. [doi: 10.21037/ace.2018.08.02]
3. Leading causes of death. Centers for Disease Control and Prevention. 2017. URL: https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm [accessed 2021-05-11]
4. Cancer statistics. National Cancer Institute. 2020. URL: https://www.cancer.gov/about-cancer/understanding/statistics [accessed 2021-05-11]
5. Clarke M, Hass A. Cancer stem cells. In: Meyers RA, editor. Reviews in Cell Biology and Molecular Medicine. Weinheim: Wiley ‐ VCH Verlag GmbH & Co KGaA; 2004:221-241.
6. Blackadar CB. Historical review of the causes of cancer. World J Clin Oncol 2016 Feb 10;7(1):54-86 [FREE Full text] [doi: 10.5306/wjco.v7.i1.54] [Medline: 26862491]
7. Walsh M, Cadoo K, Salo-Mullen E, Dubard-Gault M, Stadler Z, Offit K. Genetic factors: hereditary cancer predisposition syndromes. In: Abeloff's Clinical Oncology. Amsterdam, Netherlands: Elsevier; 2020:180-208.
8. Garber JE, Offit K. Hereditary cancer predisposition syndromes. J Clin Oncol 2005 Jan 10;23(2):276-292. [doi: 10.1200/JCO.2005.10.042] [Medline: 15637391]
9. Nagy R, Sweet K, Eng C. Highly penetrant hereditary cancer syndromes. Oncogene 2004 Aug 23;23(38):6445-6470. [doi: 10.1038/sj.onc.1207714] [Medline: 15322516]
10. Friedman AA, Letai A, Fisher DE, Flaherty KT. Precision medicine for cancer with next-generation functional diagnostics. Nat Rev Cancer 2015 Dec;15(12):747-756 [FREE Full text] [doi: 10.1038/nrc4015] [Medline: 26536825]
11. Nath AS, Pal A, Mukhopadhyay S, Mondal KC. A survey on cancer prediction and detection with data analysis. Innovations Syst Softw Eng 2019 Aug 22;16(3-4):231-243. [doi: 10.1007/s11334-019-00350-6]
12. Cancer survival in England: national estimates for patients followed up to 2017. Office for National Statistics. URL: https://www.ons.gov.uk/releases/cancersurvivalinenglandadultstageatdiagnosisandchildhoodpatientsfollowedupto2017 [accessed 2021-05-11]
13. Hawkes N. Cancer survival data emphasise importance of early diagnosis. BMJ 2019 Jan 25;364:l408. [doi: 10.1136/bmj.l408] [Medline: 30683652]
14. Tests for cancer of unknown primary. American Cancer Society. URL: https://www.cancer.org/cancer/cancer-unknown-primary/detection-diagnosis-staging/how-diagnosed.html [accessed 2021-05-11]
15. Losa F, Soler G, Casado A, Estival A, Fernández I, Giménez S, Seguí. SEOM clinical guideline on unknown primary cancer (2017). Clin Transl Oncol 2018 Jan;20(1):89-96 [FREE Full text] [doi: 10.1007/s12094-017-1807-y] [Medline: 29230692]
16. Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, et al. Electronic health records to facilitate clinical research. Clin Res Cardiol 2017 Jan;106(1):1-9 [FREE Full text] [doi: 10.1007/s00392-016-1025-6] [Medline: 27557678]
17. Denaxas SC, Morley KI. Big biomedical data and cardiovascular disease research: opportunities and challenges. Eur Heart J Qual Care Clin Outcomes 2015 Jul 01;1(1):9-16. [doi: 10.1093/ehjqcco/qcv005] [Medline: 29474568]
18. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
19. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1:18 [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
20. Prince S. Computer Vision: Models, Learning, and Inference. Cambridge, UK: Cambridge University Press; 2012.
21. Zong N, Wong RSN, Yu Y, Wen A, Huang M, Li N. Drug-target prediction utilizing heterogeneous bio-linked network embeddings. Brief Bioinform 2021 Jan 18;22(1):568-580. [doi: 10.1093/bib/bbz147] [Medline: 31885036]
22. Pai S, Bader GD. Patient similarity networks for precision medicine. J Mol Biol 2018 Sep 14;430(18 Pt A):2924-2938 [FREE Full text] [doi: 10.1016/j.jmb.2018.05.037] [Medline: 29860027]
23. Bender D, Sartipi K. HL7 FHIR: an Agile and RESTful approach to healthcare information exchange. In: Proceedings of the 26th IEEE international symposium on computer-based medical systems. 2013 Presented at: IEEE international symposium on computer-based medical systems; June 20-22; Porto, Portugal. [doi: 10.1109/cbms.2013.6627810]
24. Grover A, Leskovec J. node2vec: scalable feature learning for networks. KDD 2016 Aug;2016:855-864 [FREE Full text] [doi: 10.1145/2939672.2939754] [Medline: 27853626]
25. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016 Feb 17:899-908 [FREE Full text] [doi: 10.1093/jamia/ocv189] [Medline: 26911829]
26. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. Stud Health Technol Inform 2006;121:279-290. [Medline: 17095826]
27. McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem 2003 Apr;49(4):624-633 [FREE Full text] [Medline: 12651816]
28. International Statistical Classification of Diseases and Related Health Problems (ICD). World Health Organization. URL: http://www.who.int/classifications/icd/en/ [accessed 2021-05-11]

29. Decker S, Melnik S, van Harmelen F, Fensel D, Klein M, Broekstra J, et al. The Semantic Web: the roles of XML and RDF. IEEE Internet Comput 2000;4(5):63-73. [doi: 10.1109/4236.877487]
30. Kaggal VC, Elayavilli RK, Mehrabi S, Pankratz JJ, Sohn S, Wang Y, et al. Toward a learning health-care system - knowledge delivery at the point of care empowered by big data and NLP. Biomed Inform Insights 2016;8(Suppl 1):13-22 [FREE Full text] [doi: 10.4137/BII.S37977] [Medline: 27385912]
31. Slee VN. The international classification of diseases: ninth revision (ICD-9). Ann Intern Med 1978 Mar 01;88(3):424. [doi: 10.7326/0003-4819-88-3-424]
32. Liu S, Wei Ma, Moore R, Ganesan V, Nelson S. RxNorm: prescription for electronic drug information exchange. IT Prof 2005 Sep;7(5):17-23. [doi: 10.1109/mitp.2005.122]
33. Hong N, Wen A, Shen F, Sohn S, Wang C, Liu H, et al. Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data. JAMIA Open 2019 Dec;2(4):570-579 [FREE Full text] [doi: 10.1093/jamiaopen/ooz056] [Medline: 32025655]
34. 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]
35. Sohn S, Clark C, Halgrim SR, Murphy SP, Chute CG, Liu H. MedXN: an open source medication extraction and normalization tool for clinical text. J Am Med Inform Assoc 2014;21(5):858-865 [FREE Full text] [doi: 10.1136/amiajnl-2013-002190] [Medline: 24637954]
36. Lin Y, Chen H, Brown RA. MedTime: a temporal information extraction system for clinical narratives. J Biomed Inform 2013 Dec;46 Suppl:S20-S28 [FREE Full text] [doi: 10.1016/j.jbi.2013.07.012] [Medline: 23911344]
37. Zong N, Sharma DK, Yu Y, Egan JB, Davila JI, Wang C, et al. Developing a FHIR-based framework for phenome wide association studies: a case study with a pan-cancer cohort. AMIA Jt Summits Transl Sci Proc 2020;2020:750-759 [FREE Full text] [Medline: 32477698]
38. O'Hara S, Draper B. Introduction to the bag of features paradigm for image classification and retrieval. arXiv. Preprint posted online on Jan 17, 2011.
39. Liaw A, Wiener M. Classification and regression by randomForest. R news 2002;2(3):18-22 [FREE Full text]
40. Huang Y, Li L. Naive Bayes classification algorithm based on small sample set. 2011 Presented at: IEEE International Conference on Cloud Computing and Intelligence Systems; September 15-17; Beijing, China. [doi: 10.1109/ccis.2011.6045027]
41. Zhou X, Liu K, Wong STC. Cancer classification and prediction using logistic regression with Bayesian gene selection. J Biomed Inform 2004 Aug;37(4):249-259 [FREE Full text] [doi: 10.1016/j.jbi.2004.07.009] [Medline: 15465478]
42. Noble WS. What is a support vector machine? Nat Biotechnol 2006 Dec;24(12):1565-1567. [doi: 10.1038/nbt1206-1565] [Medline: 17160063]
43. Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE. A survey of deep neural network architectures and their applications. Neurocomputing 2017 Apr;234:11-26. [doi: 10.1016/j.neucom.2016.12.038]
44. Lawrence S, Giles CL, Tsoi AC, Back AD. Face recognition: a convolutional neural-network approach. IEEE Trans Neural Netw 1997;8(1):98-113. [doi: 10.1109/72.554195] [Medline: 18255614]
45. Kipf T, Welling M. Semi-supervised classification with graph convolutional networks. arXiv. Preprint posted online on February 22, 2017 [FREE Full text]
46. Azhagusundari B, Thanamani A. Feature selection based on information gain. Int J Innov Technol Explor Eng 2013;2(2):18-21 [FREE Full text]
47. node2vec. GitHub. URL: https://github.com/aditya-grover/node2vec [accessed 2011-05-11]
48. Machine Learning Group. LIBLINEAR -- a library for large linear classification. Taiwan University. URL: https://www.csie.ntu.edu.tw/~cjlin/liblinear/ [accessed 2021-05-11]
49. Weka. The University of Waikato. URL: https://www.cs.waikato.ac.nz/ml/weka/ [accessed 2021-05-10]
50. Chang CC, Line CJ. LIBSVM -- a library for support vector machines. Taiwan University. URL: https://www.csie.ntu.edu.tw/~cjlin/libsvm/ [accessed 2021-05-10]
51. Keras API reference. Keras. URL: https://keras.io/api/ [accessed 2021-05-10]
52. specktral. GitHub. URL: https://github.com/danielegrattarola/spektral [accessed 2021-05-10]
53. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning. 2006 Presented at: 23rd International Conference on Machine Learning; June 25-29; Pittsburgh, Pennsylvania p. 233-240. [doi: 10.1145/1143844.1143874]
54. Nguyen G, Bouzerdoum A, Phung S. Learning pattern classification tasks with imbalanced data sets. In: Yin PY, editor. Pattern Recognition. London, United Kingdom: InTech Open; 2009:193-208.
55. Jeni LA, Cohn JF, De La Torre F. Facing imbalanced data recommendations for the use of performance metrics. Int Conf Affect Comput Intell Interact Workshops 2013;2013:245-251 [FREE Full text] [doi: 10.1109/ACII.2013.47] [Medline: 25574450]
56. Roc. GitHub. URL: https://github.com/kboyd/Roc [accessed 2021-05-10]

57. Holmes G, Donkin A, Witten I. WEKA: a machine learning workbench. 1994 Presented at: Second Australian and New Zealand Conference on Intelligent Information Systems; November 29-December 2; Brisbane. [doi: 10.1109/anziis.1994.396988]
58. Woolson R. Wilcoxon signed-rank test. In: D'Agostino R, Massaro J, Sullivan L, editors. Wiley Encyclopedia of Clinical Trials. Hoboken, New Jersey: John Wiley & Sons, Inc; 2007.
59. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. 2017 Presented at: Advances in Neural Information Processing Systems 30; December 4-9; Long Beach, California.
60. Yang Y, Huang X, Zhou L, Deng T, Ning T, Liu R, et al. Clinical use of tumor biomarkers in prediction for prognosis and chemotherapeutic effect in esophageal squamous cell carcinoma. BMC Cancer 2019 May 31;19(1):526 [FREE Full text] [doi: 10.1186/s12885-019-5755-5] [Medline: 31151431]
61. Scarà S, Bottoni P, Scatena R. CA 19-9: biochemical and clinical aspects. Adv Exp Med Biol 2015;867:247-260. [doi: 10.1007/978-94-017-7215-0_15] [Medline: 26530370]
62. Høgdall E. Cancer antigen 125 and prognosis. Curr Opin Obstet Gynecol 2008 Feb;20(1):4-8. [doi: 10.1097/GCO.0b013e3282f2b124] [Medline: 18196998]
63. Zhang J, Huang T, Zhang F, Xu J, Chen G, Wang X, et al. Prognostic role of serum carbohydrate antigen 19-9 levels in patients with resectable hepatocellular carcinoma. Tumour Biol 2015 Apr;36(4):2257-2261. [doi: 10.1007/s13277-014-2435-6] [Medline: 25787748]
64. Wu H, Xiao J, Xiao S, Cheng Y. KRAS: a promising therapeutic target for cancer treatment. Curr Top Med Chem 2019;19(23):2081-2097. [doi: 10.2174/1568026619666190905164144] [Medline: 31486755]
65. Hankey W, Frankel WL, Groden J. Functions of the APC tumor suppressor protein dependent and independent of canonical WNT signaling: implications for therapeutic targeting. Cancer Metastasis Rev 2018 Mar;37(1):159-172 [FREE Full text] [doi: 10.1007/s10555-017-9725-6] [Medline: 29318445]
66. Du W, Lin C, Chen W. High expression of is an unfavorable prognostic biomarker in T4 gastric cancer patients. World J Gastroenterol 2019 Aug 21;25(31):4452-4467 [FREE Full text] [doi: 10.3748/wjg.v25.i31.4452] [Medline: 31496624]
67. Yassin A, AlRumaihi K, Alzubaidi R, Alkadhi S, Al Ansari A. Testosterone, testosterone therapy and prostate cancer. Aging Male 2019 Dec;22(4):219-227. [doi: 10.1080/13685538.2018.1524456] [Medline: 30614347]
68. Michaud JE, Billups KL, Partin AW. Testosterone and prostate cancer: an evidence-based review of pathogenesis and oncologic risk. Ther Adv Urol 2015 Dec;7(6):378-387 [FREE Full text] [doi: 10.1177/1756287215597633] [Medline: 26622322]
69. Hershman JM. Falling levels of thyroglobulin antibody after treatment for DTC predict no structural recurrence. Clin Thyroidol 2016 Mar;28(3):79-81. [doi: 10.1089/ct.2016;28.79-81]
70. Peiris AN, Medlock D, Gavin M. Thyroglobulin for monitoring for thyroid cancer recurrence. JAMA 2019 Mar 26;321(12):1228. [doi: 10.1001/jama.2019.0803] [Medline: 30912839]
71. Kim ES, Lim DJ, Baek KH, Lee JM, Kim MK, Kwon HS, et al. Thyroglobulin antibody is associated with increased cancer risk in thyroid nodules. Thyroid 2010 Aug;20(8):885-891. [doi: 10.1089/thy.2009.0384] [Medline: 20465529]
72. Santhanam P, Ladenson PW. Surveillance for differentiated thyroid cancer recurrence. Endocrinol Metab Clin North Am 2019 Mar;48(1):239-252. [doi: 10.1016/j.ecl.2018.11.008] [Medline: 30717906]
73. Jin A, Xu J, Wang Y. The role of TERT promoter mutations in postoperative and preoperative diagnosis and prognosis in thyroid cancer. Medicine (Baltimore) 2018 Jul;97(29):e11548 [FREE Full text] [doi: 10.1097/MD.0000000000011548] [Medline: 30024548]
74. Li X, Dai D, Chen B, Tang H, Xie X, Wei W. Clinicopathological and prognostic significance of cancer antigen 15-3 and carcinoembryonic antigen in breast cancer: a meta-analysis including 12,993 patients. Dis Markers 2018;2018:9863092 [FREE Full text] [doi: 10.1155/2018/9863092] [Medline: 29854028]
75. Liu H, Xu Y, Xiang J, Long L, Green S, Yang Z, et al. Targeting alpha-fetoprotein (AFP)-MHC complex with CAR T-cell therapy for liver cancer. Clin Cancer Res 2017 Jan 15;23(2):478-488 [FREE Full text] [doi: 10.1158/1078-0432.CCR-16-1203] [Medline: 27535982]
76. Bethune G, Bethune D, Ridgway N, Xu Z. Epidermal growth factor receptor (EGFR) in lung cancer: an overview and update. J Thorac Dis 2010 Mar;2(1):48-51 [FREE Full text] [Medline: 22263017]
77. Widschwendter M, Jones A, Evans I, Reisel D, Dillner J, Sundström K, FORECEE (4C) Consortium. Epigenome-based cancer risk prediction: rationale, opportunities and challenges. Nat Rev Clin Oncol 2018 May;15(5):292-309. [doi: 10.1038/nrclinonc.2018.30] [Medline: 29485132]
78. Lee C, Peng C, Li R, Chen Y, Tsai H, Hung Y, et al. Risk evaluation for the development of cervical intraepithelial neoplasia: development and validation of risk-scoring schemes. Int J Cancer 2015 Jan 15;136(2):340-349 [FREE Full text] [doi: 10.1002/ijc.28982] [Medline: 24841989]
79. Teschendorff AE, Jones A, Fiegl H, Sargent A, Zhuang JJ, Kitchener HC, et al. Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. Genome Med 2012 Mar 27;4(3):24 [FREE Full text] [doi: 10.1186/gm323] [Medline: 22453031]
80. Machine learning repository. UCI. URL: https://archive.ics.uci.edu/ml/index.php [accessed 2021-05-14]

81. Huang M, Chen C, Lin W, Ke S, Tsai C. SVM and SVM ensembles in breast cancer prediction. PLoS One 2017;12(1):e0161501 [FREE Full text] [doi: 10.1371/journal.pone.0161501] [Medline: 28060807]
82. Kumari M, Singh V. Breast cancer prediction system. Procedia Computer Science 2018;132:371-376. [doi: 10.1016/j.procs.2018.05.197]
83. Mostavi M, Chiu Y, Huang Y, Chen Y. Convolutional neural network models for cancer type prediction based on gene expression. BMC Med Genomics 2020 Apr 03;13(Suppl 5):44 [FREE Full text] [doi: 10.1186/s12920-020-0677-2] [Medline: 32241303]
84. Xiao Y, Wu J, Lin Z, Zhao X. A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data. Comput Methods Programs Biomed 2018 Nov;166:99-105. [doi: 10.1016/j.cmpb.2018.10.004] [Medline: 30415723]
85. Hou Q, Bing Z, Hu C, Li M, Yang K, Mo Z, et al. RankProd combined with genetic algorithm optimized artificial neural network establishes a diagnostic and prognostic prediction model that revealed C1QTNF3 as a biomarker for prostate cancer. EBioMedicine 2018 Jun;32:234-244 [FREE Full text] [doi: 10.1016/j.ebiom.2018.05.010] [Medline: 29861410]
86. Liu J, Wang X, Cheng Y, Zhang L. Tumor gene expression data classification via sample expansion-based deep learning. Oncotarget 2017 Dec 12;8(65):109646-109660 [FREE Full text] [doi: 10.18632/oncotarget.22762] [Medline: 29312636]
87. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep 2016 Dec 17;6:26094 [FREE Full text] [doi: 10.1038/srep26094] [Medline: 27185194]
88. Wang X, Zhang Y, Hao S, Zheng L, Liao J, Ye C, et al. Prediction of the 1-year risk of incident lung cancer: prospective study using electronic health records from the State of Maine. J Med Internet Res 2019 May 16;21(5):e13260 [FREE Full text] [doi: 10.2196/13260] [Medline: 31099339]
89. Wu Y, Burnside E, Cox J, Fan J, Yuan M, Yin J. Breast cancer risk prediction using electronic health records. 2017 Presented at: IEEE International Conference on Healthcare Informatics; August 23-36; Park City, Utah. [doi: 10.1109/ichi.2017.62]
90. Muhammad W, Hart GR, Nartowt B, Farrell JJ, Johung K, Liang Y, et al. Pancreatic cancer prediction through an artificial neural network. Front Artif Intell 2019 May 3;2(2):1. [doi: 10.3389/frai.2019.00002]
91. Wan J, Chen B, Kong Y, Ma X, Yu Y. An early intestinal cancer prediction algorithm based on deep belief network. Sci Rep 2019 Nov 22;9(1):17418 [FREE Full text] [doi: 10.1038/s41598-019-54031-2] [Medline: 31758076]
92. Pavlidis N, Pentheroudakis G. Cancer of unknown primary site. Lancet 2012 Apr 14;379(9824):1428-1435. [doi: 10.1016/S0140-6736(11)61178-1] [Medline: 22414598]
93. Mikolov T, Karafiát M, Burget L, ?ernocký J, Khudanpur S. Recurrent neural network based language model. 2010 Presented at: Eleventh Annual Conference of the International Speech Communication Association; September 26-30; Makuhari, Japan.
94. Gers FA, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM. Neural Comput 2000 Oct;12(10):2451-2471. [doi: 10.1162/089976600300015015] [Medline: 11032042]
95. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
96. cancer-prediction-on-fhir-rdf. GitHub. URL: https://github.com/fhircat/cancer-prediction-on-fhir-rdf [accessed 2021-05-10]

Abbreviations

AUPRC: area under the precision–recall curve
AUROC: area under the receiver operating characteristic curve
ICD: International Statisitical Classification of Diseases
FHIR: Fast Healthcare Interoperability Resources
RDF: Resource Description Framework
ReLU: rectified linear unit

Edited by G Eysenbach; submitted 17.08.20; peer-reviewed by JF Silva, X Zhou; comments to author 13.11.20; revised version received 07.01.21; accepted 27.01.21; published 25.05.21.

Please cite as:
Zong N, Ngo V, Stone DJ, Wen A, Zhao Y, Yu Y, Liu S, Huang M, Wang C, Jiang G
Leveraging Genetic Reports and Electronic Health Records for the Prediction of Primary Cancers: Algorithm Development and Validation Study
JMIR Med Inform 2021;9(5):e23586
URL: https://medinform.jmir.org/2021/5/e23586
doi:10.2196/23586
PMID:34032581

Corrigenda and Addenda

Correction: Extracting Family History Information From Electronic Health Records: Natural Language Processing Analysis

Maciej Rybinski[1], PhD; Xiang Dai[1,2], MSc; Sonit Singh[1,3], MSc; Sarvnaz Karimi[1], PhD; Anthony Nguyen[4], PhD

[1]Commonwealth Scientific and Industrial Research Organisation, Sydney, Australia

[2]University of Sydney, Sydney, Australia

[3]Macquarie University, Sydney, Australia

[4]Commonwealth Scientific and Industrial Research Organisation, Brisbane, Australia

Corresponding Author:
Maciej Rybinski, PhD
Commonwealth Scientific and Industrial Research Organisation
Marsfield
Sydney
Australia
Phone: 61 293724222
Email: maciek.rybinski@csiro.au

Related Article:

Correction of: https://medinform.jmir.org/2021/4/e24020

(JMIR Med Inform 2021;9(5):e30153) doi:10.2196/30153

In “Extracting Family History Information From Electronic Health Records: Natural Language Processing Analysis” (JMIR Med Inform 2021;9(4):e24020) one correction was made.

Due to a system error, five extraneous figures were added to the Methods section of the paper at the time of publication. These have been removed from the corrected version.

The correction will appear in the online version of the paper on the JMIR Publications website on May 3, 2021, 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 03.05.21; this is a non–peer-reviewed article;accepted 03.05.21; published 03.05.21.

Please cite as:
Rybinski M, Dai X, Singh S, Karimi S, Nguyen A
Correction: Extracting Family History Information From Electronic Health Records: Natural Language Processing Analysis
JMIR Med Inform 2021;9(5):e30153
URL: https://medinform.jmir.org/2021/5/e30153
doi:10.2196/30153
PMID:33939618

Corrigenda and Addenda

Correction: A Patient Journey Map to Improve the Home Isolation Experience of Persons With Mild COVID-19: Design Research for Service Touchpoints of Artificial Intelligence in eHealth

Qian He[1*], BSc; Fei Du[1*], BSc; Lianne W L Simonse[1*], MSc, PhD

Department of Design Organisation & Strategy, Faculty of Industrial Design Engineering, Delft University of Technology, Delft, Netherlands

[*] all authors contributed equally

Corresponding Author:
Lianne W L Simonse, MSc, PhD
Department of Design Organisation & Strategy
Faculty of Industrial Design Engineering
Delft University of Technology
Landbergstraat 15
Delft, 2628CE
Netherlands
Phone: 31 15 27 ext 89054
Email: L.W.L.Simonse@tudelft.nl

Related Article:

Correction of: https://medinform.jmir.org/2021/4/e23238/

(JMIR Med Inform 2021;9(5):e29794) doi:10.2196/29794

In “A Patient Journey Map to Improve the Home Isolation Experience of Persons With Mild COVID-19: Design Research for Service Touchpoints of Artificial Intelligence in eHealth” (JMIR Med Inform 2021;9(4):e23238) the authors noted one error.

In the originally published manuscript, Multimedia Appendix captions incorrectly appeared as follows:

> *Multimedia Appendix 1: Patient journey map of persons with mild COVID-19 during home isolation.*
>
> *Multimedia Appendix 2: Visual summary.*
>
> *Multimedia Appendix 3: Video purpose and comments coding trees.*
>
> *Multimedia Appendix 4: Personal video story coverage and experienced symptoms during home isolation.*

In the corrected version of the manuscript, Multimedia Appendix captions have been corrected to:

> *Multimedia Appendix 1: Personal video story coverage and experienced symptoms during home isolation.*
>
> *Multimedia Appendix 2: Patient journey map of persons with mild COVID-19 during home isolation.*
>
> *Multimedia Appendix 3: Video purpose and comments coding trees.*
>
> *Multimedia Appendix 4: Visual summary of design research.*

The correction will appear in the online version of the paper on the JMIR Publications website on May 4, 2021, 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 20.04.21; this is a non–peer-reviewed article;accepted 20.04.21; published 04.05.21.

Please cite as:
He Q, Du F, Simonse LWL
Correction: A Patient Journey Map to Improve the Home Isolation Experience of Persons With Mild COVID-19: Design Research for Service Touchpoints of Artificial Intelligence in eHealth
JMIR Med Inform 2021;9(5):e29794
URL: https://medinform.jmir.org/2021/5/e29794
doi:10.2196/29794
PMID:33945495

Original Paper

Predicting Semantic Similarity Between Clinical Sentence Pairs Using Transformer Models: Evaluation and Representational Analysis

Mark Ormerod[1], BEng; Jesús Martínez del Rincón[1], PhD; Barry Devereux[1], PhD

[1]Institute of Electronics, Communications & Information Technology, School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, United Kingdom

Corresponding Author:
Mark Ormerod, BEng
Institute of Electronics, Communications & Information Technology
School of Electronics, Electrical Engineering and Computer Science
Queen's University Belfast
Queen's Road
Queen's Island
Belfast, BT3 9DT
United Kingdom
Phone: 44 28 9097 1700
Email: mormerod01@qub.ac.uk

Abstract

Background: Semantic textual similarity (STS) is a natural language processing (NLP) task that involves assigning a similarity score to 2 snippets of text based on their meaning. This task is particularly difficult in the domain of clinical text, which often features specialized language and the frequent use of abbreviations.

Objective: We created an NLP system to predict similarity scores for sentence pairs as part of the Clinical Semantic Textual Similarity track in the 2019 n2c2/OHNLP Shared Task on Challenges in Natural Language Processing for Clinical Data. We subsequently sought to analyze the intermediary token vectors extracted from our models while processing a pair of clinical sentences to identify where and how representations of semantic similarity are built in transformer models.

Methods: Given a clinical sentence pair, we take the average predicted similarity score across several independently fine-tuned transformers. In our model analysis we investigated the relationship between the final model's loss and surface features of the sentence pairs and assessed the decodability and representational similarity of the token vectors generated by each model.

Results: Our model achieved a correlation of 0.87 with the ground-truth similarity score, reaching 6th place out of 33 teams (with a first-place score of 0.90). In detailed qualitative and quantitative analyses of the model's loss, we identified the system's failure to correctly model semantic similarity when both sentence pairs contain details of medical prescriptions, as well as its general tendency to overpredict semantic similarity given significant token overlap. The token vector analysis revealed divergent representational strategies for predicting textual similarity between bidirectional encoder representations from transformers (BERT)–style models and XLNet. We also found that a large amount information relevant to predicting STS can be captured using a combination of a classification token and the cosine distance between sentence-pair representations in the first layer of a transformer model that did not produce the best predictions on the test set.

Conclusions: We designed and trained a system that uses state-of-the-art NLP models to achieve very competitive results on a new clinical STS data set. As our approach uses no hand-crafted rules, it serves as a strong deep learning baseline for this task. Our key contribution is a detailed analysis of the model's outputs and an investigation of the heuristic biases learned by transformer models. We suggest future improvements based on these findings. In our representational analysis we explore how different transformer models converge or diverge in their representation of semantic signals as the tokens of the sentences are augmented by successive layers. This analysis sheds light on how these "black box" models integrate semantic similarity information in intermediate layers, and points to new research directions in model distillation and sentence embedding extraction for applications in clinical NLP.

(JMIR Med Inform 2021;9(5):e23099) doi:10.2196/23099

KEYWORDS

natural language processing; biomedical NLP; transformer models; representation learning; clinical text

Introduction

Clinical Semantic Textual Similarity

Semantic textual similarity (STS) has long been an important task in natural language processing (NLP) research. Early work built document-level models for textual similarity that used an unsupervised approach, primarily for the purpose of indexing documents for search [1,2]. These models generally relied on the assumption that greater overlap in terms indicated greater interdocument similarity. This body of work was enriched by Lee et al [3] who also modeled similarity at the document level but elicited human semantic judgments of similarity to create a small data set of interest to NLP researchers and cognitive scientists. It was not until *SemEval-2012 Task 6* [4] that the first sentence-based STS data set was released, featuring 2000 training and 750 test sentence pairs that were rated by humans on a scale of 0-5 (from low to high similarity). Since then, there have been many new *SemEval* STS tasks, building on the initial task to encompass new domains of text [5] and cross-lingual similarity [6,7]. Researchers have used these models in a diverse set of applications such as discovering links between data sets [8] and identifying arguments in online discourse [9]. Recognizing both the potential of STS for processing eHealth records and the need for specialized data sets to account for clinical domain knowledge and handle the use of medical abbreviations, Rastegar-Mojarad et al [10] introduced a corpus of clinical sentence pairs that were assigned semantic similarity labels on a 0-5 scale by medical experts. This data set of 1068 annotated sentence pairs, as well as an expanded corpus of 174,629 unannotated sentence pairs, was released as MedSTS [11]. As with previous STS tasks, performance on this data set is measured by the Pearson correlation between the predicted labels and the ground-truth similarity scores. In general, the best systems in the *BioCreative/OHNLP Challenge STS task* used ensembles of traditional machine learning models and deep learning models [12], with the overall top-performing model achieving a correlation of 0.83 on the test set. The clinical STS task tackled in this paper, the *2019 n2c2/OHNLP Track on Clinical Semantic Textual Similarity* [13], uses an expansion of the *BioCreative/OHNLP Challenge STS task* data set.

Transformer Models

In this work we train different types of transformer language models [14]. One of the types of transformer models that we train is bidirectional encoder representations from transformers (BERT) [15], which uses a masked language modeling task to train fully on bidirectional context without the decoder component of the original transformer architecture. Recently there has been much work in further training BERT on data from specialized domains, including biomedical text [16] and clinical documents [16-18]. We also further fine-tune these models on the task of STS. The last type of transformer model that we fine-tune is XLNet [19], which performs autoregressive language modeling while also capturing bidirectional context by sampling different possible word orders.

Interpreting Deep Neural Networks

After we train our models, we explore the representations that they build of clinical semantic similarity to identify any systematic biases or heuristics they may have learned that we can then work toward addressing to improve future clinical STS transformer architectures. There is a substantial literature that uncovers the kind of linguistic representations deep neural networks learn by experimentally perturbing the model's input and carefully analyzing the failure cases [20-22]. Another approach uses "decoding" to try to predict task-relevant information from intermediate representations generated from the model [23-25]. Recently there has been further work on interpreting the representations in deep neural models using attention weights [26,27]. While this approach is intuitive, there is still an ongoing debate about the extent to which the attention mechanism can be used to interpret a model's decision-making process [28,29]. As such, we focus our layer-wise analysis on our models' hidden token vectors [24]. Other relevant work on layer-wise analyses of BERT representations include [30] and [31].

One method we use to analyze the representational geometry of our models is representational similarity analysis (RSA) [32], which compares models that represent stimuli using vectors with different numbers of dimensions by measuring the correlation of second-order dissimilarity matrices with each other (ie, how dissimilar each pair of sentences is to each other pair by some metric). RSA has been used recently to analyze linguistic properties of deep learning models [33,34]. We use basic RSA to correlate various representations that we extract from each layer of our fine-tuned models with a matrix that corresponds to the ground-truth dissimilarity patterns found in the test set. This allows us to measure the strength of a clinical semantic signal through the layers of our networks and compare this signal across both models and choices of representation. We also employ a version of RSA that involves reweighting and linearly recombining the representational dissimilarity matrices (RDMs) [35] to build a representational model that best explains the ground-truth dissimilarity patterns in the test set. To our knowledge, this is the first use of this framework to explore the representational space of a deep neural language model.

Contributions

This work presents the following contributions:

- A transformer ensemble that achieves very competitive results on a new clinical STS task (with predictions producing a correlation of 0.87 with ground-truth similarity scores compared with the state-of-the-art correlation of 0.9), serving as a very strong deep learning baseline for this task.
- An extensive qualitative analysis of the transformer ensemble's error cases in the task of clinical semantic similarity that highlights the inability of popular transformer models to capture fine-grained differences between

medicinal sentence pairs, despite being trained on clinical or biomedical text.

- A quantitative error analysis framework for STS that reveals the shallow heuristics that transformer models learn to rely on for this task.
- The application of linear decoding and RSA to measure the semantic similarity signal in intermediate token representations of 5 popular transformer models, showing convergent and divergent representational strategies that reflect the models' performance on this task.
- The first application (to the authors' knowledge) of a reweighted and recombined version of RSA to neural language models, indicating that better representations of sentence pairs may be synthesized by combining 2 layers from a relatively poorly performing biomedical transformer with a simple textual feature signal, and suggesting new directions for research in sentence embedding extraction.

Methods

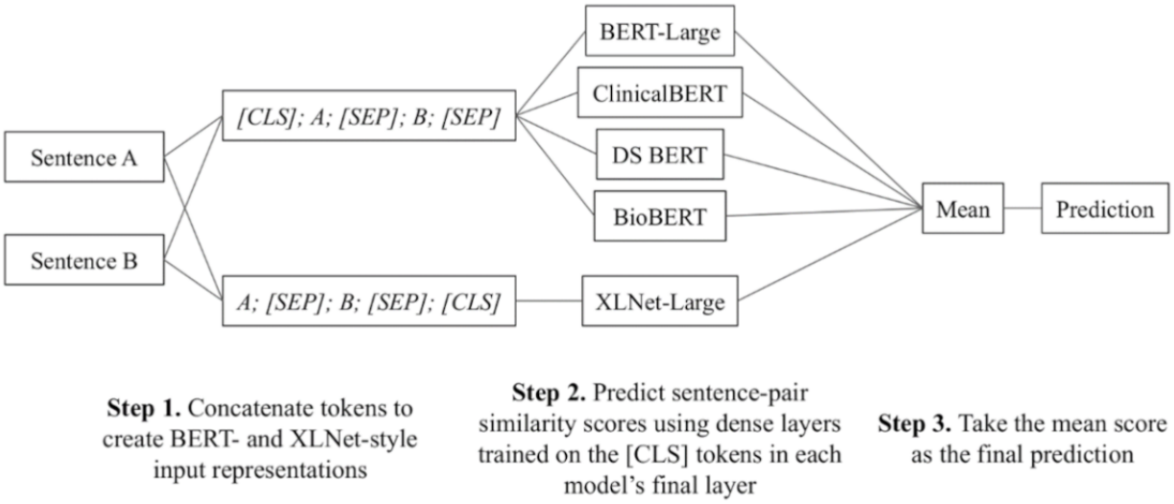
Data

The training data for this task were made up of 1642 sentence pairs and their associated similarity scores and the test set was made up of 412 sentence pairs. The similarity scores are floats on a scale of 0 to 5, ranging from no similarity to semantically identical. The annotations were performed by 2 medical experts (Donna Ihrke and Gang Liu [13]). The task is evaluated by the Pearson correlation between the predictions of a model and the ground-truth similarity scores.

Models

We fine-tuned 5 transformer [14] models. These include BERT-Large [15], 3 variants of BERT that were fine-tuned on text from the clinical domain, and XLNet-Large [19]. The 3 BERT variants were BioBERT [16], ClinicalBERT [17,18], and Discharge Summary BERT (DS BERT) [17,18]. We also created a *mean_score* model by taking the average prediction of the 5 transformer models. A linear layer was added on top of the pooled output for each model to perform the regression. The input for the BERT models was *[CLS]* + *A* + *[SEP]* + *B* + *[SEP]*, where *[CLS]* is the classification token, *A* and *B* are the 2 text snippets, and *[SEP]* is the separator token. The input for XLNet was *A* + *[SEP]* + *B* + *[SEP]* + *[CLS]*. We set the maximum sequence length for each model to 128. As we add 3 additional tokens to the input, any sentence pairs with over 125 tokens in total were shortened. This affected 5 sentence pairs, all of which were in the training set (with an average of 7.6 removed tokens). Each model was trained over 23 epochs using a batch size of 32. These models were trained using the PyTorch-Transformers library [36]. Our system architecture is depicted in Figure 1. We submitted the predictions of 3 models for evaluation on the n2c2 2019 Track 1 task: those from ClinicalBERT, XLNet, and the mean_score model.

Figure 1. Our system architecture for predicting the semantic textual similarity between two sentences using an ensemble of five Transformer models.


Sentence A
Sentence B
[CLS]; A; [SEP]; B; [SEP]
A; [SEP]; B; [SEP]; [CLS]
BERT-Large
ClinicalBERT
DS BERT
BioBERT
XLNet-Large
Mean
Prediction
Step 1. Concatenate tokens to create BERT- and XLNet-style input representations
Step 2. Predict sentence-pair similarity scores using dense layers trained on the [CLS] tokens in each model's final layer
Step 3. Take the mean score as the final prediction

Results

Overview

Our best performing model, the mean_score ensemble, achieved a correlation of 0.87, reaching 6th place out of 33 teams in the n2c2 2019 Track 1 task. The best model on the task achieved a correlation of 0.9 [37]. Our results are presented in Table 1.

The correlation between the predictions of each of 5 transformer models with all others is presented in Table 2. While the 3 models that have been fine-tuned with biomedical or clinical text (BioBERT, ClinicalBERT, and DS BERT) are more correlated with each other than with both XLNet and BERT, the predictions of all models generally correlate strongly with each other.

Table 1. Pearson correlation between the ground-truth labels and the predicted labels for each model.

Model	BERT	BioBERT	ClinicalBERT	DS BERT	XLNet	Mean score
Correlation	0.817	0.855	0.854	0.867	0.837	0.870

Table 2. Correlation between the predictions of each transformer model on the test set.

Model	BERT	BioBERT	ClinicalBERT	DS BERT	XLNet
BERT	1	0.92	0.92	0.92	0.91
BioBERT	0.92	1	0.95	0.96	0.92
ClinicalBERT	0.92	0.95	1	0.96	0.92
DS BERT	0.92	0.96	0.96	1	0.93
XLNet	0.91	0.92	0.92	0.93	1

Error Analysis

Error Cases Investigation

Rather than only evaluating our transformer ensemble by the correlation between its predictions and the ground-truth similarity scores, we carried out an extensive investigation into the error cases of this ensemble to shed light on any trends in the biases and heuristics that the component models may have learned from the training data. In this endeavor we carried out both qualitative and quantitative error analyses. Both analyses use a measure of loss that is calculated as the squared error between the models' prediction and the ground-truth similarity score.

Qualitative Analysis

We first carried out a qualitative analysis by grouping the sentence pairs that were most difficult to predict for the transformer ensemble by the primary lexical, syntactic, or semantic feature that we consider to be most salient and distinguishing. By identifying common error clusters, we can better understand our models' biases and attempt to mitigate these issues in future iterations of the clinical STS system. A list of these error categories as well as example sentences can be found in Table 3. We took 100 sentence pairs from the test data set with the highest loss and manually analyzed them to find possible explanations for incorrect predictions. The main categories that were identified are shown in Figures 2 and 3. We divided the errors into 2 cases: those where the transformer ensemble overpredicted sentence similarity with respect to the ground truth (Figure 2, which includes 77 sentence pairs) and those where the models underpredicted sentence similarity (Figure 3, which includes 23 sentence pairs).

Table 3. Example sentence pairs and error type (ie, whether the transformer ensemble overpredicted or underpredicted semantic similarity with respect to the ground truth) for each error category selected for the qualitative analysis.

Error type	Category	Example sentence pair	Notes
Overprediction	Medical prescription	(1) Ibuprofen [MOTRIN] 400 mg tablet 1 tablet by mouth every 4 hours as needed. (2) Gabapentin [NEURONTIN] 300 mg capsule 1 capsule by mouth every bedtime.	
Overprediction	Lexical overlap	(1) Patient to call to schedule additional treatment sessions as needed otherwise patient dismissed from therapy. (2) Patient tolerated session without adverse reactions to therapy.	
Overprediction	Semantic overlap	(1) The client verbalized understanding and consented to the plan of care. (2) The patient consented to the possibility of blood transfusion.	Some semantic overlap despite low ground-truth similarity score of 0
Overprediction	Reuse of phrase template	(1) male who presents for evaluation of Knee Pain (right). (2) female who presents for evaluation of Ear Infection/ Ear Pain.	Common phrase structures often feature lexical overlap, as well as strong syntactic similarity
Overprediction	Similar punctuation	(1) "Left upper extremity: Inspection, palpation examined and normal." (2) "Abdomen: Liver and spleen, bowel sounds examined and normal."	Note quotation marks within original text
Overprediction	Unknown	(1) "Mental: Alert and oriented to person, place and time." (2) She demonstrated understanding and agreed to proceed as noted.	The ensemble predicted a score of 2.55/5 for this example sentence pair
Underprediction	Unknown	(1) He denies any shortness of breath or difficulty breathing. (2) Patient denies any chest pain or shortness of breath.	
Underprediction	Different punctuation	(1) "Thank you for choosing the Name, M.D.. care team for your health care needs!" (2) Thank you for choosing Location for your health care and wellness needs.	
Underprediction	Lack of lexical overlap	(1) The above has been discussed and reviewed in detail with the patient. (2) The family was advised that the content of this interview will be shared with the health care team.	Semantic similarity with little lexical overlap

Figure 2. Common categories of error for cases when the model over-predicts similarity as identified by manual analysis of the 100 worst predictions.

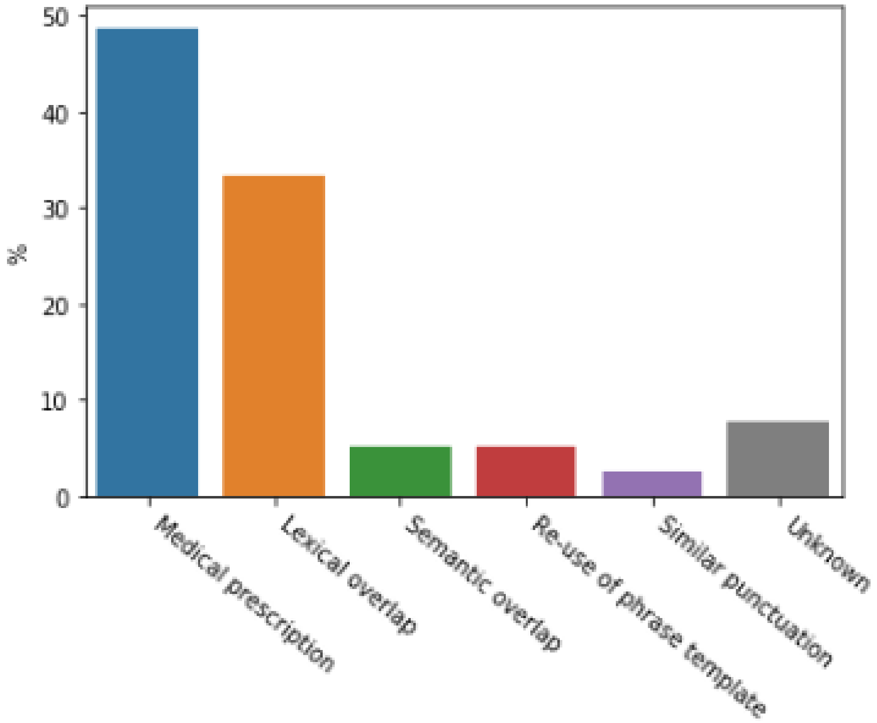
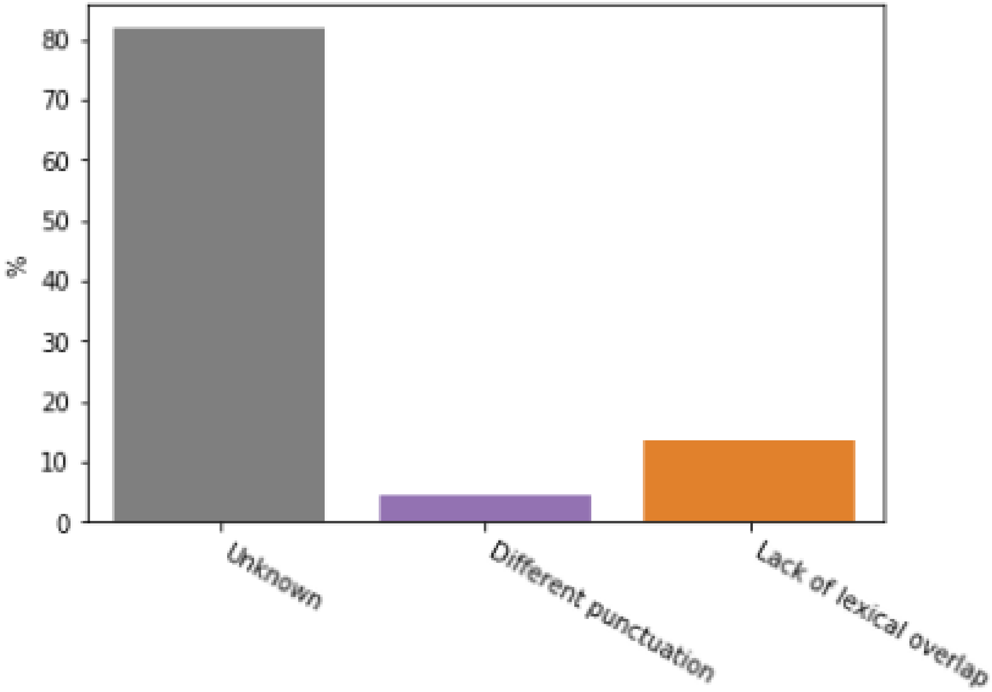


Figure 3. Common categories of error for cases when the model under-predicts similarity as identified by manual analysis of the 100 worst predictions.



Quantitative Analysis

To complement our qualitative analysis, we developed a simple STS quantitative analysis framework that allows us to investigate the relationship between surface features of the sentence pairs and our model's performance. This involves measuring the correlation between model loss and various features of the sentence pairs. In addition to providing the results for all labels, we present correlations (measured using Spearman rho) between the loss and pair features for each similarity score in the test set. The results are shown in Table 4. Below is an explanation of each sentence-pair feature that we investigated:

- Average sentence length: The total amount of tokens across the 2 sentences.
- Scaled total token frequency: The number of times each token in the sentence pair appears in the training set divided by the average sentence length, calculated after we removed stop words.

- Scaled unseen tokens per pair: The number of tokens in the sentence pair that do not appear in the training corpus, divided by the average sentence length.
- Scaled difference in token frequency: The difference between the training corpus token frequency across the 2 sentences, divided by the average sentence length, calculated after we removed stop words.
- Jaccard distance: The distance between the token sets of 2 sentences in a pair measured as

$$1 - (|A \cap B|)/(|A \cup B|)$$

Table 4. Correlation (Spearman rho) between the model's loss (mean score) per sentence pair and various sentence-pair features.

Label[a]	Average sentence length	Scaled total token frequency	Scaled unseen tokens per pair	Scaled difference in token frequency	Jaccard distance
All	−0.132	0.142	0.020	0.074	−0.025
0.0	−0.310	0.391	−0.263	0.219	−0.554 (<.001)[b]
0.5	0.102	−0.114	−0.249	−0.010	−0.202
1.0	0.067	−0.043	0.047	−0.033	−0.074
1.5	0.004	−0.151	0.033	−0.281	−0.153
2.0	0.118	0.441	0.012	0.354	−0.338
2.5	−0.018	0.014	−0.238	0.070	0.109
3.0	−0.453	0.432	−0.098	−0.026	0.119
3.5	−0.440	−0.051	0.257	−0.046	0.587
4.0	−0.088	0.138	0.268	0.052	0.171
4.5	−0.181	−0.266	−0.221	0.033	0.468
5.0	−0.040	0.789 (.042)	−0.242	0.590	0.596

[a] Labels are ground-truth similarity scores.

[b] Significant *P* value is reported in parenthesis after Bonferroni correction.

Layer-wise Token Representation Decoding

Given the difficulty of analyzing how these models build representations of clinical STS by looking at their loss alone, we next performed a layer-wise decoding analysis by training linear regression models to predict between-sentence semantic similarity given representations from each transformer across different layers of the model. By decoding the semantic signal in the intermediate layers of each model, we can uncover the mechanisms that transformer models use to predict clinical semantic similarity. We can then investigate whether any representational strategies correspond to better performance on this task, shedding light on why certain constituent models of the transformer ensemble perform worse, and potentially indicating directions for sentence-pair embedding extraction for STS. In the case of 12-layer models we used each layer and in the case of the larger 24-layer models, we used every other layer. This allows for direct comparison of representations by relative depth through the network.

We chose a variety of representations to decode. As we have many tokens per sentence pair, there are many different possible ways to map this list of vectors to a fixed-length representation. We aimed to choose representations that can reveal potential strategies and heuristics that our models use to predict semantic similarity. In doing so, we may also reveal how different types of models (ie, those trained on clinical versus general domain text, or those with BERT/XLNet-style architectures) diverge or converge in their representational transformation strategies. The chosen representations were

- [CLS]: The token vector corresponding to the classification token input.
- avg_reps_concat: Concatenation of the mean-pooled token vector representations of sentences A and B.
- max_reps_concat: Concatenation of max-pooled token vectors within sentences A and B.
- sent_avg_difference: The absolute difference in average token vector representations in sentences A and B.
- sent_max_difference: The absolute difference in max-pooled token vector representations in sentences A and B.
- sent_a_avg_max_concat: Concatenation of mean- and max-pooled token vectors from sentence A.
- sent_b_avg_max_concat: Concatenation of mean- and max-pooled token vectors from sentence B.

The linear regression models were evaluated using 10-fold cross-validation. Table 5 shows the overall best representations for decoding similarity score. Figures 4 and 5 feature layer-wise correlation plots for representations based on the classification token vector (Figure 4) and the absolute difference between the average token vectors in each sentence (Figure 5).

Table 5. The overall top decoding scores ranked in descending order. All the top-performing representations were extracted from XLNet and are mostly made up of the concatenation of the max-/mean-pooled token representations in the 2 sentences that were extracted from middle-late layers.

Model	Representation	Layer	Correlation
XLNet-large	max_reps_concat	18	0.9
XLNet-large	sent_a_avg_max_concat	18	0.89
XLNet-large	avg_reps_concat	18	0.88
XLNet-large	max_reps_concat	20	0.88
XLNet-large	avg_reps_concat	16	0.88
XLNet-large	avg_reps_concat	20	0.88
XLNet-large	sent_b_avg_max_concat	18	0.87
XLNet-large	sent_b_avg_max_concat	14	0.87
XLNet-large	max_reps_concat	14	0.87
XLNet-large	max_reps_concat	16	0.87

Figure 4. Pearson correlation between linear regression models' predictions of a sentence pair's semantic similarity and the ground-truth score (10-fold cross-validated on test-set) using [CLS] token pair representations.

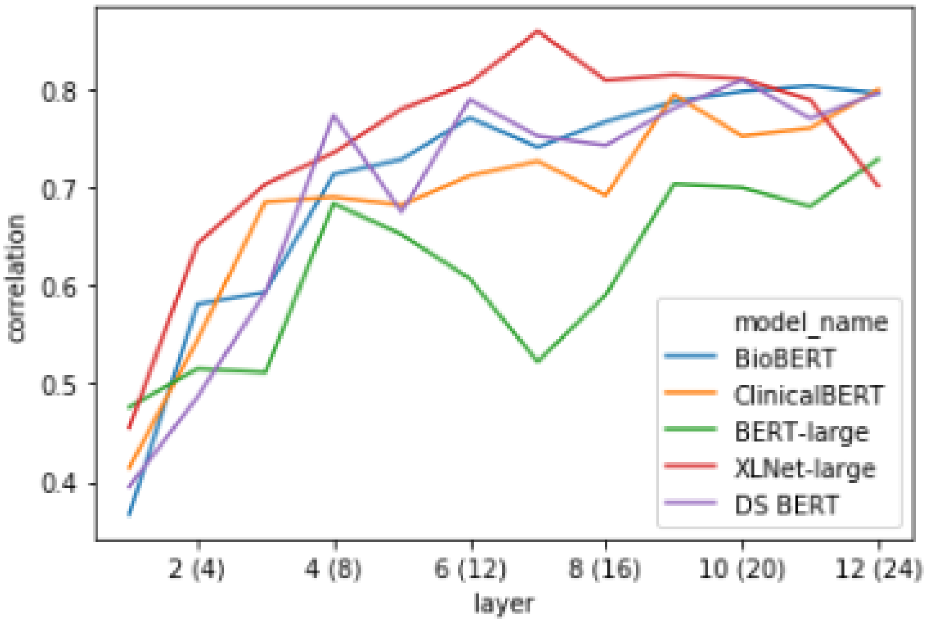
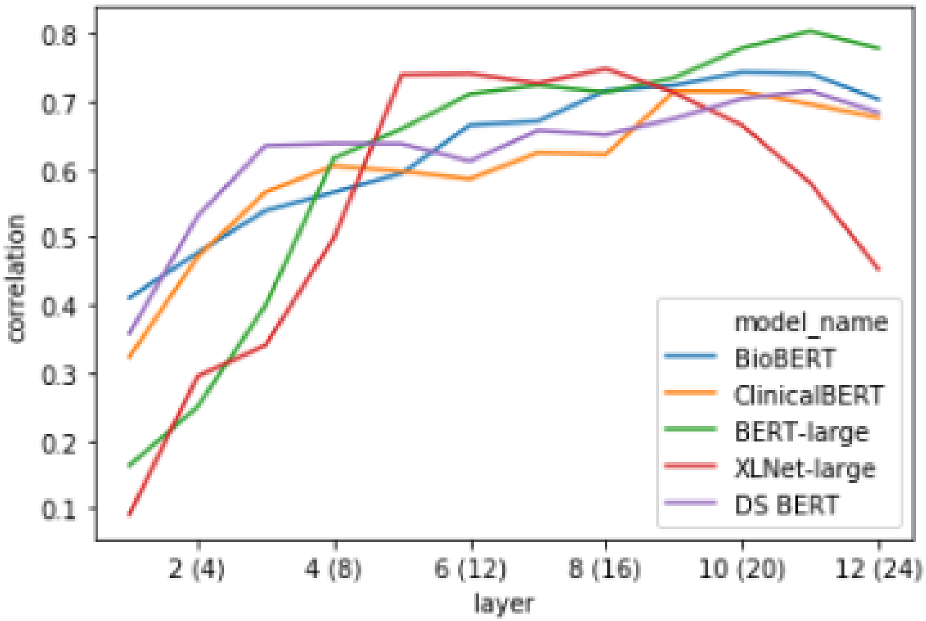


Figure 5. Pearson correlation between linear regression models' predictions of a sentence pair's semantic similarity and the ground-truth score (10-fold cross-validated on test-set) using the absolute difference between each sentence's mean-pooled token vector.



Representational Similarity Analysis

Overview

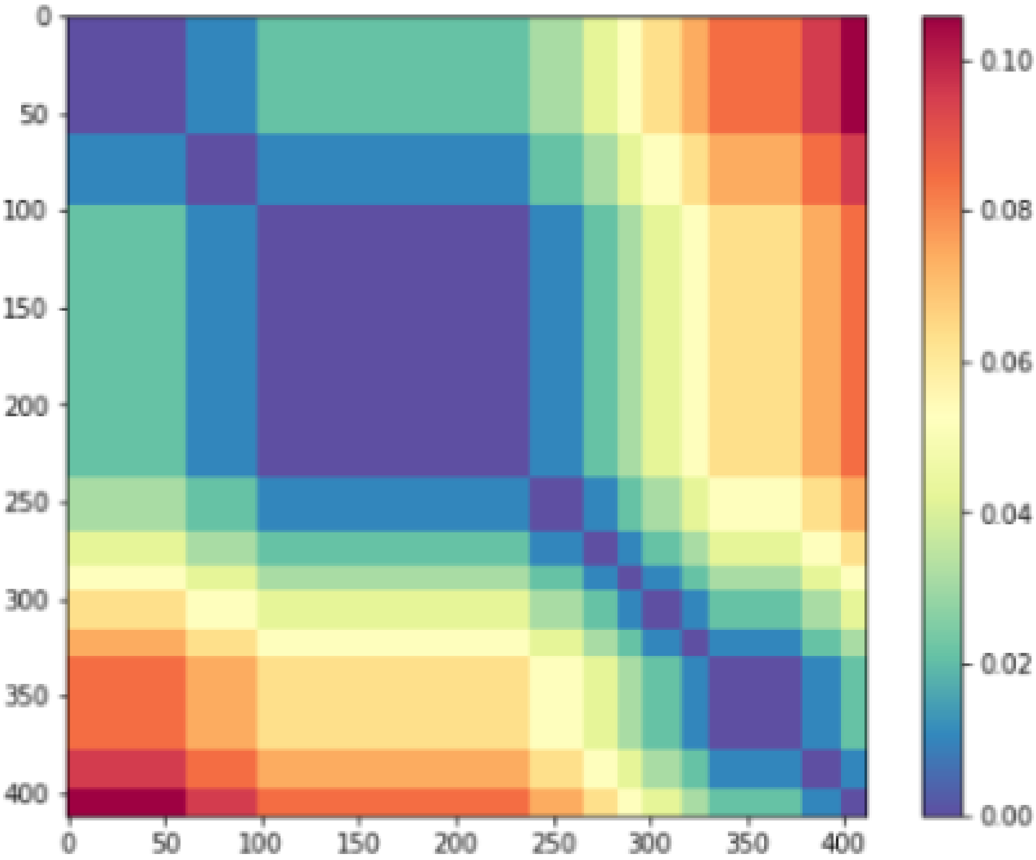
To find which representations learned by our models best explain the representational geometry of the semantic similarity task, we carried out 2 types of investigations within the framework of RSA. We use RSA to complement our layer-wise linear probing analysis, as it can reveal second-order representational patterns across many samples, while the layer-wise probing analysis relies on identifying particular dimensions of the representational space that predict semantic similarity. By taking these methods together, we can reach more robust conclusions about how transformer models build representations of semantic similarity and use this information to understand the performance of these models and identify how we can improve them. The data RDMs that we compared with the ground-truth RDM were extracted from each layer of each of the 5 transformer models, for each of the pair representations defined in the previous decoding analysis as well as 3 additional potential explanatory representations:

- avg_representation: The average across all token vectors.
- avg_sent_cosine_dist: The cosine distance between the mean-pooled token vector representations in sentences A and B.
- max_sent_cosine_dist: The cosine distance between the max-pooled token vector representations in sentences A and B.

Basic RSA

In our first RSA experiment, we performed a basic analysis in which we measure the Spearman correlation between a model RDM (calculated using the distance between all the samples in the test set measured by their ground-truth similarity score) and various representations elicited from our transformer models. Using the 412 test sentence pairs we produced the 412 × 412 matrix shown in Figure 6.

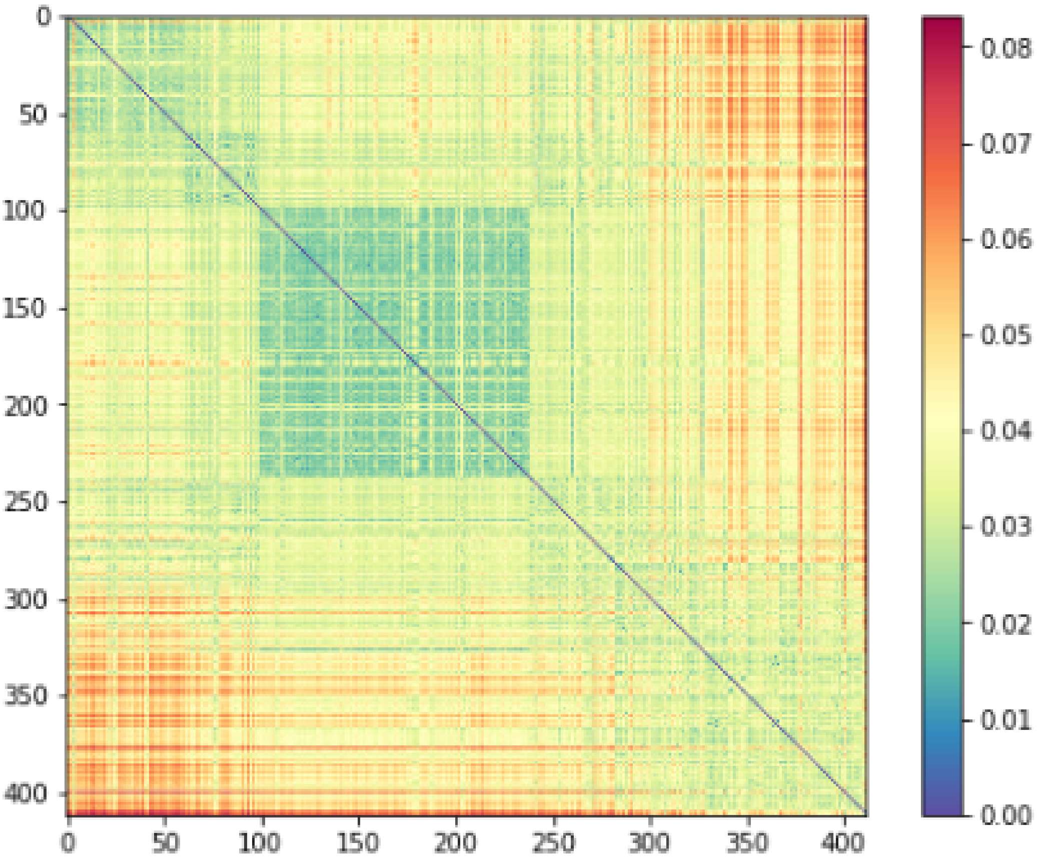
Figure 6. Model representational dissimilarity matrix for 412 test sentence pairs measured by distance between ground-truth semantic similarity scores. The dimensions of the dissimilarity matrix are sorted by each sentence-pair's ground-truth semantic similarity score.



Reweighted and Recombined RSA

We then found a combination of representations from all layers of each of the separate 5 transformer models and an RDM made up of text features (detailed in the "Quantitative Analysis" section) that best explains the ground-truth model when linearly recombined. Each explanatory RDM in a given trial had an associated weight that altogether summed to 1. These weights were found using a non-negative least squares (NNLS) solver using 10-fold cross-validation. This analysis revealed that the best performing explanation model was BioBERT. The final BioBERT-reweighted explanatory RDM is shown in Figure 7.

Figure 7. The final best-fitting re-weighted and linearly re-combined explanatory model found using NNLS and representations from BioBERT, achieving a correlation of 0.54 with the ground-truth model. The dimensions of the dissimilarity matrix are sorted by each sentence-pair's ground-truth semantic similarity score.



Layer-wise Reweighted RSA

In the final part of our reweighted RSA, we revisited the representations of BERT-Large to investigate why the classification token suddenly becomes less representative of the ground-truth similarity score around layers 12-16 as measured by linear regression probing (Figure 4) and RSA correlation (Figure 8). We reran the NNLS solver for the BERT-Large representations (using 10-fold cross-validation) but this time we excluded the text features RDM and used token vectors from only 1 layer at a time. We performed this analysis for the even layers, from layers 2 to 24 (as we had previously extracted every other layer of the 24-layer models to directly compare representations with 12-layer models based on relative depth through the network), and retrieved the values used to reweight the RDM for each layer. The plot of weights associated with each representation can be seen in Figure 9.

Figure 8. Correlation between the ground-truth model RDM and explanatory RDMs constructed from [CLS] token pair representations.

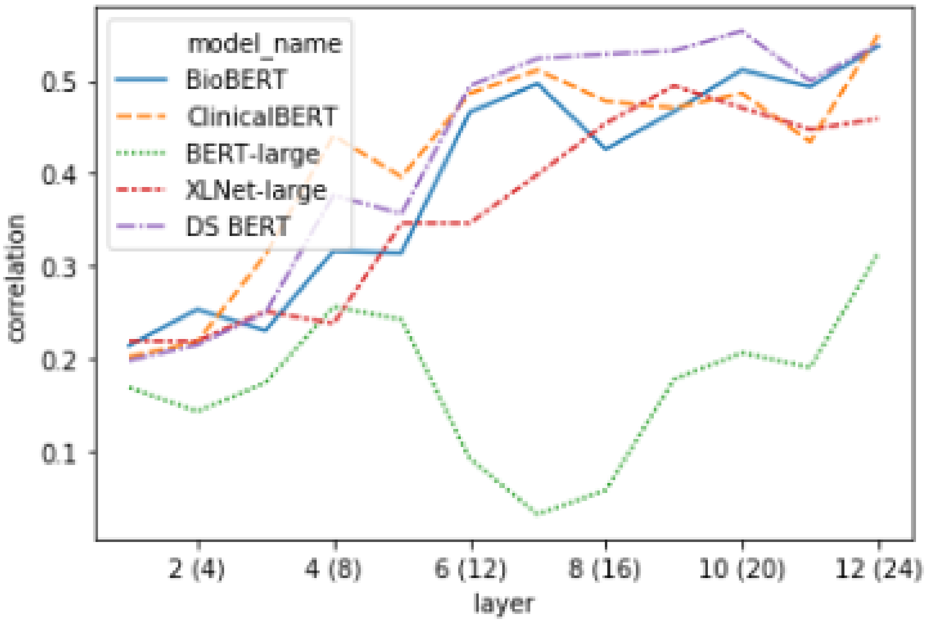
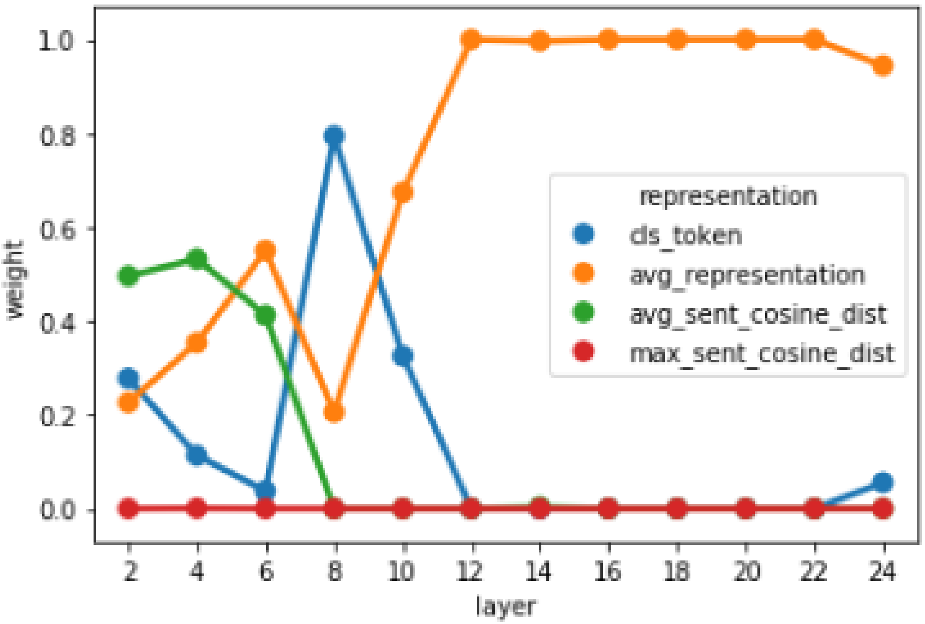


Figure 9. Weights associated with sentence-pair representations of BERT-Large found using NNLS to minimise the distance between a linearly re-combined set of RDMs and the ground-truth model RDM for each layer.



Discussion

Principal Results

Qualitative Error Analysis

In the case of sentence pairs that caused our ensemble to overpredict semantic similarity (Figure 2), the most obvious problem with our ensemble was its failure to model the semantic similarity of 2 sentences which contain details of medical prescriptions. This is likely because our models do not have the advanced level of domain knowledge necessary to correctly model this problem. As these sentences are usually very similar (apart from the name of a drug and dosage), the models overpredict similarity. The second biggest issue when overpredicting similarity is when there is a lexical overlap without semantic overlap. This suggests that our models over-rely on surface features such as token overlap. In most cases when our model underpredicts similarity, there is no obvious possible explanation. However, in the interpretable samples the issue was usually that synonyms were used, again suggesting an over-reliance on lexical overlap, and potentially motivating a concept normalization preprocessing step. In any case, the qualitative approach to analysis error is relatively limited for interpreting the instances of underprediction of semantic similarity for this ensemble. This limitation is mitigated by the fact that overpredictions made up the majority of the largest errors (77 out of 100). By taking both the cases of underprediction and overprediction together, it is clear that simple heuristics, such as predicting similarity given lexical overlap, are prominent within the transformer ensemble, and that these transformer models still lack the ability to produce the extremely fine-grained clinical semantic representations that are required to implicitly calculate semantic distances between medical concepts (eg, particular drugs) given a relatively small task set. Any future work would have to address these issues; for example, by augmenting the data using a concept normalization preprocessing step, or by enriching the ensemble's domain knowledge by incorporating a clinical terms resource.

Quantitative Error Analysis

Overall, Table 4 shows a weak negative correlation between the average sentence length and loss. This relationship is relatively strong for entirely dissimilar sentence pairs and moderately similar sentence pairs and may be explained by the fact that longer sentences provide more contextual information that can be used to decide whether 2 sentences are semantically similar. Another trend is for the loss to increase with the scaled total token frequency (ie, how often the words in the pair appear in the training corpus), particularly in the case where the 2 sentences are semantically identical. This relationship is difficult to interpret, but additional analysis could investigate the extent to which the loss can be explained using the relative frequency of the words given a more general corpus (such as Wikipedia), to separate the effect of clinical term frequency.

We also see that Jaccard distance is negatively correlated with loss for sentence pairs that are less semantically similar and positively correlated with loss for pairs that are more semantically similar. One possible explanation for this observation is that our deep transformer models have learned an appropriate strategy of predicting low similarity scores given token overlap for the extreme case when sentence pairs are dissimilar and have little overlap. However, the model seems unable to apply such a shallow heuristic in cases where sentence pairs are very semantically similar. Further analysis showed Jaccard distance to be very significantly negatively correlated with the ground-truth label (*P*<.001), which may indicate that a deep ensemble model could benefit from the presence of traditional machine learning models that are trained on simple features of the text such as relative overlap between tokens.

The quantitative analysis approach has both verified the existence of overall heuristics that use surface features of the sentence pairs to predict semantic similarity as noted in the previous qualitative analysis and allowed to us examine these trends as they occur within certain ranges of semantic similarity scores. This approach to quantitative analysis of STS errors has thus produced a richer view of these biases, while still suggesting that these deep transformer models use a set of relatively shallow strategies for this task.

Layer-wise Token Representation Decoding

The first striking pattern to note in Figure 4 is that the BERT models tend to drop in performance on the CLS token task in the middle of the network, thereafter reaching their apexes (in the extreme case this is amplified in BERT-Large), whereas XLNet tends to steadily increase to its highest point before dropping off over the rest of the network. This indicates that in BERT-style models, the [CLS] token does not serve as the primary representation of semantic similarity in the middle layers. Second, the correlation between linear model predictions and ground-truth scores held-out folds almost always monotonically increases for the difference between average sentence representations for all BERT-style models (Figure 5). This contrasts with the performance on the XLNet sent_avg_diff representation, which caps half-way through the network, then drops off steadily beginning a few layers later. It appears that XLNet builds a good representation based on the mean-pooled token representations, but that this information is integrated in the middle of processing and subsequently discarded around layer 18.

All the top 10 best decoding scores across all representations were extracted from XLNet (Table 5). Overall, XLNet did best using the max_reps_concat, reaching a correlation of 0.9 in layer 18, which represents a 7.5% increase in that model's initial performance on the test set. This demonstrates that given the initial representations of a large deep model, it may be possible to increase its performance very inexpensively and massively on small amounts of held-out data using a simple linear model and the correct choice of representation.

It is clear from the linear decoding experiment that the representational strategies of the transformers fine-tuned with biomedical or clinical documents tend to align, with each model gradually building better representations of STS over the course of their layers in an almost always monotonic fashion, in both the [CLS] token and the absolute difference between mean-pooled sentence representations. This is in contrast to the relatively erratic differences between decodability over layers seen with BERT-Large and XLNet, where decodability will rapidly gain or fall over the course of 1-2 layers, especially when looking at the distance between mean-pooled sentence vectors representation. This result suggests that models with more clinical domain knowledge (and better performance on this task) learn to build robust representations of clinical semantic similarity (ie, not only using the [CLS] token or the distance between mean-pooled vectors) and that this information is gradually recovered in a steady, step-wise manner.

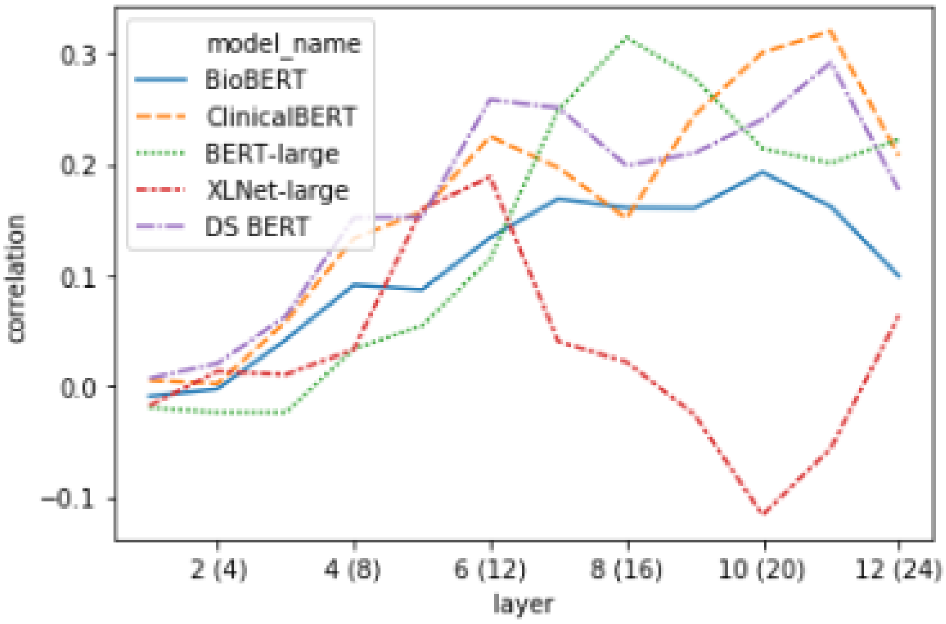
Representational Similarity Analysis

Basic RSA

In carrying out the single-correlation RSA task, we found confirmation for some of the representational trends identified during the decoding task. Two of such trends are presented in Figures 8 and 10, which include the correlation of the model RDM with data RDMs built using classification tokens (Figure 8) and the absolute difference between average token vectors from the 2 sentences in a pair (Figure 10). As was previously shown in Figure 4, BERT-Large diverges drastically from the other models in how representative the classification token is of a sentence pair's semantic similarity score around layers 12-16, while all other models generally generate progressively better [CLS] tokens throughout the network, with only slight loss in performance around the middle of the network. The performance of BERT-Large [CLS] representations on this task again reflects its final score, which was the lowest of the 5 models. We further analyzed the representational geometry of

BERT-Large in our reweighting analysis later in the current section to better understand this observation. The confirmation of this considerable drop in decodability performance shows that this trend does not simply reflect the inability of the linear regression models to predict semantic information due to the small amount of data. Likewise, the correlation plot featured in Figure 10 presents more evidence for our previous finding that BERT-style models seem to represent across-sentence similarity by minimizing the average difference in token vectors. While these correlations are positive from layers 4 to 12, this signal is not as strong as would be indicated by the probing analysis, suggesting that this strategy may not be a primary heuristic. In any case, taken together, these 2 layer-wise correlation plots show that the probing task produces robust metrics of representational trends, and that probing and basic RSA are complementary approaches to the analysis of transformations in token vectors of deep transformer models.

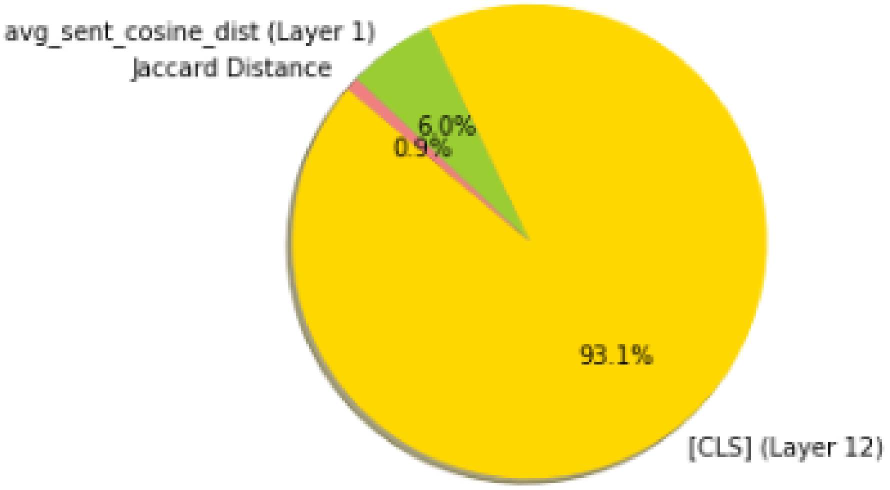
Figure 10. Correlation between the ground-truth model RDM and explanatory RDMs constructed using the absolute difference between each sentence's mean-pooled token vector.



Reweighted and Recombined RSA

After performing the next stage in our RSA, reweighting and recombining a set of RDMs (using all layers using all representations, as well as the text features RDM) for each transformer to minimize the distance between the new RDM and the ground-truth representation, we found that the best choice of model was BioBERT. Figure 7 shows visual confirmation that much of the ground-truth dissimilarity patterning (Figure 6) has been reproduced by this explanatory model. This result was somewhat unexpected, given that this model did not perform best on the test set. This finding suggests that when generating sentence-pair vectors, it may in some cases be better to reweight and combine representations from runner-up models, rather than using the single best model. The weights learned for each RDM in the BioBERT model (Figure 11) show that the RDM is mostly made up of the final layer's [CLS] token, although it has been reweighted using the cosine distance between the average token vector of the 2 sentences in a given pair and the Jaccard distance between the 2 sentences. We believe that beyond revealing how well each representation explains the ground-truth semantic similarity, this technique has promising potential for generating sentence embeddings for downstream tasks.

Figure 11. Proportion of weights learned for the best explanatory model (which used BioBERT representations and text features).



Layer-wise Reweighted RSA

By looking at the weights learned for each component of the layer-wise BERT explanatory model (Figure 9), we find that after layer 8, the weight associated with the average token representation drastically increases and this representation becomes dominant for the remaining layers, whereas the explanatory weight of the [CLS] token peaks at layer 8 before rapidly declining. We link this result to our finding that the worst linear probing and RSA correlations for BERT's [CLS] tokens start to occur after layer 8 (Figures 4 and 8). This suggests that in middle to late layers, BERT-Large focuses on building better mean-pooled representations of the sentence pairs, an interpretation which is in line with the dramatic increase in correlation between BERT-Large's representations and the ground-truth model when using the absolute difference between the average token vector of each sentence as the data RDM (Figure 10). This interpretation is also compatible with the increase in linear regression performance when using BERT-Large token vectors and taking the absolute difference between the average token vectors in each sentence as input (Figure 5).

Limitations and Future Work

While we employed the use of cross-validation for our linear probing and NNLS RSA tasks, it should be noted that our test set of 412 sentence pairs represents a relatively small amount of data and as such it may be difficult to assess whether our results would generalize to more data-rich contexts. One potential method for partially mitigating this problem would be to cross-validate our results across the full set of 2054 sentence pairs, rather than restricting the analysis to the original test set from the clinical STS task. While this approach may lead to insights into the robustness of our interpretation, we consider it to be outside of the scope of this work as we aim to analyze the errors and representational strategies that both result from the inductive biases of transformer models and reflect biases learned from the task's data. Restricting our analysis to the original 412 sentence-pair test set thus enables direct comparison with other models trained on the same data. Another issue with cross-validating across the whole data set is that we will always be limited to a relatively small amount of data for this task, as even testing on a slice of 50% of the total data would still only allow for 1027 sentence pairs for evaluation. It could also be insightful to carry out our analysis on models trained using larger general domain semantic similarity tasks that feature more sentence pairs. We again consider this line of research to be out of scope for this work.

In future work we wish to investigate to what extent we can directly use a layer's token representations to automatically learn interpretable explanations that minimize the distance between a reweighted RDM and the ground-truth model RDM. We expect that incorporating our models' attention weights will be essential at that level of analysis. Additionally, we wish to set alternative target RDMs to examine how we can recombine the token vectors in a sentence pair to best explain the model's classification token, thereby further exploring the inner representational dynamics of fine-tuned transformer models.

Conclusion

We tackled a recent clinical STS task using a variety of transformer models, including both those trained on general domain language and models that were further trained on clinical text. After achieving a high correlation between the predictions of a mean-pooled ensemble of these models and the test-set ground truth, we analyzed the error cases of our model both qualitatively and quantitatively, finding groups of semantically related sentences that are generally difficult for our transformers to model and identifying surface features of the sentence pair that significantly correlate with loss for particular ranges of the semantic similarity space. These findings suggest potential avenues for further improvement, for example, by augmenting our models to allow them to directly take traditional NLP textual features into account.

We then carried out 2 types of representational analyses, namely, linear decoding and RSA, to shed light on the heuristics on which these models have learned to rely. These approaches were shown to be complementary and revealed divergent representational strategies for predicting textual similarities between BERT-style models and XLNet. Furthermore, our search through the representational space for the best explanatory model of the ground-truth data suggests that a large amount of this information can be captured using a combination of a classification token and the cosine distance between sentence-pair representations in the first layer of a transformer model that did not produce the best predictions on the test set, suggesting interesting directions for research in model distillation and sentence embedding extraction.

Acknowledgments

This research was funded by a Northern Ireland Health and Social Care Board eHealth Directorate grant (grant no. 24F-1801) to BD.

Conflicts of Interest

None declared.

References

1. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci 1990 Sep;41(6):391-407. [doi: 10.1002/(sici)1097-4571(199009)41:6<391::aid-asi1>3.0.co;2-9]
2. Salton G. Automatic text processing: the transformation, analysis, and retrieval of information by computer. Choice Reviews Online 1989 Sep 01;27(01):27-0351-27-0351. [doi: 10.5860/choice.27-0351]
3. Lee M, Pincombe B, Welsh M. An Empirical Evaluation of Models of Text Document Similarity. In: Proceedings of the Annual Meeting of the Cognitive Science Society. 2005 Jul Presented at: Annual Conference of the Cognitive Science Society; 21-23 July 2005; Stresa, Italy p. 1254-1259 URL: https://escholarship.org/uc/item/48g155nq
4. Agirre E, Cer D, Diab M, Gonzalez-Agirre A. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity (SemEval@NAACL-HLT). 2012 Jun Presented at: *SEM 2012: The First Joint Conference on Lexical and Computational Semantics; June 7-8, 2012; Montréal, QC, Canada p. 285-393. [doi: 10.18653/v1/s17-2001]
5. Agirre E, Cer D, Diab M, Gonzalez-Agirre A, Guo W. *SEM 2013 shared task: Semantic Textual Similarity (*SEM@NAACL-HLT). 2013 Jun Presented at: Second Joint Conference on Lexical and Computational Semantics (*SEM); June 2013; Atlanta, GA, USA p. 32-43.
6. Agirre E, Banea C, Cer D, Diab M, Gonzalez-Agirre A, Mihalcea R, et al. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation (SemEval@NAACL-HLT). 2016 Jun Presented at: 10th International Workshop on Semantic Evaluation (SemEval-2016); June 2016; San Diego, CA p. 497-511. [doi: 10.18653/v1/s16-1081]
7. Cer D, Diab M, Agirre E, Lopez-Gazpio I, Specia L. SemEval-2017 Task 1: semantic textual similarity - multilingual and cross-lingual focused evaluation. arXiv. 2017 Aug. URL: https://arxiv.org/abs/1708.00055 [accessed 2021-04-19]
8. McCrae J, Buitelaar P. Linking datasets using semantic textual similarity. Cybernetics and information technologies. 2018 Mar. URL: https://sciendo.com/article/10.2478/cait-2018-0010 [accessed 2021-04-19]
9. Boltužić F, Šnajder J. Identifying prominent arguments in online debates using semantic textual similarity. In: Proceedings of the 2nd Workshop on Argumentation Mining. 2015 Jun 4 Presented at: NAACL HLT 2015; 4 June 2015; Denver, Colorado, USA p. 110-115. [doi: 10.3115/v1/w15-0514]
10. Rastegar-Mojarad M, Liu S, Wang Y, Afzal N, Wang L, Shen F, et al. Biocreative/OHNLP challenge 2018. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. 2018 Aug 15 Presented at: ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics; August 2018; Washington, DC p. 575-575. [doi: 10.1145/3233547.3233672]
11. Wang Y, Afzal N, Fu S, Wang L, Shen F, Rastegar-Mojarad M, et al. MedSTS: a resource for clinical semantic textual similarity. Lang Resources & Evaluation 2018 Oct 24;54(1):57-72. [doi: 10.1007/s10579-018-9431-1]
12. Wang Y, Afzal N, Liu S, Rastegar-Mojarad M, Wang L, Shen F, et al. Overview of the BioCreative/OHNLP challenge 2018 task 2: clinical semantic textual similarity. In: Proceedings of the BioCreative/OHNLP Challenge 2018. 2018 Aug Presented at: BioCreative/OHNLP Challenge 2018; August 29, 2018; Virtual Conference p. 575. [doi: 10.13140/RG.2.2.26682.24006]
13. Wang Y, Fu S, Shen F, Henry S, Uzuner O, Liu H. The 2019 n2c2/OHNLP Track on Clinical Semantic Textual Similarity: Overview. JMIR Med Inform 2020 Nov 27;8(11):e23375. [doi: 10.2196/23375] [Medline: 33245291]
14. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. 2017 Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); December 4-9, 2017; Long Beach, CA p. 1-11 URL: https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
15. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019 Jun Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 3-5, 2019; Minneapolis, MN, USA p. 4171-4186. [doi: 10.18653/v1/N19-1423]

16. 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;36(1):1234-1240. [doi: 10.1093/bioinformatics/btz682]
17. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv. 2019. URL: https://arxiv.org/abs/1904.05342 [accessed 2020-11-29]
18. Alsentzer E, Murphy J, Boag W, Weng W, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. arXiv. Jun. URL: https://arxiv.org/abs/1904.03323 [accessed 2021-04-19]
19. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le Q. XLNet: generalized autoregressive pretraining for language understanding. arXiv. 2019 Jun 19. URL: https://arxiv.org/abs/1906.08237 [accessed 2020-01-02]
20. Linzen T, Dupoux E, Goldberg Y. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. TACL 2016 Dec;4:521-535. [doi: 10.1162/tacl_a_00115]
21. Gulordava K, Bojanowski P, Grave E, Linzen T, Baroni M. Colorless green recurrent networks dream hierarchically. arXiv. 2018 Mar. URL: https://arxiv.org/abs/1803.11138 [accessed 2018-03-29]
22. McCoy R. Right for the wrong reasons: diagnosing syntactic heuristics in natural language inference. arXiv. 2019 Feb. URL: https://arxiv.org/abs/1902.01007 [accessed 2019-06-04]
23. Hupkes D, Veldhoen S, Zuidema W. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. Journal of Artificial Intelligence Research 2018 Apr 30:907-926 [FREE Full text] [doi: 10.24963/ijcai.2018/796]
24. van Aken B, Winter B, Löser A, Gers FA. How Does BERT Answer Questions?: A Layer-Wise Analysis of Transformer Representations. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019 Nov Presented at: ACM International Conference on Information and Knowledge Management; November 3-7, 2019; Beijing, China p. 1823-1832. [doi: 10.1145/3357384.3358028]
25. Tenney I, Xia P, Chen B, Wang A, Poliak A, McCoy RT, et al. What do you learn from context? Probing for sentence structure in contextualized word representations. arXiv. 2019 May 15. URL: https://arxiv.org/abs/1905.06316 [accessed 2019-05-15]
26. Lin Y, Tan Y, Frank R. Open sesame: getting inside BERT's linguistic knowledge. arXiv. 2019 Jun 04. URL: https://arxiv.org/abs/1906.01698 [accessed 2020-06-04]
27. Voita E, Serdyukov P, Sennrich R, Titov I. Context-aware neural machine translation learns Anaphora resolution. arXiv. 2018 May 25. URL: https://arxiv.org/abs/1805.10163 [accessed 2018-05-25]
28. Jain S, Wallace B. Attention is not Explanation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019 Jun Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 3-5, 2019; Minneapolis, MN, USA p. 3543-3556.
29. Wiegreffe S, Pinter Y. Attention is not not Explanation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019 Nov Presented at: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); November 3-7, 2019; Hong Kong, China p. 11-20. [doi: 10.18653/v1/D19-1002]
30. Reif E, Yuan A, Wattenberg M, Viegas FB, Coenen A, Pearce A, et al. Visualizing and measuring the geometry of BERT. In: Advances in Neural Information Processing Systems. 2019 Dec Presented at: Advances in Neural Information Processing Systems; December 2019; Vancouver, BC, Canada.
31. Hewitt J, Manning C. A Structural Probe for Finding Syntax in Word Representations. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019 Jun Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 3-7, 2019; Minneapolis, MN p. 4129-4138.
32. Kriegeskorte N, Mur M, Bandettini P. Representational similarity analysis - connecting the branches of systems neuroscience. Front Syst Neurosci 2008;2:4 [FREE Full text] [doi: 10.3389/neuro.06.004.2008] [Medline: 19104670]
33. Abnar S, Beinborn L, Choenni R, Zuidema W. Blackbox meets blackbox: Representational Similarity and Stability Analysis of Neural Language Models and Brains. arXiv. 2019 Jun 04. URL: https://arxiv.org/abs/1906.01539 [accessed 2019-06-04]
34. Abdou M, Kulmizev A, Hill F, Low DM, Søgaard A. Higher-order comparisons of sentence encoder representations. arXiv. 2019 Sep 1. URL: https://arxiv.org/abs/1909.00303 [accessed 2019-09-01]
35. Khaligh-Razavi S, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput Biol 2014 Nov;10(11):e1003915 [FREE Full text] [doi: 10.1371/journal.pcbi.1003915] [Medline: 25375136]
36. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's transformers: state-of-the-art natural language processing. arXiv. 2019 Oct 09. URL: https://arxiv.org/abs/1910.03771 [accessed 2019-10-09]
37. Mahajan D, Poddar A, Liang J, Lin Y, Prager J, Suryanarayanan P, et al. Identification of Semantically Similar Sentences in Clinical Notes: Iterative Intermediate Training Using Multi-Task Learning. JMIR Med Inform 2020 Nov 27;8(11):e22508 [FREE Full text] [doi: 10.2196/22508] [Medline: 33245284]

Abbreviations

BERT: bidirectional encoder representations from transformers
NLP: natural language processing
NNLS: non-negative least squares
RDM: representational dissimilarity matrix
RSA: representational similarity analysis
STS: semantic textual similarity

Edited by Y Wang; submitted 31.07.20; peer-reviewed by M Torii, L Ferreira; comments to author 13.11.20; revised version received 07.01.21; accepted 23.01.21; published 26.05.21.

Please cite as:
Ormerod M, Martínez del Rincón J, Devereux B
Predicting Semantic Similarity Between Clinical Sentence Pairs Using Transformer Models: Evaluation and Representational Analysis
JMIR Med Inform 2021;9(5):e23099
URL: https://medinform.jmir.org/2021/5/e23099
doi:10.2196/23099
PMID:34037527

Original Paper

Application of Intelligent Computer-Assisted Taylor 3D External Fixation in the Treatment of Tibiofibular Fracture: Retrospective Case Study

Hongfeng Sheng[1], MM; Weixing Xu[1], MD; Bin Xu[1], MM; Hongpu Song[1], MM; Di Lu[1], MM; Weiguo Ding[1], MM; Henry Mildredl[2], PhD

[1]Department of Orthopaedics, Tongde Hospital of Zhejiang Province, Hangzhou, China

[2]Federal Institute for Drugs and Medical Devices, Medical Devices Division, Bonn, Germany

Corresponding Author:
Hongpu Song, MM
Department of Orthopaedics
Tongde Hospital of Zhejiang Province
234 Gu-cui Road
Hangzhou, 310012
China
Phone: 86 0571 89972000
Email: hongpusongzj@yeah.net

Abstract

Background: With the development of modern society, severe and complex tibial fractures caused by high-energy injuries such as traffic accidents have gradually increased. At present, the commonly used methods for the treatment of tibial fractures include plate fixation, intramedullary nail fixation, and external fixation. Most of these fractures are open wounds with severe soft tissue injury and wound contamination, and some involve bone defects, which makes internal fixation treatment difficult.

Objective: This study aims to explore the use of intelligent computer-assisted Taylor 3D external fixation for the treatment of tibiofibular fractures.

Methods: In total, 70 patients were included and divided into the Taylor 3D external fixation (TSF) group (28 patients with severe tibial fractures treated with TSF) and the internal fixation group (42 patients with complicated tibiofibular fractures treated by internal fixation). After the treatment, the follow-up evaluation of TSF for the treatment of tibiofibular fractures noted the incidence of complications, as well as the efficacy and occurrence of internal fixation for the treatment of tibial fractures in our hospital.

Results: The results showed that TSF was superior to orthopedics in the treatment of tibiofibular fractures in terms of efficacy and complications.

Conclusions: TSF for the treatment of tibiofibular fractures is more effective than internal fixation and the incidence of complications is low. This is a new technology for the treatment of tibiofibular fractures that is worthy of clinical promotion.

(JMIR Med Inform 2021;9(5):e21455) doi:10.2196/21455

KEYWORDS

intelligent computer-assisted instruction; Taylor three-dimensional external fixation; tibial fracture; internal fixation; external fixation

Introduction

With the development of modern society, severe and complex tibial fractures caused by high-energy injuries such as traffic accidents have gradually increased. At present, the commonly used methods for the treatment of tibial fractures include plate fixation, intramedullary nail fixation, and external fixation. Most of these fractures are open wounds with severe soft tissue injury and wound contamination, and some involve bone defects, which makes internal fixation treatment difficult. Potential complications include postoperative wound infection, chronic osteomyelitis, delayed fracture healing, and fracture nonunion. The incidence of malunion healing is high, often resulting in treatment failure [1]. External fixation technology is a good method for the treatment of such fractures. External fixation can reduce the damage to soft tissue, and reduce the risk of postoperative wound infection, osteomyelitis, delayed fracture healing, and fracture nonunion. While complications occur, fracture fixation can be performed in the early stages after an injury, which provides a better prognosis for soft tissue repair, limb care, and early functional exercise [2]. However, in the past, external fixation stents for the treatment of complicated tibiofibular fractures have had poor stability for fracture fixation. They can only be used as a temporary fixation method. Most of the latter require secondary surgery to replace internal fixation, which makes the treatment period prolonged and significantly increases the cost of fracture treatment. Additionally, the fracture healing time is prolonged.

Based on the Ilizarov circular fixator (ICF), a previous study [3] applied the Stewart platform and Charles theory to the field of orthopedics, and combined these with computer software to invent the Taylor 3D space frame. Taylor 3D external fixation (TSF) is a good complement to the deficiencies of the ICF for multidimensional planar fractures and deformity correction. TSF has the following advantages in the treatment of tibial fractures: TSF is quick and easy to learn; accurate closed reduction of fractures can be achieved with computer software assistance during or after surgery; TSF is a better option for ensuring fixation stability of fractures; and the external fixator can be used as a long-term fixation method. The stent is maintained during the entire process of fracture healing. The needle is fixed during the installation process. It does not cause secondary damage to the local soft tissue. The risk of postoperative infection is low, and the rate of fracture nonunion is low. Fracture surgery can be performed soon after the injury to achieve early functional exercise. Postoperative bone defects can be repaired by adjusting the external frame. There are many reports on the use of TSF in the treatment of limb deformities, although there is little literature on the use of TSF in the treatment of tibial fractures.

To further explore the efficacy and possible complications of this technique in the treatment of severely complex fractures, this study retrospectively analyzed 28 cases of severe tibial fractures treated with TSF. This study will provide a theoretical basis for the clinical application and improvement of this technology. The follow-up data of 42 patients with severe tibiofibular fractures treated with internal fixation were compared with that of the TSF group to further evaluate the efficacy of TSF.

Methods

General Information

The TSF group included 28 patients with severe tibial fractures treated with TSF in our department from May 2015 to June 2018. These cases included 23 males and 5 females, aged 19 to 65 years (mean 38.5 years). These cases included 18 traffic accidents, 6 heavy bruises, 4 high fall injuries, 17 open fractures (according to Gustilo classification: 12 of type II and 5 of type III), and 11 closed fractures (according to Tscherne classification: there were 8 level 2 cases and 3 level 3 cases). According to the fracture line classification, there were 10 cases of transverse shape fracture line, 6 cases of oblique shape, 3 cases of spiral shape, 5 cases of comminuted fracture, and 4 cases of multiple fractures. According to the location of the fracture, there were 7 cases in the proximal one-third of the bone, 5 cases in the middle one-third, 11 cases in the middle and distal junctions, and 5 cases in the distal one-third. Compartment syndrome occurred and 4 cases underwent open decompression.

The internal fixation group included 42 patients with severe complicated tibiofibular fractures treated by internal fixation from January 2011 to March 2017, including 33 males and 9 females aged 17 to 70 years (mean 40.3 years old). There were 26 cases of traffic injuries, 10 cases of heavy bruises, 6 cases of high fall injuries, 22 cases of open fractures (according to Gustilo classification: 12 cases of type II, 10 cases of type III), 20 cases of closed fractures (according to Tscherne classification: 13 cases of grade 2, 7 cases of grade 3). According to fracture line classification, there were 16 cases of transverse fracture, 9 cases of oblique, 6 cases of spiral, 7 cases of comminuted, and 4 cases of multiple fractures. According to the location of fracture, there were 11 cases in the proximal one-third, 9 cases in the middle one-third, 14 cases in the middle and distal junctions, and 8 cases in the distal one-third. Compartment syndrome was treated with incision decompression in 6 cases. There were 18 cases treated with steel plates and 24 cases treated with intramedullary nails.

Inclusion Criteria and Exclusion Criteria

Inclusion criteria were the following: (1) high-intensity injury resulting in severe soft tissue injury or open severe complex tibiofibular fracture, Gustilo type II-III or Tscherne grade 2-3 and (2) follow-up time ≥6 months. Exclusion criteria were the following: (1) Simple low-energy tibiofibular fracture, Gustilo type I or Tscherne grade 1. (2) Total tibial plateau, pilon fracture, and other cumulative articular surface fracture patients. (3) Follow-up time <6 months. (4) Cases involving serious internal medicine. (5) Patients with interruption of follow-up or impaired case data. Finally, (6) patients with severe neurovascular injury.

Surgical Methods

TSF Group

Epidural anesthesia is often used in the TSF group, and general anesthesia can be used in patients with other combined injuries. The patient is often placed in a supine position to reduce ischemia and reperfusion injury; it is generally not recommended to use a tourniquet. Open wound treatment involves the following: emergency (6 to 8 hours) wound debridement, Taylor frame fixation, for a small wound surface; if contamination is not serious, the wound can be closed in one stage; if the wound is large, heavy contamination can be removed with a vacuum sealing drainage (VSD) negative pressure device. A second-stage skin graft or flap transfer is used to close the wound. Closed fracture treatment involves the following: generally, you do not need to wait for the swelling to subside; rather, the fracture can be fixed in the early stage of the Taylor frame, resulting in early exercise of the limbs, and gradual exercise after 2 to 3 weeks.

The Taylor external fixator installation procedure is the following: after the affected limb is sterilized, the C-arm machine monitors the axial traction of the distal end of the affected limb, roughly resets the fracture end displacement (shortening, angulation, and rotational displacement), and initially restores the length of the tibia. Cantering on the fracture line, a TSF ring is inserted into the distal and proximal fractures (if the fracture segment is ≥3 cm, ensure that each fracture segment is fitted with one ring, because the TSF ring and the tibia are placed ≥2 cm from the fracture line). When the anatomical axis is vertical, at least 2 full needles or olive needles are inserted into the safety channel of each level of the tibia, and connected with the TSF ring. If the stability is poor, a half needle can be implanted to increase stability and, according to the adjacent TSF ring, install 6 adjustable connecting rods. Under the perspective of the C-arm machine, the fracture displacement parameters were preliminarily calculated and the fracture was repaired by adjusting the 6 connecting rods. If the fracture is difficult to reset and the local soft tissue can be finitely cut at the fracture end, a comminuted fracture can be transformed into a relatively simple fracture by temporary fixation with Kirschner wire or by using plate fixation to maximize the recovery of the bone shaft. The tubular morphology was further reset after surgery with computer software.

Internal Fixation Group

The anesthesia method and the surgical position of the internal fixation group are the same as in the TSF group. Open fracture emergency (6 to 8 hours) wound debridement is performed, followed by temporary external fixation of the fractured unilateral outer frame; if there is a small wound surface and contamination is not serious, the wound can be closed in one stage; if the wound is large and contamination is heavy, VSD negative pressure is used and a second-stage skin graft or flap transfer is used to close the wound. After the soft tissue recovers, open reduction and internal fixation are performed. After the swelling of the closed fracture subsides (indicated by the appearance of dermatoglyphics) and the local soft tissue recovers, open reduction and internal fixation are performed once tension blisters have subsided. The open reduction and internal fixation process is the following: the surgical approach is determined according to the soft tissue condition and the type of fracture. In the process of fracture reduction, as much as possible is done to protect and reduce soft tissue damage, including avoiding using long incisions to pursue excessive anatomical reduction. For the reduction, it is required to fully reduce the longitudinal, axial, and rotational displacement of the tibia.

Postoperative Treatment

In the TSF group, the standard lateral radiograph was improved. The fracture displacement parameters were measured according to the x-ray film. The parameters were input into the TSF computer software system, and 6 adjustable connecting rods were used to make adjustments. After the fracture was reset, the film was reviewed again. If the reset was not good, the fracture displacement parameter could be measured again and imported into the computer software to adjust the parameters again until ideal. Postoperative nail dressing and regular dressing care were provided. Knee and ankle joint functional exercise began the first day after surgery, with gradual weight-bearing 2 to 3 weeks after surgery, and a monthly review following filming; the external fixator was dismantled after the fracture healed.

When evaluating fracture reduction and fixation, the internal fixation group had an improved positive lateral radiograph. After the operation, the affected limb promoted blood return. On the second day, active and passive functional exercises of the knee and ankle joints were increased. Regular incision dressing was provided (if the wound dressing had exudation, the dressing was changed), and the wound was not bandaged after exudation. The incision healing was complete 10 to 14 days after surgery. If the wound became infected, the secretion was assessed in the laboratory for bacteria, and an antibiotic was intravenously provided according to the result; the drug was changed frequently, and if necessary, vacuum suction treatment was used. The patient avoided weight-bearing activities for 6 weeks after surgery, and then gradually added weight with the help of progressive ablation. In this study, the x-ray films were reviewed in January, February, March, and June. One year later, fracture healing was judged according to the films. After the fracture healed, the internal fixation was removed.

Observation Indicators

The main follow-up details recorded were the patient's surgical preparation time, operation time, fracture healing time, total weight-bearing time, length of hospital stays and expenses, postoperative complications, as well as other indicators.

Statistical Methods

We used SPSS Statistics software (Version 21.0; SPSS Inc) for the following statistical analysis: the categorical variable data was analyzed by chi-square test, the countable data was analyzed by *t* test, and the test standard was α=.05.

Results

Clinical Follow-up Results of the TSF Treatment Group

All 28 cases were followed up for an average of 23.5 months (range 10-48 months); the average preoperative preparation time was 3.5 days (range 0.5-8 days); the average operation time was 112.3 minutes (range 90-131 minutes); and 4 cases of bone defects occurred. Bone grafting and internal fixation were used to obtain healing. In 3 cases of delayed fracture healing, late adjustment of the external frame fracture resulted in good healing. The fracture healing rate was 85.71%, with an average fracture healing time of 20.3 weeks (range 16-48 weeks). The external fixation frame was worn for an average of 26 weeks (range 17-48 weeks). Overall, 4 cases of compartment syndrome occurred, emergency decompression was given, the TSF external fixation frame was installed after the wound was closed, and the patient was discharged after adjustment and resetting. The average weight-bearing time was 90.5 days (range 65-180 days); the average number of days spent in hospital was 10.8 days (range 6-24 days); and the average hospitalization cost was 5.6

million (range 3.8-97 million). Postoperative wound infection occurred in 2 cases, and 13 cases occurred in the needle. Infections were cured after wound dressing and oral antibiotics. No cases of chronic osteomyelitis occurred. In 1 case, there was another fracture after the removal of the external frame, and the fracture was healed after internal fixation. At the last follow-up, all patients could step onto the ground and 21 patients could participate in daily housework. There were no patients with joint stiffness.

Clinical Follow-up Results of the Internal Fixation Group

A total of 42 patients were followed up for an average of 19.5 months (range 7-34 months). This included 18 patients in the plate fixation group and 24 patients in the intramedullary nail fixation group. The average preoperative preparation time was 10.5 days (range 6-24 days) and the average operation time was 152.4 minutes (range 120-185 minutes). Overall, 3 cases of nonunion occurred, which healed after internal bone grafting. In total, 4 cases had delayed fracture healing. The fracture healing rate was 92.86% and the average fracture healing time was 23.8 weeks (range 17-54 weeks). A total of 6 cases of compartment syndrome occurred; acute incision decompression was provided and internal fixation was performed after the closure of the wound. All healed well and the average time to weight-bearing was 110.3 days (range 60-185 days). The average hospital stay was 18.2 days (range 14-33 days) and hospitalization costs averaged 6.2 million (range 5.3-11.2 million). Postoperative wound infection occurred in 20 cases; infections were cured after dressing change, intravenous antibiotic, and/or VSD negative pressure treatment. In total, 5 cases of chronic osteomyelitis occurred. One patient's bone fractured after internal fixation. There was no joint stiffness among patients.

Comparison of the Efficacy of TSF and Internal Fixation in the Treatment of Severe Tibiofibular Fractures

The fracture healing rate was 85.71% (24/28) in the TSF treatment group and 92.86% (39/42) in the internal fixation group (Table 1). In the TSF treatment group, the time spent on preoperative preparation time, operation time, fracture healing time, total time to full weight-bearing, and hospitalization stays were shorter than those in the internal fixation group, and the hospitalization cost was lower; the difference was statistically significant (*P*<.05).

Table 1. Comparison of surgical-related indicators between the two groups ( ).

Group	TSF[a] group (n=28)	Internal fixation group (n=42)	*t* value	*P* value
Preoperative preparation time (days)	3.5 (1.8)	10.5 (3.2)	10.50	<.001
Operation time (minutes)	112.3 (27.5)	152.4 (40.3)	4.59	<.001
Fracture healing time (week)	20.3 (3.4)	23.8 (2.7)	4.79	<.001
Full weight-bearing time (days)	90.5 (10.3)	110.3 (14.5)	6.24	<.001
Hospital stay (days)	10.8 (2.8)	18.2 (3.1)	6.24	<.001
Hospitalization expenses (million)	5.6 (1.3)	6.8 (2.2)	2.33	.02

[a] TSF: Taylor 3D external fixation.

Comparison of Postoperative Complications Between the TSF and Internal Fixation Groups

The incidence of postoperative infection and osteomyelitis was lower in the TSF group than in the internal fixation group (*P*<.05). There was no significant difference in the probability of nonunion and refracture (*P*>.05; Table 2).

Table 2. Comparison of postoperative complications between the two groups.

Project	TSF[a] group (n=28)	Internal fixation group (n=42)	χ^2 value	*P* value
Postoperative infection	2 (7.14)	20 (47.62)	12.772	<.001
Delayed fracture healing	3 (10.71)	4 (9.52)	N/A[b]	>.99[c]
Nonunion	4 (14.29)	3 (7.14)	N/A	.43[c]
Osteomyelitis	0 (0.00)	5 (11.90)	N/A	.08[c]
Refracture	1 (3.57)	1 (2.38)	N/A	>.99[c]

[a] TSF: Taylor 3D external fixation.

[b] N/A: not applicable.

[c] This was determined using the Fisher exact probability method.

Typical Cases

Figures 1 and 2 showed the open fracture of left tibia and postoperative recovery of patient A.

Figure 1. Patient A: car accident leading to open fracture of the left tibia.

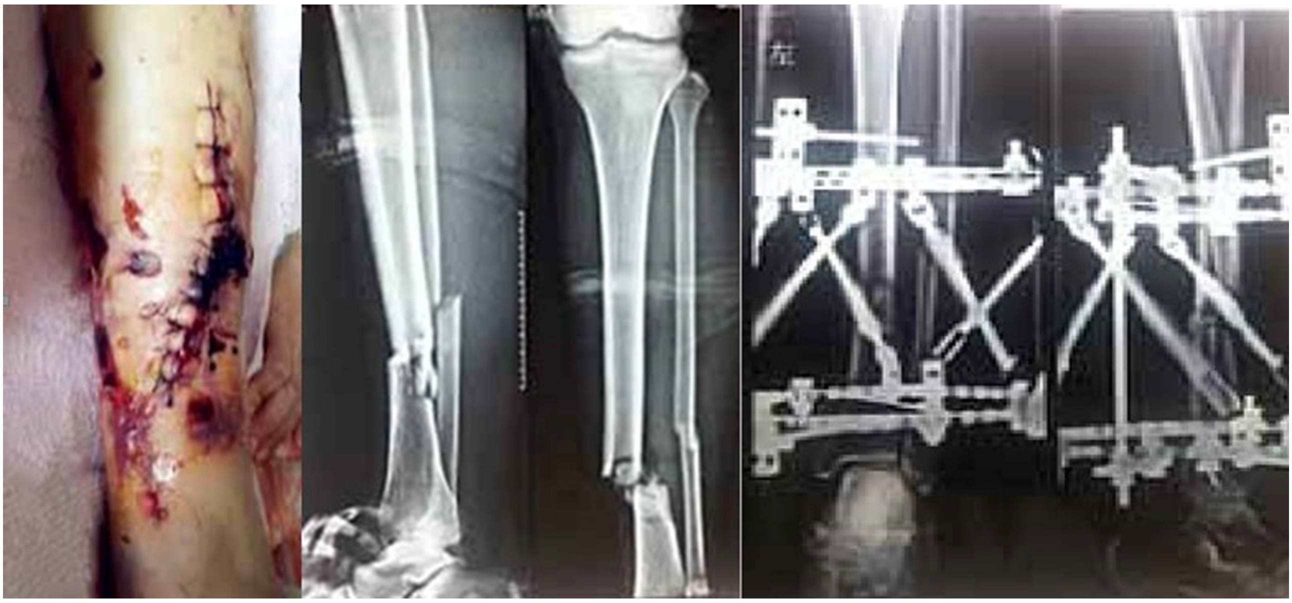
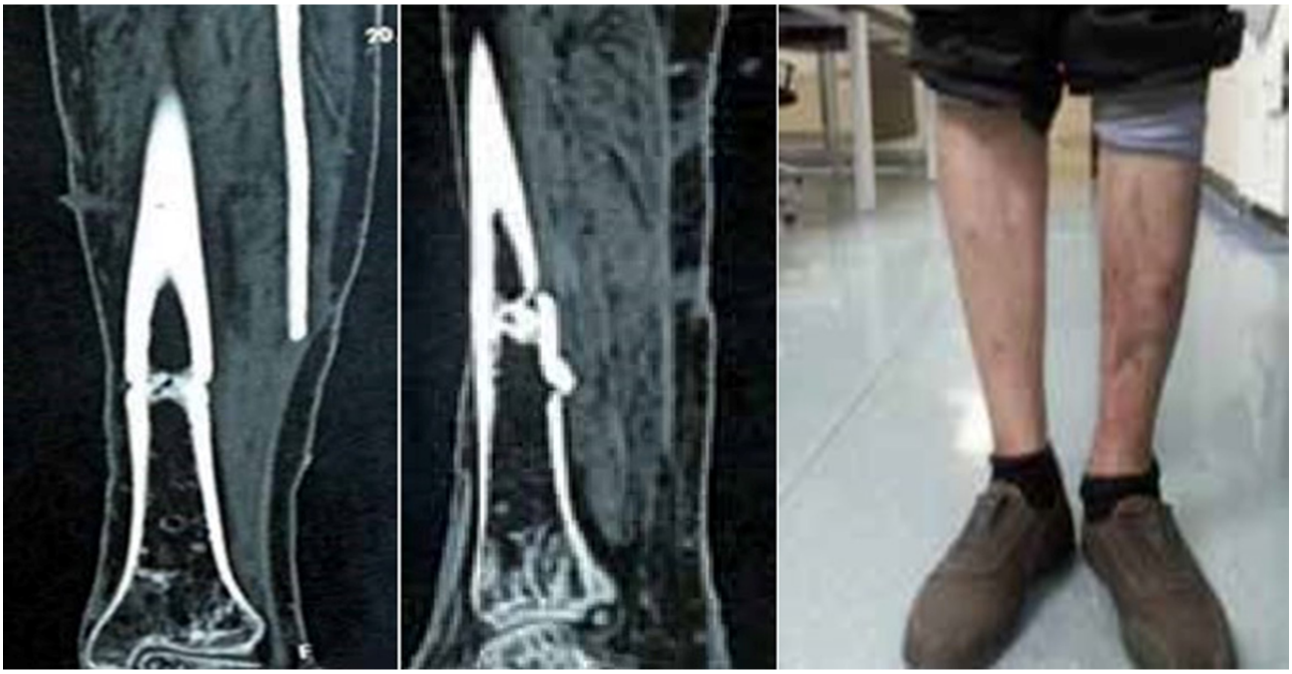


Figure 2. Postoperative recovery of patient A.



Case-Related Information

Clinicians at an external hospital completed the wound debridement and closure. Later, in our hospital, the soft tissue injury was found to be considerable. The preoperative x-ray showed a fracture. The third day of admission, TSF external fixation was performed. Software-assisted adjustments were performed to achieve a good reduction of the fracture end. Finally, 10 months later, a review of computed tomography scans showed good fracture healing; the affected limb had normal function 11 months after the removal of the outer frame.

Figure 3 showed the comminuted fracture of the left tibia of patient B with severe soft tissue injury. Figures 4 and 5 showed the patient's TSF treatment and postoperative recovery.

Figure 3. Patient B had a car accident–caused comminuted fracture of the left tibia, combined with severe soft tissue injury.

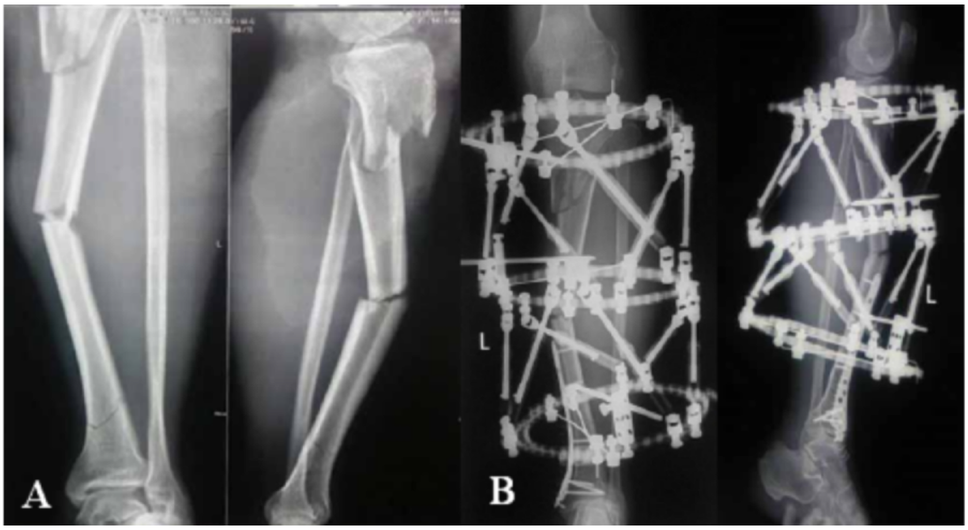


Figure 4. Patient with TSF treatment. TSF: Taylor 3D external fixation.

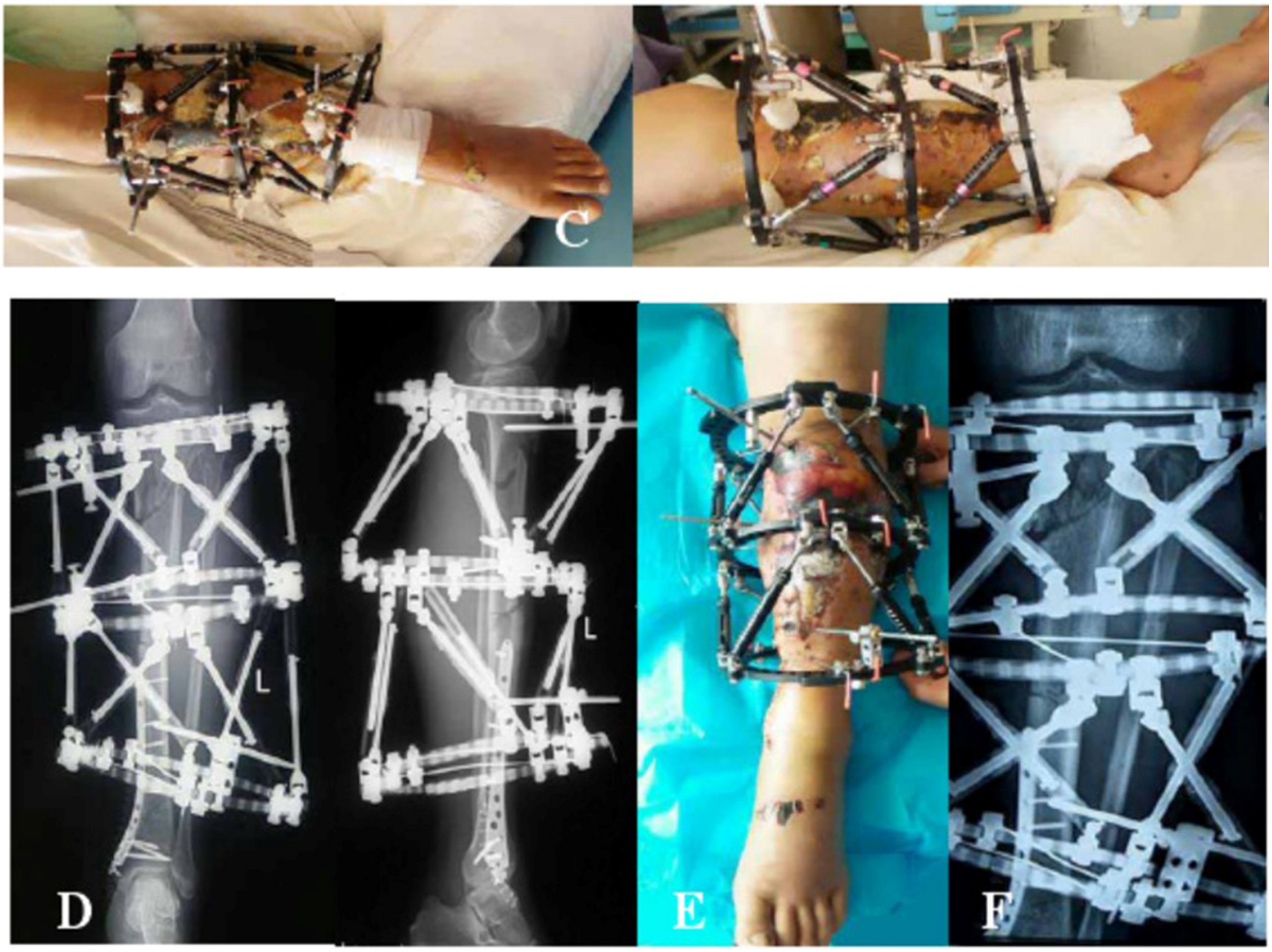
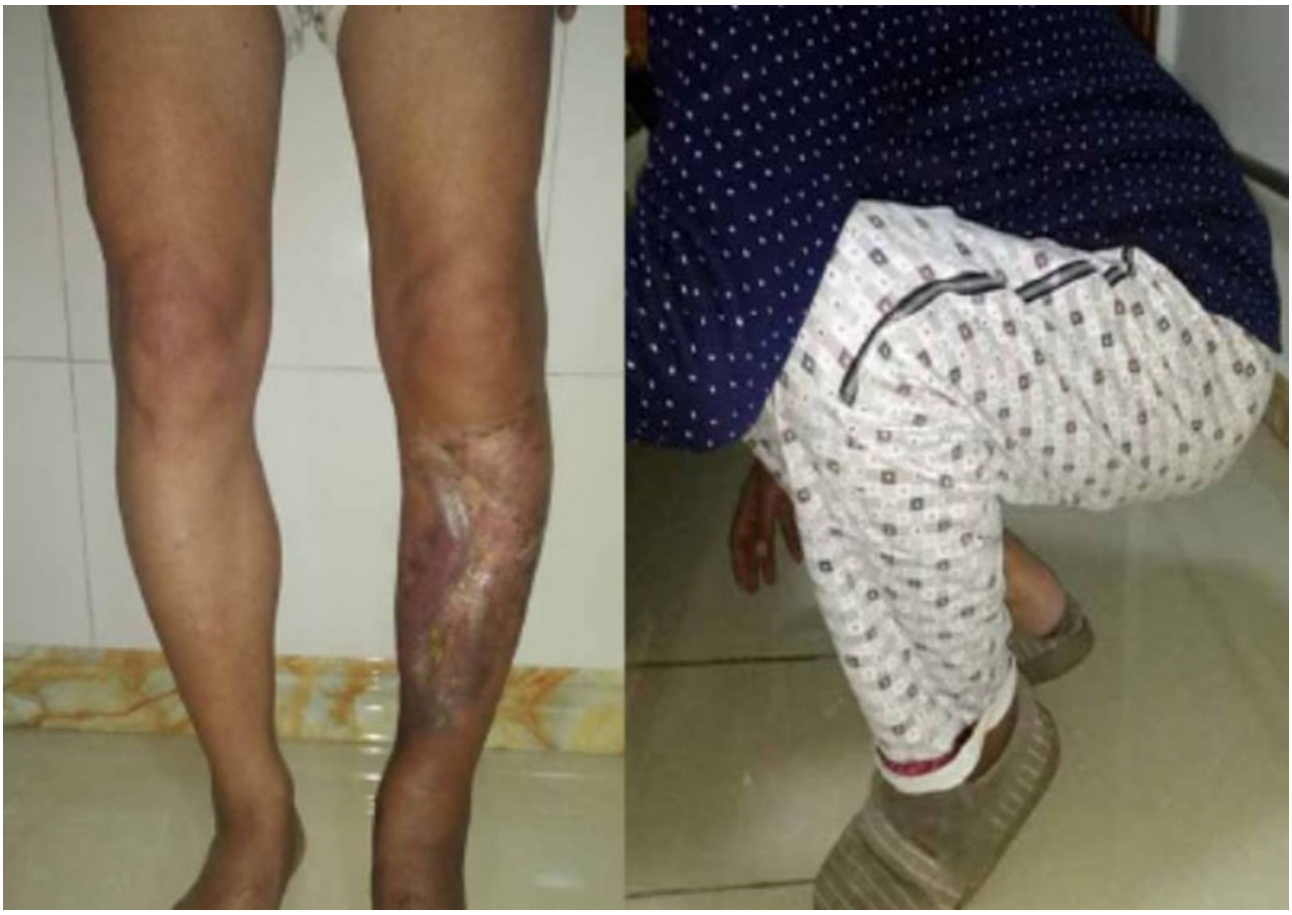


Figure 5. Postoperative recovery of patients.



Patient-Related Information

The patient was 48 years old. A car accident caused injuries of the left tibia including severe soft tissue injury and three comminuted fractures (Tscherne grade 3). The lateral position of the left tibia showed three fractures of the proximal, middle, and distal bone, and the proximal and middle fractures were clearly displaced. On the fourth day of admission, the proximal and middle fractures were treated with TSF external fixation, the distal fracture was fixed with internal fixation, and a lateral x-ray was performed. After the operation, soft tissue damage was severe and many tensional blood vessels could be seen; in addition, the anterior tibial skin was black and necrotic. Combined with the TSF computer software, the positive lateral radiographs after fracture reduction were good. Finally, 11 months after the operation, the skin was restored to a good condition with soft tissue treatment such as skin grafting. The limb functioned well 20 months after surgery, and knee joint function was good.

Discussion

Principal Findings

In trauma orthopedics, tibiofibular fractures are common, accounting for about 12% of total long bone fractures. The prognosis after fracture is affected by the energy level of the injury. When the damage energy is higher, the probability of an open fracture, the degree of fracture complications, and the degree of soft tissue injury increase accordingly, increasing the incidence of postoperative complications. High-energy damage is mainly seen in traffic accidents, falls from high places, and direct injuries by heavy objects. In contrast, low-energy injuries are more common in sports (about 80.1%) and regular falls. As countries develop, there is a corresponding increase in the incidence of traffic accidents, leading to an increase in high-energy fractures. Since there are fewer subcutaneous tissues on the anterior aspect of the lower leg, these fractures are prone to be open fractures, and account for 9.72% to 13.7% of open fractures.

An epidemiological survey of 523 cases of tibiofibular fractures showed that 400 cases (76.5%) involved closed fractures and 123 cases (23.5%) involved open fractures [4]. These were a result of traffic accident injuries (37.5%), falling (17.8%), sports (30.9%), and beatings or direct hits (4.5%). The majority of the blood supply of the tibia is provided by the nourishing artery. This artery enters the tibia from the upper one-third, and the trophoblast descends into the skeletal cortex. In a fracture, most of the arteries providing cortical nourishment are broken, resulting in insufficient blood supply to the distal one-third of the tibia, which slows down healing and is not conducive to the patient's recovery. There are many treatment methods for tibia fractures, each with its own advantages and disadvantages. The choice is mainly based on the way the injury occurred, the fracture type, other injuries, and the patient's condition. Gypsum or splint fixation is generally suitable for stable fractures from a low-energy injury and those without obvious displacement. Due to the risk of calf compartment syndrome and venous thrombosis, it is currently used for fractures of the tibia. If there is a lower risk of calf compartment syndrome and venous thrombosis, there is less indication that the fracture was caused by a high energy injury.

Open reduction and internal fixation treatment can achieve good fracture reduction, which is beneficial for early functional training of the limb. The intramedullary nail is currently the preferred treatment for humerus shaft fractures. This technique has many advantages and a long history in the treatment of tibial fractures. It has a central fixed biomechanical advantage, involves a minimally invasive operation away from the fracture end, retains the hematoma at the fracture end, and involves less soft tissue exfoliation, which is conducive to fracture healing. Therefore, it is widely used in clinical practice. However, any given treatment is not perfect, and the intramedullary nail still has its limitations. In a previous study, 32 cases of proximal humeral fractures were treated with intramedullary nails and the malunion rate was 19%, indicating the treatment was not satisfactory [5]. Kumar et al [6] compared the biomechanical characteristics of the treatment of tibiofibular fractures with steel plate, interlocking intramedullary nail, and external fixation. The results from the intramedullary nail treatment are better than those of the other two techniques, but this treatment is associated with malunion.

The orthopedic surgeon's philosophy of fracture treatment has gone from Association for the Study of Internal Fixation (AO)-led anatomical reduction to strong internal fixation to the promotion of biological fixation. The four principles of treatment of fractures as proposed by the AO concept are as follows: (1) anatomical reduction, (2) compression fixations at the fracture end, (3) protection of blood supply, and (4) early functional exercise. Early dynamic compression plate (DCP) treatment increased friction at the end of the fracture through the compression of the fracture end, and achieved first-stage healing of the fracture. The DCP is in close contact with the bone, and the fracture is stabilized by increasing friction, which destroys the blood supply at the fracture end. Influenced by the AO concept, many orthopedic surgeons remove large amounts of soft tissue to destroy blood supply and achieve anatomical reduction. Strong fixation causes stress shielding, and the bone is prone to refracture after the removal of internal fixation. The concept of biological fixation puts more emphasis on the protection of local soft tissue and the blood supply of the fracture.

The biological fixation principle is as follows. First, fracture reduction is performed as far as possible from the fracture end, to protect local soft tissue. To minimize soft tissue dissection, comminuted fracture block reduction cannot excessively destroy the blood supply. Fracture fixation involves a low elastic modulus and good biocompatibility. The contact area between the built-in material and the bone surface is minimized, avoiding excessive fixation and causing stress to discourage refracture. The patient is preoperatively fully evaluated and there is a preoperative design process, which shortens operation time and reduces surgical exposure.

Through the transformation of the fracture fixation concept and related biomechanical research, the locking compression plate (LCP) came into being. It combines two completely different fixation techniques, both a compression plate as well as steel plates and nail tails. The locking component between them is used as an inner bracket. The LCP provides both angular and axial stability to prevent the screws from slipping. In one study, 28 patients with fractures of the lower tibia were treated with DCP and 20 patients were treated with LCP [7]. The fracture healing time was 16.2 months for DCP and 15.4 months for LCP. The LCP effect was better than the DCP result.

In another study, 25 cases of tibiofibular fracture were treated with the minimally invasive percutaneous plate osteosynthesis (MIPPO) technique, which is considered safe and effective [8]. The most ideal fracture treatment should be as minimally invasive as possible and avoid the use of implants such as steel plates and intramedullary nails. The internal fixation treatment is beneficial for the anatomical reduction of the fracture, but at the same time, this invasive operation increases the risk of infection. Due to the development of internal fixation equipment and the improvement of fracture fixation, the results of internal fixation for the treatment of severe tibial fractures have been greatly improved. However, due to the high incidence of complications such as postoperative infection and osteomyelitis, the combination of open and severe soft tissue injury, and the treatment of multiple comminuted fractures and infected tibiofibular fractures, this technique is challenging.

External fixation is a good solution to the abovementioned shortcomings of internal fixation in the treatment of severe tibial fractures. The external fixator is simple to install and the technique is easy to learn. It causes minor secondary damage to soft tissue and can be used for early fixation of open tibiofibular fractures or fractures with severe soft tissue injury. It is beneficial for the early care of the affected limb, such as functional exercise, adjacent joint function, and exercise. This study compared the postoperative complications of TSF external fixation and internal fixation in the treatment of severe complex tibiofibular fractures (Table 2). The incidence of postoperative wound infection and osteomyelitis was significantly lower in TSF than in the internal fixation group (P<.05), and there were no significant differences in the rates of delayed fracture healing, nonunion, and refracture. A study by Herrera-Pérez et al [9] included a total of 14 internal fixation and external fixation cases for the treatment of severe tibial fractures. A meta-analysis showed that the difference in the rate of refracture following either external or internal fixation was not statistically significant. The 41 patient cases in this study showed that external fixation was better than internal fixation in the treatment of open tibial fractures.

The commonly used single-sided and half-needle external fixator is quick and easy to operate, and is often used for postdebridement fracture fixation of open fractures. However, its fracture stability is not ideal, so it is also often used for temporary fixation, and other fixation methods are used later on. Sabesan et al [10] reported that patients over 12 years of age with humeral shaft fractures were treated with a unilateral external fixator due to the risk of lost fracture reduction.

In this study, 28 patients with severe tibiofibular fractures were treated with TSF and achieved good results (a fracture healing rate of 85.71%). TSF can achieve early and accurate reduction of fractures aided by a computer. TSF can be the final method of fracture fixation due to good fracture stability. It enables the precise treatment of tibial fractures caused by complex high-energy injuries. This study compared the preoperative preparation time, operation time, fracture healing time, total time to weight-bearing, hospitalization days, and other outcomes of TSF and internal fixation in the treatment of tibial fractures. The operation and healing times of the TSF group were shorter than in the internal fixation group [11].

The authors' experience in the treatment of complex tibial fractures with TSF includes the following insights: postoperative computer-assisted adjustment of external frame fractures requires the addition of two parallel links on both sides of the outer ring to increase stability; it is important to measure the fracture displacement parameters; the distance between the proximal and distal rings in the TSF installation process needs to be prejudged to avoid the longest distance between the two rings being greater than the longest model connecting rod, or shorter than the minimum model connecting rod, as the short length makes postoperative resetting impossible. Overall, our experience shows that TSF is effective in treating patients with severe tibial fractures caused by high-energy injuries. Compared with the internal fixation method, the incidence of postoperative wound infection and osteomyelitis was reduced. TSF can enable early fracture fixation surgery and early functional exercise, shorten hospitalization time, and reduce treatment costs.

Conclusions

TSF has a low complication rate, with the advantages of fracture closure and accurate reduction, providing a new treatment method for complex tibiofibular fractures. Compared with the internal fixation method, it has a shorter preoperative preparation time, operation time, fracture healing time, and total time to weight-bearing, as well as shorter hospital stays and lower hospitalization costs for the treatment of severe complex tibial fractures. There is a lower chance of complications such as postoperative infection and osteomyelitis, and there is no significant difference in the incidence of nonunion, delayed healing, and refracture.

Conflicts of Interest

None declared.

References

1. Laux CJ, Grubhofer F, Werner CML, Simmen H, Osterhoff G. Current concepts in locking plate fixation of proximal humerus fractures. J Orthop Surg Res 2017 Sep 25;12(1):137 [FREE Full text] [doi: 10.1186/s13018-017-0639-3] [Medline: 28946902]
2. Craig E. Reverse Shoulder Arthroplasty in the Treatment of Proximal Humeral Fractures. Techniques in Shoulder & Elbow Surgery 2015;16(4):99-102. [doi: 10.1097/bte.0000000000000062]
3. Ribeiro FR, Takesian FH, Bezerra LEP, Filho RB, Júnior ACT, da Costa MP. Impacted valgus fractures of the proximal humerus. Revista Brasileira de Ortopedia (English Edition) 2016 Mar;51(2):127-131. [doi: 10.1016/j.rboe.2016.01.004]
4. Oura K, Kunihiro O, Okada K, Tanaka H, Murase T. Corrective osteotomy assisted by computer simulation for a malunited intra-articular fracture of the distal humerus: two case reports. Arch Orthop Trauma Surg 2016 Aug 17;136(11):1499-1505. [doi: 10.1007/s00402-016-2555-0]
5. Lowry KJ, Hamson KR, Bear L, Peng YB, Calaluce R, Evans ML, et al. Polycaprolactone/glass bioabsorbable implant in a rabbit humerus fracture model. J Biomed Mater Res 1997 Sep 15;36(4):536-541. [doi: 10.1002/(sici)1097-4636(19970915)36:4<536::aid-jbm12>3.0.co;2-8]
6. Kumar S, Singh S, Kumar D, Kumar N, Verma R. Intercondylar humerus fracture- parallel plating and its results. J Clin Diagn Res 2015 Jan;9(1):RC01-RC04 [FREE Full text] [doi: 10.7860/JCDR/2014/12137.5479] [Medline: 25738046]
7. Silverstein MP, Yirenkyi K, Haidukewych G, Koval KJ. Analysis of Failure with the Use of Locked Plates for Stabilization of Proximal Humerus Fractures. Bull Hosp Jt Dis 2015 Jul;73(3):185-189 [FREE Full text] [Medline: 26535597]
8. Cruickshank D, Lefaivre KA, Johal H, MacIntyre NJ, Sprague SA, Scott T, et al. A scoping review of biomechanical testing for proximal humerus fracture implants. BMC Musculoskelet Disord 2015 Jul 30;16(1). [doi: 10.1186/s12891-015-0627-x]
9. Herrera-Pérez M, Boluda-Mengod J, Muñoz-Ortus R, Gutiérrez-Morales MJ, Pais-Brito J. Continuous pain and swelling after humerus fracture in an 86-years-old woman. Acta Ortop Mex 2017;31(1):30-34 [FREE Full text] [Medline: 28741325]
10. Sabesan VJ, Lombardo D, Petersen-Fitts G, Weisman M, Ramthun K, Whaley J. National trends in proximal humerus fracture treatment patterns. Aging Clin Exp Res 2017 Jan 25;29(6):1277-1283. [doi: 10.1007/s40520-016-0695-2]
11. Manoli A, Capriccioso C, Konda S, Egol K. Total shoulder arthroplasty for proximal humerus fracture is associated with increased hospital charges despite a shorter length of stay. Orthopaedics & Traumatology: Surgery & Research 2016 Feb;102(1):19-24. [doi: 10.1016/j.otsr.2015.11.003]

Abbreviations

AO: Association for the Study of Internal Fixation
BO: biological fixation
DCP: dynamic compression plate
ICF: Ilizarov circular fixator

LCP: locking compression plate
MIPPO: minimally invasive percutaneous plate osteosynthesis
TSF: Taylor 3D external fixation
VSD: vacuum sealing drainage

Edited by Z Du; submitted 15.06.20; peer-reviewed by J Ding, Y Oannidis, D Thyros; comments to author 22.06.20; revised version received 26.06.20; accepted 27.06.20; published 14.05.21.

Please cite as:
Sheng H, Xu W, Xu B, Song H, Lu D, Ding W, Mildredl H
Application of Intelligent Computer-Assisted Taylor 3D External Fixation in the Treatment of Tibiofibular Fracture: Retrospective Case Study
JMIR Med Inform 2021;9(5):e21455
URL: https://medinform.jmir.org/2021/5/e21455
doi:10.2196/21455
PMID:33988516

Original Paper

Evaluation of the Privacy Risks of Personal Health Identifiers and Quasi-Identifiers in a Distributed Research Network: Development and Validation Study

SeHee Oh[1], BS; MinDong Sung[1], MD; Yumie Rhee[2], PhD; Namki Hong[2], MD; Yu Rang Park[1], PhD

[1]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea

[2]Department of Internal Medicine, Endocrine Research Institute, Yonsei University College of Medicine, Seoul, Republic of Korea

Corresponding Author:
Yu Rang Park, PhD
Department of Biomedical Systems Informatics
Yonsei University College of Medicine
50 Yonsei-ro
Seodaemun-gu
Seoul, 03722
Republic of Korea
Phone: 82 2 2228 2493
Email: yurangpark@yuhs.ac

Abstract

Background: Privacy should be protected in medical data that include patient information. A distributed research network (DRN) is one of the challenges in privacy protection and in the encouragement of multi-institutional clinical research. A DRN standardizes multi-institutional data into a common structure and terminology called a common data model (CDM), and it only shares analysis results. It is necessary to measure how a DRN protects patient information privacy even without sharing data in practice.

Objective: This study aimed to quantify the privacy risk of a DRN by comparing different deidentification levels focusing on personal health identifiers (PHIs) and quasi-identifiers (QIs).

Methods: We detected PHIs and QIs in an Observational Medical Outcomes Partnership (OMOP) CDM as threatening privacy, based on 18 Health Insurance Portability and Accountability Act of 1996 (HIPPA) identifiers and previous studies. To compare the privacy risk according to the different privacy policies, we generated limited and safe harbor data sets based on 16 PHIs and 12 QIs as threatening privacy from the Synthetic Public Use File 5 Percent (SynPUF5PCT) data set, which is a public data set of the OMOP CDM. With minimum cell size and equivalence class methods, we measured the privacy risk reduction with a trust differential gap obtained by comparing the two data sets. We also measured the gap in randomly sampled records from the two data sets to adjust the number of PHI or QI records.

Results: The gaps averaged 31.448% and 73.798% for PHIs and QIs, respectively, with a minimum cell size of one, which represents a unique record in a data set. Among PHIs, the national provider identifier had the highest gap of 71.236% (71.244% and 0.007% in the limited and safe harbor data sets, respectively). The maximum size of the equivalence class, which has the largest size of an indistinguishable set of records, averaged 771. In 1000 random samples of PHIs, Device_exposure_start_date had the highest gap of 33.730% (87.705% and 53.975% in the data sets). Among QIs, Death had the highest gap of 99.212% (99.997% and 0.784% in the data sets). In 1000, 10,000, and 100,000 random samples of QIs, Device_treatment had the highest gaps of 12.980% (99.980% and 87.000% in the data sets), 60.118% (99.831% and 39.713%), and 93.597% (98.805% and 5.207%), respectively, and in 1 million random samples, Death had the highest gap of 99.063% (99.998% and 0.934% in the data sets).

Conclusions: In this study, we verified and quantified the privacy risk of PHIs and QIs in the DRN. Although this study used limited PHIs and QIs for verification, the privacy limitations found in this study could be used as a quality measurement index for deidentification of multi-institutional collaboration research, thereby increasing DRN safety.

(JMIR Med Inform 2021;9(5):e24940) doi:10.2196/24940

KEYWORDS

distributed research network; Observational Medical Outcomes Partnership common data model; privacy risk quantification; personal health identifier; quasi-identifier

Introduction

As medical data include sensitive personal patient information, various challenges are being studied to protect patient information and optimize research results, including artificial intelligence, federated learning, and distributed research networks (DRNs) [1-11]. Among the above challenges, the DRN is a multi-institutional collaboration network [1] for standardizing the data of participating institutions into a common structure, terminology, and software called a common data model (CDM) [12-16]. In such research networks, data are not shared directly, and only analysis results are shared [1,3,6,17]. In research where sharing sensitive patient information has limitations or where large-scale data privacy needs to be preserved, the DRN structure is applied to standardize the data, terminology, and software [4-6]. There are several CDMs in DRNs, including the Observational Medical Outcomes Partnership (OMOP) CDM of Observational Health Data Sciences and Informatics (OHDSI), Sentinel CDM of the Food and Drug Administration, and Patient - Centered Outcomes Research Network of the Patient-Centered Outcomes Research Institute [18,19].

A DRN was recently recognized as a platform for protecting large-scale data [16,20-22]. DRN-based studies have argued two factors that enable the DRN infrastructure to mitigate privacy issues relative to other data sharing–based studies [1,6,23-29]. First, a DRN process protects patient information without directly sharing data [1,3,6,17]. Second, a CDM structure excludes some direct identifiers that could threaten the privacy of patient information, such as names and exact birthdays, by complying with the Health Insurance Portability and Accountability Act (HIPAA) [30-33]. Therefore, a DRN protects patient information through processes and structures.

However, previous studies have revealed limitations of DRNs in terms of data privacy. First, a DRN in a single site has privacy issues similar to a conventional database owing to repeated reuse [34-41]. Second, DRN privacy may be threatened when the remaining age and local information are used, even if direct identifiers are removed [34-43]. DRN researchers have recognized that there are no satisfactory solutions to privacy risk [43]. Despite such privacy risks, few studies have objectively measured these risks as compared to conventional data sharing–based studies [44-46]. To mitigate the possible risk to a DRN, an objective measurement of the privacy risk should be performed.

Thus, this study aimed to quantify DRN privacy risk by comparing different deidentification levels focusing on personal health identifiers (PHIs) and quasi-identifiers (QIs) of patient information. The key research questions in this study are as follows: (1) What PHIs and QIs are included in a DRN, and how many exist? (2) Using a PHI and QI, when comparing the deidentification level of a CDM to a safe harbor policy, how much will be the decrease in the DRN privacy risk? and (3) What is the true privacy risk of the PHI or QI itself when adjusted for the number of records?

Methods

Data Sources

We used the Synthetic Public Use File 5 Percent (SynPUF5PCT) data set, which is a sample data set of the OMOP CDM. The OMOP CDM (version 5.2.2), which was developed by OHDSI [18,47], is a database of relational schema and consists of 37 tables with demographic information, disease natural history, health care cost, etc [48]. The SynPUF5PCT is a synthetic data set with 5% random sampling from a synthetic public use file of the Centers for Medicare and Medicaid Services [49] and complies with the limited data set policy of the HIPAA [32]. The SynPUF5PCT consists of 33 of 37 OMOP CDM tables and is provided from the OHDSI [50]. We used only 12 tables with patient information without missing and null variables from the SynPUF5PCT [51].

Target PHIs and QIs

In this study, PHIs and QIs were focused on as privacy-threatening patient information by referencing previous studies [52-54]. For the PHIs, we manually matched the structure of the OMOP CDM based on 18 HIPAA identifiers (Figure 1) [55]. For the QIs, we selected the target range in demographic variables (eg, year of birth and gender) and clinical variables (eg, clinical order code) based on previous studies on the privacy risk of QIs [52-54,56,57]. In the 18 HIPAA identifiers, however, dates (excluding the year) and zip codes are defined as PHIs with a QI characteristic [56]. We prioritized the 18 HIPAA identifiers and fixed the dates and zip codes as PHIs instead of QIs. Forty-five PHIs and 17 QIs were detected from the OMOP CDM structure (Multimedia Appendix 1) [58]. Because there were missing tables in the SynPUF5PCT compared to the OMOP CDM, 16 PHIs and 12 QIs were targeted from the SynPUF5PCT (Figure 1 and Table 1). Detailed information for the 28 targeted variables is presented in Multimedia Appendix 2.

Figure 1. Study workflow. PHI: personal health identifier; QI: quasi-identifier.

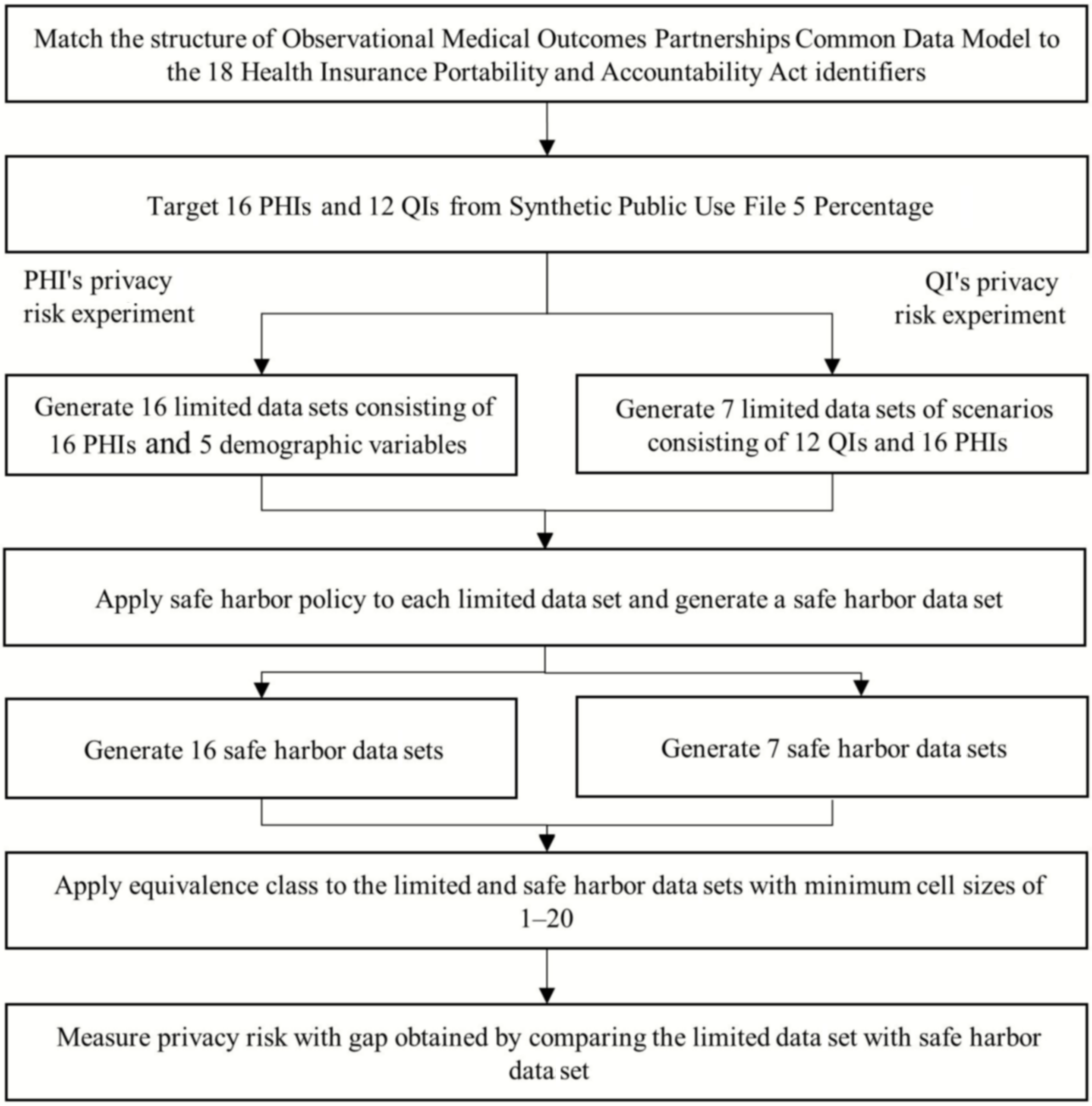

Match the structure of Observational Medical Outcomes Partnerships Common Data Model to the 18 Health Insurance Portability and Accountability Act identifiers
Target 16 PHIs and 12 QIs from Synthetic Public Use File 5 Percentage
PHI's privacy risk experiment
QI's privacy risk experiment
Generate 16 limited data sets consisting of 16 PHIs and 5 demographic variables
Generate 7 limited data sets of scenarios consisting of 12 QIs and 16 PHIs
Apply safe harbor policy to each limited data set and generate a safe harbor data set
Generate 16 safe harbor data sets
Generate 7 safe harbor data sets
Apply equivalence class to the limited and safe harbor data sets with minimum cell sizes of 1–20
Measure privacy risk with gap obtained by comparing the limited data set with safe harbor data set

Table 1. Sixteen personal health identifiers and 12 quasi-identifiers targeted in the Observational Medical Outcomes Partnership common data model based on not null values in the Synthetic Public Use File 5 Percent data set.

Standard clinical tables in the OMOP[a] CDM[b]	Variable of personal health identifier	Demographic variable of quasi-identifier	Clinical variable of quasi-identifier
Person	Month_of_birth and Day_of_birth	Year_of_birth, Gender_concept_id, Race_concept_id, and Ethnicity_concept_id	N/A[c]
Death	Death_date	N/A	N/A
Device_exposure	Device_exposure_start_date and Device_exposure_end_date	N/A	Device_concept_id
Drug_exposure	Drug_exposure_start_date and Drug_exposure_end_date	N/A	Drug_concept_id
Location	County	State	N/A
Measurement	Measurement_date	N/A	Measurement_concept_id
Observation	Observation_date	N/A	Observation_concept_id
Procedure_occurrence	Procedure_date	N/A	Procedure_concept_id
Visit_occurrence	Visit_start_date and Visit_end_date	N/A	N/A
Condition_occurrence	Condition_start_date and Condition_end_date	N/A	Condition_concept_id
Provider	NPI[d]	N/A	N/A
Care_site	N/A	N/A	Place_of_service_concept_id

[a] OMOP: Observational Medical Outcomes Partnership.

[b] CDM: common data model.

[c] N/A: not applicable.

[d] NPI: national provider identifier.

Study Design

We conducted privacy risk experiments of the PHIs and QIs. We generated data sets for each experiment. The workflow for this study is shown in Figure 1. In the privacy risk experiment of the PHIs, 16 limited data sets were generated, with each comprising one of the 16 PHIs merged with five common demographic variables (Year_of_birth, Gender_concept_id, Race_concept_id, Ethnicity_concept_id, and State), as in previous clinical studies [53,54]. For example, Condition_start_date, which is the name of data set 1 of the 16 limited data sets, consists of one PHI (Condition_start_date variable) and five common demographic variables. Another example is the Procedure_date data set consisting of one PHI (Procedure_date variable) and five common demographic variables. Thus, each limited data set consists of six variables.

In the QI privacy risk experiment, we mocked up seven scenarios based on the core tables of the OMOP CDM [16,59-61], which are frequently used in the real world. The seven scenarios are as follows: (1) diagnosis, (2) procedure, (3) drug treatment, (4) lab test, (5) device treatment, (6) death, and (7) medical history (Multimedia Appendix 3). Based on the scenarios, seven limited data sets were generated: 10 PHIs and seven QIs were assigned according to the characteristics of each scenario differently, and five demographic variables and six PHIs were used as common variables (Multimedia Appendix 3). For example, the diagnosis scenario consisted of 14 variables as follows: two PHIs (Condition_start_date and Condition_end_date) and one QI (Condition_concept_id), which followed the characteristics of the diagnosis scenario, and 11 common variables were merged.

To compare different deidentification levels for the same data set, we applied the safe harbor policy to the 16 limited data sets. For example, when the safe harbor policy was applied to the limited data set, the PHIs were partially or completely masked. The date type (such as start date, end date, and death date) was masked from “YYYY-MM-DD” to “YYYY-**-**.” In other words, they used only the “year”. The others (such as Month_of_birth, Day_of_birth, NPI, and County) were completely masked. We additionally generated 16 and seven safe harbor data sets for PHIs and QIs, respectively, by applying the safe harbor policy on the limited data sets.

Privacy Risk Evaluation Metrics

An equivalence class (EC) denotes a group of indistinguishable record forms with common attributes. The common attribute sizes that are included in each group can be represented as the calculated size of the EC [46]. An EC size of one represents the highest possibility of privacy disclosure for a certain patient’s information [56]. In contrast, if the size is maximum, it indicates the highest deidentification level of the data set. In previous studies, the minimum cell size was an empirically defined threshold with the calculated EC size [56,57]. The minimum cell size determines the level of deidentification and measures the privacy risk in the data set. The most commonly used minimum cell size in practice is five, and a larger size, such as

20, is used for data sets that include highly sensitive patient information [56]. The minimum cell size, calculated by the EC, was compared for both the limited and safe harbor data sets.

The trust differential mechanism represents the privacy risk of a data set with a gap obtained by comparing two different deidentification levels [54]. The gap represents the following two factors: (1) the quantified difference of the deidentification level and (2) the degree of decrease in privacy risk. In other words, when a certain privacy policy applies to the data set that complies with another privacy policy, a gap will occur between the two different privacy policies, which have different deidentification levels. Therefore, the gap indicates that the data set's privacy level with the lower deidentification privacy policy could be protected as the difference that arises when the higher privacy policy is applied.

Through the PHI and QI privacy risk experiments, we measured privacy risk in terms of the following two aspects: (1) measurements based on the number of total records in each data set and (2) measurements based on the identical number of records through random sampling from each data set. In the first aspect, we considered that clinical studies perform analysis with clinical tables according to clinical scenarios [16,59-61]; thus, we measured privacy risk with the number of total records in the data set generated by referring to previous studies [53,54]. With the number of total records, we compared the limited and safe harbor data sets based on the total records of each PHI and QI. Then, we measured with different minimum cell sizes from each PHI and QI experiment. To measure PHI privacy risk, we compared the limited and safe harbor data sets with the maximum EC size and a minimum cell size of one. In the QI privacy risk experiment, we compared the limited and safe harbor data sets with a minimum cell size of 1 to 20. In the second aspect, we extracted 1000, 10,000, 100,000, and 1 million random samples from each limited and safe harbor data set and iterated them 100 times. With the iterated random samples, we calculated the average of the minimum cell size 1 and then compared the limited and safe harbor data sets for PHIs and QIs.

Results

Overview

Overall, when compared with the limited and safe harbor data sets, privacy risk was reduced in both PHIs and QIs according to the trust differential gap. For the trust differential gap of a minimum cell size of one, there are two overall results. In the number of total records, the trust differential gaps of PHIs and QIs averaged 31.448% and 73.798%, respectively. In the random samples, the trust differential gaps of PHIs and QIs averaged 18.869% and 6.493% (1000 samples), 50.730% and 33.248% (10,000 samples), 74.013% and 60.306% (100,000 samples), and 50.744% and 71.868% (1,000,000 samples), respectively (Table 2).

Table 2. The averaged trust differential gap according to total records and random samples.

Number of total records[a] and sample[b]	Trust differential gap[c] with a minimum cell size of one[d]	
	Personal health identifier (mean percentage)	Quasi-identifier (mean percentage)
Number of total records	31.448%	73.798%
Sample		
1000	18.869%	6.493%
10,000	50.730%	33.248%
100,000	74.013%	60.306%
1,000,000	50.744%	71.868%

[a] Number of total records is each personal health identifier's total record.

[b] Sample is the number of random samples (ie, 1000, 10,000, 100,000, or 1 million) from the limited and safe harbor data sets.

[c] Trust differential gap is the difference obtained by comparing two data sets to measure privacy risk.

[d] Minimum cell size of one is the percentage of unique records. This can be expressed with the number of unique records as the numerator and the number of total records as the denominator.

Evaluation of the Personal Health Identifier Privacy Risk of the DRN

In the number of total record results of the limited data set, the variable with the most included minimum cell size of one was Death_date, which was 98.787% (1141/1155). In addition, the maximum EC size of two for Death_date means that every record consists of only two value types. In Death_date of the safe harbor data set, the minimum cell size of one was 87.359% (1009/1155), and the maximum EC size was three. Even though the safe harbor policy was applied, privacy was still threatened. In the Death_date trust differential gap, the gap with a minimum cell size of one was 11.428%, and the maximum EC size was one. The maximum EC size of one is the lowest trust differential gap among all the maximum EC size gaps. In the limited data set, the variable with the least minimum cell size of one was Condition_end_date, which was 4.540% (146,727/3,231,730). In Condition_end_date from the safe harbor data set, the minimum cell size of one was 0.003% (125/3,231,730). Even though the safe harbor policy was applied, the records of a minimum cell size of one did not significantly decrease. In the Condition_end_date trust differential gap, the minimum cell size of one was 4.536%, and the maximum EC size was 2348. This maximum EC size of 2348 was the highest trust differential gap among all the maximum EC size gaps. In the trust differential gaps with a minimum cell size of one, the NPI variable had the highest trust differential gap of 71.236%, which was the difference between the limited (71.244%) and safe harbor (0.007%) data sets. For Drug_exposure_start_date and Drug_exposure_end_date, both data sets exhibited the same maximum EC size and a minimum cell size of one.

Day_of_birth consists of the day part of the date of birth and was already deidentified as “1” in the SynPUF5PCT data set (eg, “dd” to “1”); thus, every patient had the exact same Day_of_birth value. Because it was the same deidentified method as for the safe harbor policy, the Day_of_birth trust differential gap was zero (Table 3). It could be provided as a statistical baseline for five demographic variables without any PHI variables. When the measured result of the Day_of_birth variable (13.079%) was compared with that of the Condition_end_date variable, the result of the Condition_end_date variable was lower by 8.539 percentage points (from 13.079% to 4.540%), and when it was compared with that of the Death_date variable, the result of the Death_date variable was higher by 85.708 percentage points (from 13.079% to 98.787%).

Table 3. Comparison of 16 personal health identifier variables and five demographic variables of the SynPUF5PCT with limited and safe harbor data sets in terms of a minimum cell size of one and the maximum size of the equivalence class.

Variable[a]	Number of total records[b]	Limited data set			Safe harbor data set			Trust differential gap[c]	
		Number of unique records[d]	Minimum cell size of one[e] (%)	Maximum size of the equivalence class[f]	Number of unique records[d]	Minimum cell size of one[e] (%)	Maximum size of the equivalence class[f]	Minimum cell size of one (%)	Maximum size of the equivalence class
Visit_start_date	1,218,881	771,684	63.310	10	581	0.047	888	62.952	878
Visit_end_date	1,218,881	771,891	63.327	10	581	0.0395	889	62.960	879
Death_date	1155	1141	98.787	2	1009	87.359	3	11.428	1
Condition_start_date	3,231,730	146,828	4.543	45	137	0.004	2391	4.538	2346
Condition_end_date	3,231,730	146,727	4.540	45	125	0.003	2393	4.536	2348
Procedure_date	3,024,452	257,161	8.502	64	201	0.006	2180	8.495	2116
Measurement_date	741,161	168,180	22.691	43	595	0.080	575	22.610	532
Observation_date	420,986	182,497	43.349	28	983	0.233	335	43.115	307
Device_exposure_start_date	47,655	13,232	27.766	40	3190	6.693	218	21.073	178
Device_exposure_end_date	47,655	13,219	27.739	40	3191	6.696	187	21.043	147
Drug_exposure_start_date	158,316	55,042	34.767	45	2845	1.797	409	32.970	364
Drug_exposure_end_date	158,316	55,042	34.767	45	2845	1.797	409	32.970	364
Month_of_birth	25,200	14,508	57.571	8	3296	13.079	49	44.492	41
Day_of_birth	25,200	3296	13.079	49	3296	13.079	49	0	0
NPI[g]	1,215,317	865,840	71.244	70	91	0.007	2247	71.236	2177
County	25,200	18,103	71.837	12	3296	13.079	49	58.757	37
Average	N/A[h]	N/A	40.488	34.75	N/A	8.999	829.437	31.448	771.937

[a] Variable refers to the variable targeted from the Observational Medical Outcomes Partnership common data model as the personal health identifier.

[b] Number of total records is each personal health identifier’s total record.

[c] Trust differential gap is the difference obtained by comparing two data sets to measure privacy risk.

[d] Number of unique records is the number of records with a common attribute size of one within the total record.

[e] Minimum cell size of one is the percentage of unique records. This can be expressed with the number of unique records as the numerator and the number of total records as the denominator.

[f] Maximum size of the equivalence class is the largest size of the indistinguishable common attributes.

[g] NPI: national provider identifier.

[h] N/A: not applicable.

In randomly sampled PHIs, privacy risk reduction was different depending on the number of samples (Table 4 and Multimedia Appendix 4). The variables with a highly ranked trust differential gap were Device_exposure_start_date (1000 samples) (33.730%; 87.705% and 53.975% in the limited and safe harbor data sets, respectively), NPI (10,000 samples)

(83.852%; 98.945% and 15.094% in the limited and safe harbor data sets, respectively), Visit_start_date (100,000 samples) (92.566%; 95.583% and 3.016% in the limited and safe harbor data sets, respectively), and NPI (1,000,000 samples) (73.588%; 73.599% and 0.011% in the limited and safe harbor data sets, respectively).

Overall, for 1000 random samples, both data sets consisted primarily of the minimum cell size of one. In the limited data set, the variables with the most and fewest included minimum cell size of one records were Visit_end_date (99.978%) and Day_of_birth (73.754%), respectively. In the safe harbor data set, the variables with the most and fewest included minimum cell size of one records were Death_date (89.044%) and NPI (67.377%), respectively (Table 4). For Visit_end_date in the limited data set with the most included minimum cell size of one records, after applying the safe harbor policy, the minimum cell size of one records of the Visit_end_date variable decreased to 86.171% (861.710/1000). Even though the safe harbor policy was applied, the minimum cell size of one records did not decrease significantly. Death_date, with the most included minimum cell size of one records in the safe harbor data set, had a trust differential gap of 9.862% (98.906% and 89.044% in the limited and safe harbor data sets, respectively). The privacy risk did not decrease significantly after applying the safe harbor policy. In the trust differential gap, the variable with the highest gap was Device_exposure_start_date (33.730%; 87.705% and 53.975% in the limited and safe harbor data sets, respectively). When the safe harbor policy was applied, the Death_date privacy risk could be significantly reduced. In the number of total records of the limited and safe harbor data sets, with a minimum cell size of one, the most privacy-threatening variables were Death_date (98.787%) and Death_date (87.359%), respectively. However, in the random sample of 1000, it was Visit_end_date (99.978%) and Death_date (89.044%), respectively. Therefore, we verified that privacy-threatening variables could differ depending on the number of records. Detailed random sampled results are displayed in Multimedia Appendix 4.

Table 4. Comparison of records with a minimum cell size of one between the limited and safe harbor data sets from 16 personal health identifier data sets.

Sample[a] and variable[b]	Limited data set		Safe harbor data set		Trust differential gap[c] (%)
	Number of minimum cell sizes of one[d]		Number of minimum cell sizes of one[d]		
	Mean[e] (SD[f])	Percentage[g] (%)	Mean[e] (SD[f])	Percentage[g] (%)	
1000 samples					
Visit_start_date	999.26 (1.125)	99.926	859.68 (16.229)	85.968	13.958
Visit_end_date	999.78 (0.629)	99.978	861.71 (15.086)	86.171	13.807
Death_date	989.06 (2.155)	98.906	890.44 (7.478)	89.044	9.862
Condition_start_date	998.44 (1.766)	99.844	857.59 (15.178)	85.759	14.085
Condition_end_date	998.12 (2.006)	99.812	858.7 (16.45)	85.87	13.942
Procedure_date	998.24 (1.804)	99.824	858.61 (16.504)	85.861	13.963
Measurement_date	995.14 (2.971)	99.514	855.11 (15.321)	85.511	14.003
Observation_date	997.54 (2.162)	99.754	854.79 (14.238)	85.479	14.275
Device_exposure_start_date	877.05 (13.107)	87.705	539.75 (16.877)	53.975	33.73
Device_exposure_end_date	875.34 (16.05)	87.534	539.68 (20.112)	53.968	33.566
Drug_exposure_start_date	956.34 (8.669)	95.634	720.7 (17.729)	72.07	23.564
Drug_exposure_end_date	956.34 (8.669)	95.634	720.7 (17.729)	72.07	23.564
Month_of_birth	971.47 (7.612)	97.147	738.4 (17.707)	73.84	23.307
Day_of_birth	737.54 (17.774)	73.754	737.54 (17.774)	73.754	0
NPI[h]	998.8 (1.775)	99.88	673.77 (19.212)	67.377	32.503
County	979.69 (6.59)	97.969	738.46 (16.856)	73.846	24.123
Average	N/A[i]	N/A	N/A	N/A	18.869

[a] Sample is the number of random samples (ie, 1000, 10,000, 100,000, or 1 million) from the limited and safe harbor data sets.

[b] Variable is the variable targeted from the Observational Medical Outcomes Partnership common data model as the personal health identifier.

[c] Trust differential gap is the difference obtained by comparing two data sets to measure privacy risk.

[d] Number of minimum cell sizes of one is the number of records with a unique record among the total records.

[e] Mean is the average of the quantity with a minimum cell size of one obtained by iterating the random sampling of each variable 100 times.

[f] SD is the standard deviation of the quantity with a minimum cell size of one obtained by iterating random sampling of each variable 100 times.

[g] Percentage is the percentage of the quantity with a minimum cell size of one. The numerator is the mean of the minimum cell size of one, which was obtained from 100 iterations, and the denominator was the number of random samples.

[h] NPI: national provider identifier.

[i] N/A: not applicable.

Evaluation of the Quasi-Identifier Privacy Risk of the DRN

In the results for the number of total records, the privacy risk of the QI with a minimum cell size of 1 to 20 was measured in the limited and safe harbor data sets. As shown in Figure 2, for the minimum cell size of one, the minimum and maximum percentages in the seven scenarios were 71% and 99%, respectively, in the limited data set (Figure 2A) and 0.7% and 41%, respectively, in the safe harbor data set (Figure 2B). The QI privacy risk was represented with a minimum cell size of one to five (Multimedia Appendix 5 and Table 5). For the minimum cell size of one in the limited data set, the Diagnosis (71.465%) and Procedure (76.123%) scenarios showed lower privacy risks than the other five scenarios (Drug treatment [95.475%], Lab test [93.012%], Medical history [92.353%], Death [99.997%], and Device treatment [97.647%]). For the Death scenario, the limited data set records were concentrated in the minimum cell size of one to two. The average gaps between the limited and safe harbor data sets, with the minimum cell size of one to five decreased from 73.798% to 54.548%. For the gaps of the minimum cell size of one, the Diagnosis scenario showed the smallest gap (28.869%), whereas the Death scenario showed the largest gap (99.212%).

Figure 2. Percentage of records measuring the quasi-identifier privacy risk with a minimum cell size of 1–20 for the (A) limited and (B) safe harbor data sets. The flattened lines are expanded (inner graph).

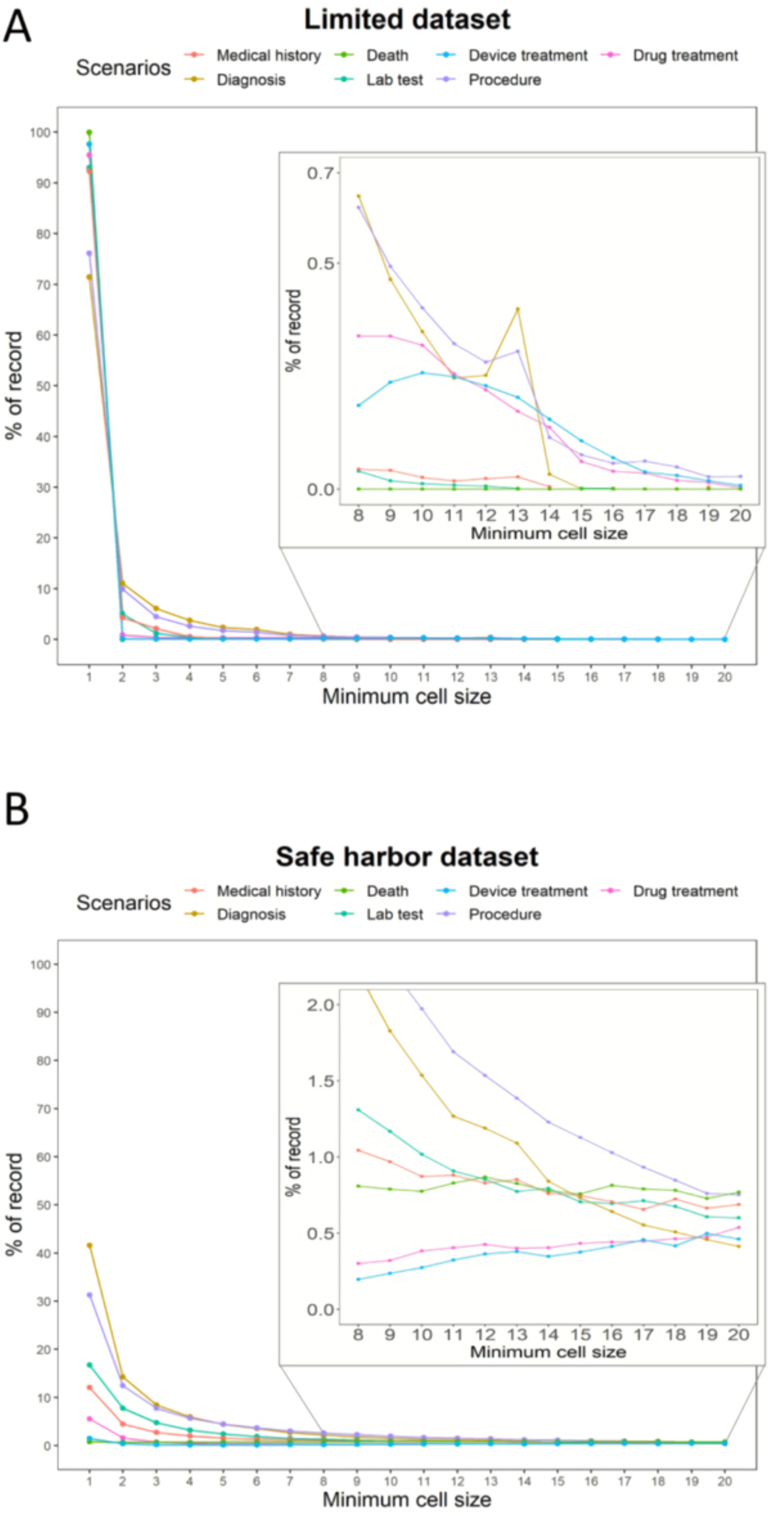


Table 5. Percentage of records measuring quasi-identifier privacy risk with gaps between the limited and safe harbor data sets with a minimum cell size of one, two, and five.

Scenarios	Number of total records[a]	Limited data set			Safe harbor data set			Trust differential gap[b]		
		Minimum cell size of one, two, and five, percentage[c] (record[d])			Minimum cell size of one, two, and five, percentage[c] (record[d])			Minimum cell size of one, two, and five, percentage[c]		
		1	2	5	1	2	5	1	2	5
Diagnosis	3,369,468	71.465 (2,407,996)	11.049 (186,162)	2.333 (15,726)	41.595 (1,401,556)	14.248 (240,043)	4.412 (29,737)	29.869	35.361	19.986
Procedure	3,105,665	76.123 (2,364,135)	9.902 (153,767)	1.731 (10,752)	31.251 (970,568)	12.472 (193,672)	4.460 (27,708)	44.871	42.301	33.173
Drug treatment	1,300,649	95.475 (1,241,796)	0.895 (5826)	0.306 (796)	5.558 (72,292)	1.625 (10,569)	0.356 (927)	89.917	89.187	88.611
Lab test	1,622,884	93.012 (1,509,486)	5.043 (40,923)	0.138 (448)	16.749 (271,819)	7.748 (62,875)	2.385 (7744)	76.263	73.558	64.958
Medical history	1,348,569	92.353 (1,245,455)	4.286 (28,900)	0.224 (606)	12.079 (162,898)	4.466 (30,115)	1.550 (4183)	80.274	80.094	76.759
Death	1,218,881	99.997 (1,218,845)	0.003 (18)	0 (0)	0.784 (9557)	0.686 (4185)	0.719 (1755)	99.212	98.529	0
Device treatment	1,247,726	97.647 (1,218,368)	0.152 (954)	0.059 (149)	1.464 (18,271)	0.380 (2372)	0.139 (348)	96.182	95.955	95.625
Average	N/A[e]	N/A	N/A	N/A	N/A	N/A	N/A	73.798	73.569	54.458

[a] Number of total records denotes each total record of the scenarios.

[b] Trust differential gap indicates the differences obtained by comparing two data sets to measure privacy risk.

[c] Minimum cell size of one, two, and five represents the percentage of records that have a common attribute size of one, two, and five, respectively. This percentage is presented as the records of minimum cell size of one, two, and five as the numerator and the total number of records as the denominator.

[d] Record is the number of records with a common attribute size of one, two, and five within the total records.

[e] N/A: not applicable.

In the random samples with a minimum cell size of one, (1) the average percentage of the limited data set decreased from 99.986% to 99.327%, (2) the average percentage of the safe harbor data set decreased from 93.493% to 21.460%, and (3) the average trust differential gap increased from 6.493% to 71.868% (Table 6). In the limited data set with 1000 to 1 million random samples, the scenario with the most included records of a minimum cell size of one was the Death scenario (1000 to 100,000 random samples had 99.999% and 1 million had 99.998%). In the safe harbor data set with 1000 to 1 million random samples, the scenario with the most included records of a minimum cell size of one was the Diagnosis scenario (1000 random samples had 99.858%, 10,000 had 98.685%, 100,000 had 89.758%, and 1 million had 60.361%). In the order of the four random samples, the scenarios with the highest trust differential gap were Device_treatment (1000 random samples: 12.980%, 99.980% and 87.000% in the limited and safe harbor data sets, respectively; 10,000 random samples: 60.118%, 99.831% and 39.713% in the limited and safe harbor data sets, respectively; 100,000 random samples: 93.598%, 98.805% and 5.207% in the limited and safe harbor data sets, respectively) and Death (1 million random samples: 99.063%, 99.998% and 0.934% in the limited and safe harbor data sets, respectively). When the safe harbor policy was applied, privacy risks were significantly reduced. In the number of total records, the most privacy-threatening scenarios were Death (99.997%) and Diagnosis (41.595%) in the limited and safe harbor data sets, respectively, with a minimum cell size of one. In the random samples with a minimum cell size of one in the limited data set, the most privacy-threatening scenario was Death, which had privacy risks of 99.999% (1000 to 100,000 random samples) and 99.998% (1 million random samples). In the safe harbor data set, Diagnosis had privacy risks of 99.858% (1000 random samples), 98.685% (10,000 random samples), 89.758% (100,000 random samples), and 60.361% (1 million random samples).

Table 6. Comparison of records with a minimum cell size of one between the limited and safe harbor data sets from seven scenarios.

Sample[a] and scenario[b]	Limited data set		Safe harbor data set		Trust differential gap[c] (%)
	Number of minimum cell sizes of one[d]		Number of minimum cell sizes of one[d]		
	Mean[e] (SD[f])	Percentage[g] (%)	Mean[e] (SD[f])	Percentage[g] (%)	
1000					
Diagnosis	999.800	99.980	998.580	99.858	0.122
Procedure	999.760	99.976	997.400	99.740	0.236
Drug_treatment	999.860	99.986	889.470	88.947	11.039
Lab_test	999.900	99.990	956.910	95.691	4.299
Medical_history	999.960	99.996	932.320	93.232	6.764
Death	999.990	99.999	899.830	89.983	10.016
Device_treatment	999.800	99.980	870.000	87.000	12.980
Average	N/A[h]	99.986	N/A	93.493	6.493
10,000					
Diagnosis	9975.850	99.759	9868.540	98.685	1.073
Procedure	9974.680	99.747	9743.830	97.438	2.309
Drug_treatment	9980.620	99.806	4642.820	46.428	53.378
Lab_test	9993.920	99.939	7320.070	73.201	26.739
Medical_history	9989.730	99.897	6226.700	62.267	37.630
Death	9999.980	99.9990	4851.230	48.512	51.487
Device_treatment	9983.140	99.831	3971.310	39.713	60.118
Average	N/A	99.854	N/A	66.606	33.248
100,000					
Diagnosis	97,724.930	97.725	89,757.540	89.758	7.967
Procedure	97,742.680	97.743	82,093.250	82.093	15.649
Drug_treatment	98,419.410	98.419	12,062.980	12.063	86.356
Lab_test	99,375.690	99.376	42,164.130	42.164	57.212
Medical_history	99,022.020	99.022	28,552.750	28.553	70.469
Death	99,999.640	99.999	9106.630	9.107	90.892
Device_treatment	98,804.960	98.805	5206.990	5.207	93.598
Average	N/A	98.727	N/A	38.420	60.306
1,000,000					
Diagnosis	846,819.090	84.682	603,607.950	60.361	24.321
Procedure	864,575.710	86.458	472,502.940	47.250	39.207
Drug_treatment	957,078.730	95.708	59,825.330	5.983	89.725
Lab_test	951,528.090	95.153	206,809.130	20.681	74.472
Medical_history	936,158.630	93.616	134,617.450	13.462	80.154
Death	999,975.900	99.998	9344.160	0.934	99.063
Device_treatment	976,802.140	97.680	15,496.020	1.550	96.131
Average	N/A	93.327	N/A	21.460	71.868

[a] Sample is the number of random samples (ie, 1000, 10,000, 100,000, or 1 million) from the limited and safe harbor data sets.

[b] Scenario is the variable targeted from the Observational Medical Outcomes Partnership common data model as the personal health identifier.

[c] Trust differential gap is the difference obtained by comparing two data sets to measure privacy risk.

[d] Number of minimum cell sizes of one is the number of records with a unique record among the total records.

[e] Mean is the average of the quantity with a minimum cell size of one obtained by iterating the random sampling of each variable 100 times.

[f] SD is the standard deviation of the quantity with a minimum cell size of one obtained by iterating random sampling of each variable 100 times.

[g] Percent is the percentage of the quantity with a minimum cell size of one. The numerator is the mean of the minimum cell size of one, which was obtained from 100 iterations, and the denominator was the number of random samples.

[h] N/A: not applicable.

Discussion

Principal Findings

In this study, we quantified the DRN privacy risk focusing on PHIs and QIs using 18 HIPAA identifiers and the findings of previous studies [34-43]. To measure the DRN privacy risk, we compared the limited data set, consisting of PHIs and QIs from the SynPUF5PCT data set, with the safe harbor data set generated by applying the safe harbor policy on the limited data set. More specifically, privacy risk was measured with the gap obtained between the two data sets, based on the trust differential, applying the threshold of the minimum cell size with the calculated size by the EC. We verified that the PHIs and QIs increased the DRN privacy risk. However, the privacy risk decreased overall when the safe harbor policy was applied to the DRN. To the best of our knowledge, this is the first study to verify that PHIs and QIs may threaten patient privacy within DRNs.

Prior studies have shown that patient privacy is threatened by PHIs and QIs within clinical databases [53,54]. The DRN of this study may have the same privacy risk as those in previous studies because the DRN at a single site follows a conventional database, although it does not share data [34-41]. Therefore, the privacy risk in a DRN should be quantified and objectively measured for three important reasons. First, because existing patient information in a CDM affects the privacy risk, the DRN privacy risk can be mitigated by providing objectively measured PHI and QI privacy risks [62]. Second, researchers can understand the mechanism of privacy risk change with the objective differences measured by comparing two different deidentification levels of data sets [63]. Finally, an objective measurement of privacy risk will contribute to the design of more secure privacy protection methods suitable for a DRN.

Consideration for Measuring Privacy Risk From Variable Characteristics

The PHI results, which measure the privacy risk, were verified in two different deidentification levels and indicated a much greater privacy risk reduction in the safe harbor data set than in the limited data set. In addition, we found that privacy risks differ depending on PHI characteristics. The privacy risk of the Visit_start_date variable, which occurs multiple times per patient, was significantly reduced after applying the safe harbor policy. However, the Death_date variable, which occurs only once per patient, still had many remaining unique records after the safe harbor policy was applied. The State variable, which is one of the demographic variables in the data set of the Death_date variable, still had unique values because it had not been deidentified by the safe harbor policy. Although the NPI variable had the highest reduction rate of privacy risk after applying the safe harbor policy, we found that it could not be used as data because it was completely masked. For the Day_of_birth as a statistical baseline, we compared the Day_of_birth with other PHI variables and could interpret a privacy risk according to the characteristics of the variable as follows. First, because each patient had multiple points for the Condition_end_date value in the SynPUF5PCT, there were fewer unique records relatively. Thus, the privacy risk of Condition_end_date was lower than that of Day_of_birth. Second, because every patient had only one point for the Death_date value, most of them had unique records. Thus, the privacy risk of Death_date was higher than that of Day_of_birth.

In the results of QI, when the limited data set had a minimum cell size of one, the privacy risk differed based on the characteristics of the scenario. In our study, we found that the QI privacy risks of the Drug treatment, Lab test, Medical history, Death, and Device treatment scenarios decreased on average 1.3 times more than those of the Diagnosis and Procedure scenarios, with a minimum cell size of one. The reason for the relatively low reduction in privacy risk under the Diagnosis and Procedure scenarios is that clinical order codes, such as Condition_concept_id and Procedure_concept_id, which used QIs, were prescribed three times on average with the same code.

The privacy risk could differ depending on the characteristics of variables, and the "balls and bins problem" theoretical basis supports our research [64]. As the number of bins increases, it could frequently take only one ball to fill than fewer bins. Similarly, the Visit_end_date variable, with 1096 distinct values ("bins"), consisted of more unique records ("only one ball") than the Month_of_birth with 12 distinct values. Consequently, a privacy protection approach must be customized or optimized by considering the characteristics of each variable.

Consideration for Measuring Privacy Risk From Record Extraction

Through the random samples, we found the following two facts: (1) Depending on the number of records, the privacy-threatening variable or scenario could differ and (2) The influence of safe harbor policy could differ depending on the number of records, because the number of unique records, which are included with PHI data sets or QI scenarios, differs according to each random sampling. Therefore, to measure the true privacy risk of PHIs and QIs, it is necessary to compare the same records through random sampling.

A minimum cell size of five, which has been a commonly used threshold in previous studies [56], may be difficult to apply as a threshold for measuring the DRN privacy risk. In the QI privacy risk experiment, the Death scenario of the limited data set was not appropriate for a minimum cell size of five because the records were concentrated in a minimum cell size of one to two. Therefore, our results reflect the fact that a minimum cell size of five may not be suitable for the current DRN. However, it should be recognized that the captured features may differ according to the data set used. Therefore, further research is required using various real-world data sets to find an appropriate minimum cell size that can contribute to the measurement of the DRN privacy risk.

Limitations

This study has some limitations. First, this study used a public data set (SynPUF5PCT), which does not handle all PHIs or QIs existing in a DRN. Therefore, we could not consider the CDM of real-world data sets generated by each institution. However, the results of this study are reliable because the SynPUF5PCT data set is an officially published data set by the OHDSI [50]. Second, when measuring the QI privacy risk, some QIs were considered based on scenarios and not based on all variables. Thus, we did not handle the privacy risk considering the combination of all QIs. However, the CDM does not use all variables because the research is based on clinical questions [59]. In addition, we focused on the frequently used scenarios. Third, we did not consider some PHIs and QIs within free text from Note and Note_nlp tables [48], because in our research methodology, PHIs and QIs are detected in the structure of OMOP CDM based on 18 HIPAA identifiers and not in the free text. However, previous studies have indicated that free text includes not only PHIs and QIs but also direct identifiers [65,66]. Therefore, further research needs to include a free text data set. Fourth, we did not consider privacy risk depending on the timespan. Because the SynPUF5PCT data set used in this study contained only 3-year records (2008-2010) and the Day_of_birth variable had already been deidentified as “1,” we could not measure privacy risk according to an extended (such as 20-year records) or a narrowed (such as single-week records) timespan. A future study should consider timespan-related privacy.

Conclusions

In this study, we validated and quantified the privacy risks of PHIs and QIs in the DRN. We objectively measured the privacy risk reduction with the gaps obtained by comparing a safe harbor policy with the DRN. In addition, we measured the true privacy risk of PHIs and QIs by random sampling to adjust for the influence of the number of records. Therefore, it is necessary to reinforce a level of privacy protection for each institution because the DRN involves big data research based on multi-institution collaboration. Our study findings can help in constructing an advanced DRN environment that protects these privacy risks as a quality measurement index.

Acknowledgments

This study was supported by the Foundational Technology Development Program (NRF2019M3E5D406468221) of the Ministry of Science and ICT, Republic of Korea, and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI19C0189).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Forty-five personal health identifiers and 17 quasi-identifiers in the structure of the Observational Medical Outcome Partnership common data model.

[DOCX File , 18 KB - medinform_v9i5e24940_app1.docx]

Multimedia Appendix 2

Detailed information of 16 personal health identifier variables and 12 quasi-identifier scenarios.

[DOCX File , 37 KB - medinform_v9i5e24940_app2.docx]

Multimedia Appendix 3

Seven scenarios with personal health identifiers and quasi-identifiers in the Observational Medical Outcome Partnership common data model, based on not null values in the Synthetic Public Use File 5 Percent data set.

[DOCX File , 17 KB - medinform_v9i5e24940_app3.docx]

Multimedia Appendix 4

Random sampling and 100 iterations were conducted to compare records with a minimum cell size of one between the limited and safe harbor data sets from 16 personal health identifier data sets.

[DOCX File , 27 KB - medinform_v9i5e24940_app4.docx]

Multimedia Appendix 5

Percentage of records measuring quasi-identifier privacy risk with gaps between the limited and safe harbor data sets with a minimum cell size of one to five.

[DOCX File , 15 KB - medinform_v9i5e24940_app5.docx]

References

1. NIH Collaboratory Distributed Research Network (DRN). Rethinking Clinical Trials. URL: https://rethinkingclinicaltrials.org/nih-collaboratory-drn/ [accessed 2020-08-01]
2. Passerat-Palmbach J, Farnan T, McCoy M, Harris JD, Manion ST, Flannery HL, et al. Blockchain-orchestrated machine learning for privacy preserving federated learning in electronic health data. 2020 Presented at: 2020 IEEE International Conference on Blockchain (Blockchain); November 2-6, 2020; Rhodes, Greece. [doi: 10.1109/Blockchain50366.2020.00080]
3. Cheu A, Smith A, Ullman J, Zeber D, Zhilyaev M. Distributed Differential Privacy via Shuffling. In: Ishai Y, Rijmen V, editors. Advances in Cryptology – EUROCRYPT 2019. EUROCRYPT 2019. Lecture Notes in Computer Science, vol 11476. Cham: Springer; 2019:375-403.
4. Konečný J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D. Federated Learning: Strategies for Improving Communication Efficiency. arXiv. 2016. URL: https://arxiv.org/abs/1610.05492 [accessed 2021-05-22]
5. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med 2020 Sep 14;3(1):119-117 [FREE Full text] [doi: 10.1038/s41746-020-00323-1] [Medline: 33015372]
6. Tkachenko O, Weinert C, Schneider T, Hamacher K. Large-Scale Privacy-Preserving Statistical Computations for Distributed Genome-Wide Association Studies. In: ASIACCS '18: Proceedings of the 2018 Asia Conference on Computer and Communications Security. 2018 Presented at: 2018 Asia Conference on Computer and Communications Security; June 2018; Incheon, Republic of Korea p. 221-235. [doi: 10.1145/3196494.3196541]
7. Tomsett R, Chan K, Chakraborty S. Model poisoning attacks against distributed machine learning systems. In: Proceedings Volume 11006, Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications. 2019 Presented at: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications; 2019; Baltimore, MD. [doi: 10.1117/12.2520275]
8. Takabi D, Podschwadt R, Druce J, Wu C, Procopio K. Privacy preserving Neural Network Inference on Encrypted Data with GPUs. arXiv. 2019. URL: https://arxiv.org/abs/1911.11377 [accessed 2021-05-22]
9. Dahl M, Mancuso J, Dupis Y, Decoste B, Giraud M, Livingstone I, et al. Private Machine Learning in TensorFlow using Secure Computation. arXiv. 2018. URL: https://arxiv.org/abs/1810.08130 [accessed 2021-05-22]
10. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell 2020 Jun 8;2(6):305-311. [doi: 10.1038/s42256-020-0186-1]
11. Salem M, Taheri S, Yuan J. Utilizing Transfer Learning and Homomorphic Encryption in a Privacy Preserving and Secure Biometric Recognition System. Computers 2018 Dec 29;8(1):3-24. [doi: 10.3390/computers8010003]
12. FitzHenry F, Resnic F, Robbins S, Denton J, Nookala L, Meeker D, et al. Creating a Common Data Model for Comparative Effectiveness with the Observational Medical Outcomes Partnership. Appl Clin Inform 2017 Dec 19;06(03):536-547. [doi: 10.4338/aci-2014-12-cr-0121]
13. Klann JG, Joss MAH, Embree K, Murphy SN. Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model. PLoS One 2019 Feb 19;14(2):e0212463-e0212413 [FREE Full text] [doi: 10.1371/journal.pone.0212463] [Medline: 30779778]
14. Gujarathi G, Ma Y. Parametric CAD/CAE integration using a common data model. Journal of Manufacturing Systems 2011 Aug;30(3):118-132. [doi: 10.1016/j.jmsy.2011.01.002]
15. Saver JL, Warach S, Janis S, Odenkirchen J, Becker K, Benavente O, et al. Standardizing the Structure of Stroke Clinical and Epidemiologic Research Data. Stroke 2012 Apr;43(4):967-973. [doi: 10.1161/strokeaha.111.634352]
16. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012 Jan 01;19(1):54-60 [FREE Full text] [doi: 10.1136/amiajnl-2011-000376] [Medline: 22037893]
17. Mamo L, Browe D, Logan H, Kim K. Patient informed governance of distributed research networks: results and discussion from six patient focus groups. AMIA Annu Symp Proc 2013;2013:920-929 [FREE Full text] [Medline: 24551383]
18. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
19. Data. The National Patient-Centered Clinical Research Network. URL: https://pcornet.org/data/ [accessed 2020-08-01]
20. Ji H, Kim S, Yi S, Hwang H, Kim J, Yoo S. Converting clinical document architecture documents to the common data model for incorporating health information exchange data in observational health studies: CDA to CDM. J Biomed Inform 2020 Jul;107:103459-103457. [doi: 10.1016/j.jbi.2020.103459] [Medline: 32470694]
21. Timbie J, Rudin R, Towe V, Chen E, Hunter L, Case S, et al. National Patient-Centered Clinical Research Network (PCORnet) Phase I: Final Evaluation Report. RAND. URL: https://www.rand.org/pubs/research_reports/RR1191.html [accessed 2021-05-22]
22. Martin-Sanchez FJ, Aguiar-Pulido V, Lopez-Campos GH, Peek N, Sacchi L. Secondary Use and Analysis of Big Data Collected for Patient Care. Yearb Med Inform 2017 Aug 19;26(01):28-37. [doi: 10.1055/s-0037-1606529]

23. Zhang Y, Steele A, Blanton M. PICCO: a general-purpose compiler for private distributed computation. In: CCS '13: Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security. 2013 Presented at: 2013 ACM SIGSAC Conference on Computer & Communications Security; November 2013; Berlin, Germany p. 813-826. [doi: 10.1145/2508859.2516752]
24. Zhang Y, Blanton M, Almashaqbeh G. Secure distributed genome analysis for GWAS and sequence comparison computation. BMC Med Inform Decis Mak 2015 Dec 21;15(S5):1-12. [doi: 10.1186/1472-6947-15-s5-s4]
25. Dubovitskaya A, Urovi V, Vasirani M, Aberer K, Schumacher M. A Cloud-Based eHealth Architecture for Privacy Preserving Data Integration. In: Federrath H, Gollmann D, editors. ICT Systems Security and Privacy Protection. SEC 2015. IFIP Advances in Information and Communication Technology, vol 455. Cham: Springer; 2015:585-598.
26. Kho A, Cashy J, Jackson K, Pah A, 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]
27. Huang L, Chu H, Lien C, Hsiao C, Kao T. Privacy preservation and information security protection for patients' portable electronic health records. Comput Biol Med 2009 Sep;39(9):743-750. [doi: 10.1016/j.compbiomed.2009.06.004] [Medline: 19589509]
28. Sahi MA, Abbas H, Saleem K, Yang X, Derhab A, Orgun MA, et al. Privacy Preservation in e-Healthcare Environments: State of the Art and Future Directions. IEEE Access 2018 Oct;6:464-478. [doi: 10.1109/access.2017.2767561]
29. Seol K, Kim Y, Lee E, Seo Y, Baik D. Privacy-Preserving Attribute-Based Access Control Model for XML-Based Electronic Health Record System. IEEE Access 2018 Feb 5;6:9114-9128. [doi: 10.1109/access.2018.2800288]
30. Blacketer C. Chapter 4 The Common Data Model. The Book of OHDSI. URL: https://ohdsi.github.io/TheBookOfOhdsi/CommonDataModel.html [accessed 2020-08-01]
31. Lee S, You SC, Park J, Cho J, Borel S, El Emam K, et al. OMOP-CDM Conversion and Anonymization of National Health Insurance Service-National Sample Cohort. Observational Health Data Sciences and Informatics. 2019. URL: https://www.ohdsi.org/2019-us-symposium-showcase-17/ [accessed 2021-05-22]
32. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. U.S. Department of Health & Human Services. 2015. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html [accessed 2020-08-01]
33. Wiley DC, Cory AC, editors. Health Insurance Portability and Accountability Act. In: Encyclopedia of School Health. Thousand Oaks, CA: SAGE Publications, Inc; 2013.
34. Murphy SN, Gainer V, Mendis M, Churchill S, Kohane I. Strategies for maintaining patient privacy in i2b2. J Am Med Inform Assoc 2011 Dec 01;18 Suppl 1(Supplement 1):i103-i108 [FREE Full text] [doi: 10.1136/amiajnl-2011-000316] [Medline: 21984588]
35. Li D. Data management system having a common database infrastructure. Google Patents. 2008. URL: https://patents.google.com/patent/US20040054675 [accessed 2021-05-22]
36. Olivier MS. Database privacy. SIGKDD Explor. Newsl 2002 Dec;4(2):20-27. [doi: 10.1145/772862.772866]
37. Blum A, Ligett K, Roth A. A learning theory approach to noninteractive database privacy. J. ACM 2013 Apr;60(2):1-25. [doi: 10.1145/2450142.2450148]
38. Bergquist T, Brandt P. Prometheus: Differential Privacy in the OMOP CDM. University of Washington. 2018. URL: https://courses.cs.washington.edu/courses/cse544/18wi/project/examples-successful-projects/psbrandt.pdf [accessed 2021-05-22]
39. Slavic A, Cordeiro M. Sharing and re-use of classification systems: the need for a common data model. The University of Arizona Libraries. 2005. URL: https://repository.arizona.edu/handle/10150/105132 [accessed 2021-05-22]
40. Makadia R, Ryan PB. Transforming the Premier Perspective Hospital Database into the Observational Medical Outcomes Partnership (OMOP) Common Data Model. EGEMS (Wash DC) 2014 Nov 11;2(1):1110-1110 [FREE Full text] [doi: 10.13063/2327-9214.1110] [Medline: 25848597]
41. Ganslandt T, Mate S, Helbing K, Sax U, Prokosch H. Unlocking Data for Clinical Research – The German i2b2 Experience. Appl Clin Inform 2017 Dec 16;02(01):116-117. [doi: 10.4338/aci-2010-09-cr-0051]
42. Liyanage H, Liaw S, Jonnagaddala J, Hinton W, de Lusignan S. Common Data Models (CDMs) to Enhance International Big Data Analytics: A Diabetes Use Case to Compare Three CDMs. Stud Health Technol Inform 2018;255:60-64. [Medline: 30306907]
43. Raisaro JL, Choi G, Pradervand S, Colsenet R, Jacquemont N, Rosat N, et al. Protecting Privacy and Security of Genomic Data in i2b2 With Homomorphic Encryption and Differential Privacy. IEEE/ACM Trans. Comput. Biol. and Bioinf 2018:1-1. [doi: 10.1109/tcbb.2018.2854782]
44. Sweeney L. k-anonymity: A model for protecting privacy. Int. J. Unc. Fuzz. Knowl. Based Syst 2012 May 02;10(05):557-570. [doi: 10.1142/s0218488502001648]
45. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. L-diversity: privacy beyond k-anonymity. 2006 Presented at: 22nd International Conference on Data Engineering (ICDE'06); April 3-7, 2006; Atlanta, GA. [doi: 10.1109/icde.2006.1]
46. Li N, Li T, Venkatasubramanian S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. 2007 Presented at: 2007 IEEE 23rd International Conference on Data Engineering; April 15-20, 2007; Istanbul, Turkey. [doi: 10.1109/icde.2007.367856]

47. Shin SJ, You SC, Park YR, Roh J, Kim J, Haam S, et al. Genomic Common Data Model for Seamless Interoperation of Biomedical Data in Clinical Practice: Retrospective Study. J Med Internet Res 2019 Mar 26;21(3):e13249 [FREE Full text] [doi: 10.2196/13249] [Medline: 30912749]
48. OHDSI/CommonDataModel. GitHub. URL: https://github.com/OHDSI/CommonDataModel/wiki [accessed 2020-08-01]
49. CMS 2008-2010 Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF). U.S. Centers for Medicare & Medicaid Services. URL: https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DE_Syn_PUF [accessed 2020-08-01]
50. Data Standardization. Observational Health Data Sciences and Informatics. URL: https://ohdsi.org/data-standardization/ [accessed 2020-08-01]
51. OHDSI Google Drive. URL: https://drive.google.com/file/d/18EjMxyA6NsqBo9eed_Gab1ESHWPxJygz/view [accessed 2020-08-01]
52. El Emam K, Arbuckle L, Koru G, Eze B, Gaudette L, Neri E, et al. De-identification methods for open health data: the case of the Heritage Health Prize claims dataset. J Med Internet Res 2012 Feb 27;14(1):e33 [FREE Full text] [doi: 10.2196/jmir.2001] [Medline: 22370452]
53. Gong M, Wang S, Wang L, Liu C, Wang J, Guo Q, et al. Evaluation of Privacy Risks of Patients' Data in China: Case Study. JMIR Med Inform 2020 Feb 05;8(2):e13046 [FREE Full text] [doi: 10.2196/13046] [Medline: 32022691]
54. Benitez K, Malin K. Evaluating re-identification risks with respect to the HIPAA privacy rule. J Am Med Inform Assoc 2010;17(2):169-177 [FREE Full text] [doi: 10.1136/jamia.2009.000026] [Medline: 20190059]
55. Garfinkel SL. De-Identification of Personal Information. National Institute of Standards and Technology. 2015. URL: https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf [accessed 2021-05-17]
56. Committee on Strategies for Responsible Sharing of Clinical Trial Data, Board on Health Sciences Policy, Institute of Medicine. Concepts and Methods for De-identifying Clinical Trial Data. In: Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Washington, DC: National Academies Press (US); 2015.
57. Lee YJ, Lee KH. What are the optimum quasi-identifiers to re-identify medical records? 2018 Presented at: 20th International Conference on Advanced Communication Technology (ICACT); February 11-14, 2018; Chuncheon, Korea (South). [doi: 10.23919/icact.2018.8323926]
58. Definition and DDLs for the OMOP Common Data Model (CDM). GitHub. URL: https://github.com/OHDSI/CommonDataModel/tree/v5.2.2 [accessed 2020-08-01]
59. Zhou X, Murugesan S, Bhullar H, Liu Q, Cai B, Wentworth C, et al. An evaluation of the THIN database in the OMOP Common Data Model for active drug safety surveillance. Drug Saf 2013 Feb 4;36(2):119-134. [doi: 10.1007/s40264-012-0009-3] [Medline: 23329543]
60. Si Y, Weng C. An OMOP CDM-Based Relational Database of Clinical Research Eligibility Criteria. Stud Health Technol Inform 2017;245:950-954 [FREE Full text] [Medline: 29295240]
61. Glicksberg B, Oskotsky B, Thangaraj P, Giangreco N, Badgeley M, Johnson K, et al. PatientExploreR: an extensible application for dynamic visualization of patient clinical history from electronic health records in the OMOP common data model. Bioinformatics 2019 Nov 01;35(21):4515-4518 [FREE Full text] [doi: 10.1093/bioinformatics/btz409] [Medline: 31214700]
62. Karr AF, Feng J, Lin X, Sanil AP, Young SS, Reiter JP. Secure analysis of distributed chemical databases without data integration. J Comput Aided Mol Des 2005 Nov 3;19(9-10):739-747. [doi: 10.1007/s10822-005-9011-5] [Medline: 16267693]
63. Domingo-Ferrer J. Microaggregation for Database and Location Privacy. In: Etzion O, Kuflik T, Motro A, editors. Next Generation Information Technologies and Systems. NGITS 2006. Lecture Notes in Computer Science, vol 4032. Berlin, Heidelberg: Springer; 2006:106-116.
64. Raab M, Steger A. "Balls into Bins" — A Simple and Tight Analysis. In: Luby M, Rolim JDP, Serna M, editors. Randomization and Approximation Techniques in Computer Science. RANDOM 1998. Lecture Notes in Computer Science, vol 1518. Berlin, Heidelberg: Springer; 1998:159-170.
65. Shin S, Park YR, Shin Y, Choi HJ, Park J, Lyu Y, et al. A De-identification method for bilingual clinical texts of various note types. J Korean Med Sci 2015 Jan;30(1):7-15 [FREE Full text] [doi: 10.3346/jkms.2015.30.1.7] [Medline: 25552878]
66. Abdalla M, Abdalla M, Rudzicz F, Hirst G. Using word embeddings to improve the privacy of clinical notes. J Am Med Inform Assoc 2020 Jun 01;27(6):901-907 [FREE Full text] [doi: 10.1093/jamia/ocaa038] [Medline: 32388549]

Abbreviations

CDM: common data model
DRN: distributed research network
EC: equivalence class
HIPAA: Health Insurance Portability and Accountability Act
NPI: national provider identifier
OHDSI: Observational Health Data Sciences and Informatics
OMOP: Observational Medical Outcome Partnership

PHI: personal health identifier
QI: quasi-identifier
SynPUF5PCT: Synthetic Public Use File 5 Percent

Edited by G Eysenbach; submitted 12.10.20; peer-reviewed by G Weber; comments to author 03.11.20; revised version received 27.12.20; accepted 11.04.21; published 31.05.21.

Please cite as:
Oh S, Sung M, Rhee Y, Hong N, Park YR
Evaluation of the Privacy Risks of Personal Health Identifiers and Quasi-Identifiers in a Distributed Research Network: Development and Validation Study
JMIR Med Inform 2021;9(5):e24940
URL: https://medinform.jmir.org/2021/5/e24940
doi:10.2196/24940
PMID:34057426

Original Paper

Use of Machine Learning Algorithms to Predict the Understandability of Health Education Materials: Development and Evaluation Study

Meng Ji[1], PhD; Yanmeng Liu[1], MA; Mengdan Zhao[1], MA; Ziqing Lyu[1,2], MA; Boren Zhang[1], MA; Xin Luo[3,4], MA; Yanlin Li[3,4], MEd; Yin Zhong[5,6], PhD

[1]School of Languages and Cultures, University of Sydney, Sydney, Australia

[2]School of Foreign Languages, Jiangsu University of Science and Technology, Zhenjiang, China

[3]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, Hong Kong

[4]The HK PolyU-PKU Research Centre on Chinese Linguistics, The Hong Kong Polytechnic University, Hong Kong, Hong Kong

[5]Department of English, The Hong Kong Polytechnic University, Hong Kong, Hong Kong

[6]Research Centre for Professional Communication in English, The Hong Kong Polytechnic University, Hong Kong, Hong Kong

Corresponding Author:
Yanmeng Liu, MA
School of Languages and Cultures
University of Sydney
Camperdown
Sydney, NSW2006
Australia
Phone: 61 449858887
Email: yanmeng.liu@sydney.edu.au

Abstract

Background: Improving the understandability of health information can significantly increase the cost-effectiveness and efficiency of health education programs for vulnerable populations. There is a pressing need to develop clinically informed computerized tools to enable rapid, reliable assessment of the linguistic understandability of specialized health and medical education resources. This paper fills a critical gap in current patient-oriented health resource development, which requires reliable and accurate evaluation instruments to increase the efficiency and cost-effectiveness of health education resource evaluation.

Objective: We aimed to translate internationally endorsed clinical guidelines to machine learning algorithms to facilitate the evaluation of the understandability of health resources for international students at Australian universities.

Methods: Based on international patient health resource assessment guidelines, we developed machine learning algorithms to predict the linguistic understandability of health texts for Australian college students (aged 25-30 years) from non-English speaking backgrounds. We compared extreme gradient boosting, random forest, neural networks, and C5.0 decision tree for automated health information understandability evaluation. The 5 machine learning models achieved statistically better results compared to the baseline logistic regression model. We also evaluated the impact of each linguistic feature on the performance of each of the 5 models.

Results: We found that information evidentness, relevance to educational purposes, and logical sequence were consistently more important than numeracy skills and medical knowledge when assessing the linguistic understandability of health education resources for international tertiary students with adequate English skills (International English Language Testing System mean score 6.5) and high health literacy (mean 16.5 in the Short Assessment of Health Literacy-English test). Our results challenge the traditional views that lack of medical knowledge and numerical skills constituted the barriers to the understanding of health educational materials.

Conclusions: Machine learning algorithms were developed to predict health information understandability for international college students aged 25-30 years. Thirteen natural language features and 5 evaluation dimensions were identified and compared in terms of their impact on the performance of the models. Health information understandability varies according to the demographic profiles of the target readers, and for international tertiary students, improving health information evidentness, relevance, and logic is critical.

(JMIR Med Inform 2021;9(5):e28413) doi:10.2196/28413

KEYWORDS

machine learning; PEMAT; health education; understandability evaluation; patient-oriented

Introduction

Background

The World Health Organization recommends a set of principles for effective health communication, including accessibility, actionability, credibility, relevance, timeliness, and understandability [1]. Health information understandability can be achieved by using familiar language and good writing practice that highlights health information directness, clearness of the desired health outcome, easy-to-follow informational organization, and discourse explicitness, that is, clear explanation of health and medical knowledge using simple, plain, and purposeful language [2-5]. Approaches to health information evaluation can be divided into 2 large categories, that is, expert-led qualitative evaluation based on clinical experiences [6-9] and automated health information analyzers using medical readability formulas or natural language processing tools [10-13]. The strengths and limitations of both approaches are well-known [14-16]. Expert-led health material evaluation draws upon the domain knowledge of medical and health professionals, which are insightful and clinically reliable. This approach, however, is costly and requires much longer evaluation timeframes when compared to automated evaluations. They have important limitations with the evaluation of health materials in large quantities or in situations that require more regular, instant evaluation such as health information updates in health emergencies. Further, this approach is not flexible with user-oriented health information evaluation that requires the evaluation criteria adjust with flexibility to align with the actual reading abilities of the patient education resource users [17,18]. For example, the same piece of health information can be of varying understandability for users with different education levels, health literacy, or existing knowledge of specific health topics. By contrast, the computerized approach of evaluating health information based on natural language features is gaining importance in health informatics [19-22].

Developing health resources of adequate understandability can have important impact on the trust, acceptance, and voluntary adherence to the health advice and recommendations delivered in the health texts [23-26]. Information simplification is an effective strategy to increase the understandability of health materials. However, with specialized health texts, oversimplification can result in critical information loss and reduced believability and persuasiveness of health information for educated readers with higher health information appraisal abilities and health risk assessment autonomies. How to maintain a balance between the understandability and the informativeness of health materials holds the key to optimal health communications. This paper leverages machine learning techniques to develop automated health information evaluation tools of English health materials for a specific group of readers, that is, students in tertiary education from non-English speaking backgrounds with intermediate English reading skills (they achieved an average 6.5 score in the International English Language Testing System test). The stress of living away from home and the adjustment difficulties among international students is known as the “foreign student syndrome” [27,28]. Previous research with students in Australia and internationally showed that international students were less likely to seek for help from health organizations than from residents [29]. English health materials available on the websites of health authorities thus provide important sources of information for international students. Whether and how health information from health authorities developed for native English readers is understandable to international students with intermediate English skills and limited health literacy remains unknown. In this study, new machine learning algorithms were developed to predict the linguistic readability of original English health information for international students. Our study illustrates the training and validation of machine learning algorithms to predict the understandability of health education materials on infectious diseases for this group of English health information users. The strength of machine learning algorithms, that is, adaptiveness and flexibility can significantly improve the cost-effectiveness and efficiency of automated health educational resource evaluation for specific user groups.

The contributions of our study are three-fold: first, we translated clinical health education material evaluation guidelines to machine algorithms to enable the quantitative evaluation of understandability of health materials. This has, for the first time, materialized the automation of health resource understandability assessment with specific reader groups, which represents a significant advance in user-oriented health information evaluation. Second, the results of the machine learning–based evaluation identified important new dimensions in information readability assessment, which are health information purposefulness and the logical structure of health texts. These new findings challenged traditional views that lack of health text readability was caused by morphological complexity and domain-specific terminology. Such views largely simplified the complex issue of the cognitive processing of health information by populations of varying education and health literacy levels and language and cultural backgrounds. Our study shows that for nonnative English readers with tertiary education and high health literacy levels, health information evidentness, logical sequence, and relevance for educational purposes weigh more than health domain knowledge and numeracy demands when assessing the understandability of health texts for readers from similar backgrounds. Lastly, our study identified textual linguistic features having large impact on the performance of machine learning algorithms. For information evidentness, these were words describing mental actions and processes (X2), general/abstract terms (A1), and for relevance to educational purposes, these were words of anatomy, physiology (B1), medicines, and medical treatment (B3). For logical sequence, these were grammatical words (Z5), negative (Z6), and conditional expressions (Z7). Different from statistics, machine learning cannot compute the regression coefficients of these variables within different models, but the large impact of these features on health text readability suggests that linguistic interventions to these features of health texts can significantly improve the performance of machine learning algorithms as automated health text readability evaluation applications.

Data Sets and Feature Selection

Data Collection and Classification

The health educational resources were collected from diverse sources, including governmental health agencies and not-for-profit health organizations in Australia. Health education resource genres are highly diverse, which may be classified into fact sheets, health topics, patient guidelines, clinical guidelines, administrative guidelines, manuals, reports, booklets, brochures, posters, leaflets, checklists, and flipcharts. In this study, we purposely selected health education resources from fact sheets, health topics, and patient guidelines, which are some of the most used health resource varieties. The main sources of credible health information were the Australian Federal and State Health departments and not-for-profit organizations: Arthritis Australia, Australian Food Safety Information Council, Australian Melanoma Research Foundation, Australian Rotary Health, Breast Cancer Network Australia, Cancer Council Australia, Diabetes Australia, National Breast Cancer Foundation, National Heart Foundation of Australia, and National LGBTI Health Alliance. The total corpus contained 1000 full-length health educational texts (running tokens of over 500,000 words). Five international students in tertiary education enrolled in Australian universities classified the collected health texts independently into easy versus hard-to-understand categories (Cohen kappa 0.705). They were aged 25-30 years with advanced English skills (International English Language Testing System test score 6.5 or above). Their mean health literacy level (16.5 [SD 1.69], IQR 13-18) was measured using the Short Assessment of Health Literacy-English [30,31], and their mean level was 87.5% over the threshold 14 of low health literacy.

Textual Features as Health Information Understandability Predictors

In order to develop automated health resource evaluation algorithms, we identified a set of key linguistic features as relevant to the understandability of written health resources. Table 1 lists some of the evaluation criteria in the Patient Education Materials Assessment Tool (PEMAT) developed by the Agency for Healthcare Research and Quality, United States Department of Health and Human Services [32]. These include the evaluation of health content, word choice and style, use of numbers, and textual organization. Each evaluation criterion was then mapped onto one or multiple semantic classes of the UCREL English Semantic Analysis System (USAS) developed by the University of Lancaster, United Kingdom [33]. We used USAS to annotate the raw English corpus texts collected. USAS is one of the most used English semantic annotation systems. It has a multi-tier structure with 21 major discourse fields covering (A) general and abstract terms, (B) the body and the individual, (C) arts and crafts, (D) emotion, (E) food and farming, (G) government and public, (H) housing and home, (I) money and commerce, (K) sports and games, (L) live and living things, (M) movement and transport, (N) numbers and measurement, (O) substances, materials, objects, and equipment, (P) education, (Q) language and communication, (S) social actions, states, and processes, (T) time, (W) world and environment, (X) psychological actions, states and processes, (Y) science and technology, and (Z) names and grammars. Within each large semantic category (A-Z), there are subcategories providing fine-grained classification of the word semantics. For example, the A category contains A1 general and abstract terms, A2 affect, A3 being, A4 classification, A5 evaluation, A6 comparison, A7 probability, A8 seem, A9 possession, and so on. These natural language features were then mapped onto the PEMAT evaluation criteria as shown in Table 1.

Table 1. Natural language features relevant to Patient Education Materials Assessment Tool guidelines.

Evaluation criteria in the Patient Education Materials Assessment Tool	Language features	Machine learning evaluation
Content		
The material makes its purpose completely evident.	A1, X1, X2, X7	Information evidentness
The material does not include information that distracts from its purpose.	B1, B3	Relevance to education purpose
Word choice and style		
Medical terms are used only to familiarize audience with the terms.	B2	Domain knowledge
Use of numbers		
The material does not expect the user to perform calculations.	N1, N2, N3	Numeracy demand
Organization		
The material presents information in a logical sequence.	Z5, Z6, Z7	Logical sequence

To quantify the PEMAT guideline item “the material makes its purpose completely evident,” 4 USAS classes were used as quantitative measures, that is, A1: general and abstract terms; X1: psychological actions, states, and processes; X2: mental actions and processes (such as think, analyze, study, look over, go over); and X7: wanting, planning, choosing (such as aim, objective, goal, target, intention, purpose, plan, idea, point). To quantify the PEMAT guideline item “the material does not include information or content that distracts from its purpose,” 2 USAS classes were used as quantitative measures, that is, B1:

anatomy and physiology and B3: medicines and medical treatment. Typical examples of content distraction include excessive detail about the equipment used for a procedure that distracts from the material's purpose or excessive detail about other procedures or treatments that are not related to the material's purpose. To quantify the PEMAT guideline item "medical terms are used only to familiarize audience with the terms," the USAS class B2: health and disease terms were used as the main quantitative measure. To quantify the PEMAT guideline item "the material does not expect the user to perform calculations," 3 USAS classes were selected from the USAS semantic tag set as relevant quantitative measures. To quantify the PEMAT guideline item "the material presents information in a logical sequence," 3 USAS classes were identified as relevant to the logical structure of health materials, that is, Z5: grammatical bin, Z6: negative, and Z7: if (conditional). In total, 13 semantic annotation classes were selected from the extensive tag set of USAS. Information evidentness of written health texts is measured by A1, X1, X2, X7; information relevance to educational purposes by B1 and B3; health domain knowledge by B2; health numeracy demand by N1, N2, and N3; and lastly, text logical sequence by grammatical and functional features Z5, Z6, and Z7.

Analysis of the Differences Between Easy and Difficult Texts

Table 2 shows the statistical results of the differences between easy and difficult health educational texts for international college students. All the predictor variables were continuous variables, and the *P* values were derived using Mann-Whitney *U* test. The result shows that statistically significant differences (*P*<.05) exist in most of the semantic features. Easy and difficult health texts, however, did not differ significantly in the semantic classes of B1 (anatomy, physiology), N1 (numbers), N2 (mathematics), and N3 (measurement). The mean values of the 7 semantic classes of easy health texts were significantly higher than those of difficult health texts. In terms of health information purposefulness, 4 semantic features contributed to the linguistic understandability of health resources, that is, A1 (14.09 easy vs 10.10 difficult), X1 (0.42 easy vs 0.18 difficult), X2 (10.41 easy vs 6.57 difficult), and X7 (3.24 easy vs 1.79 difficult). This suggests that the increased use of words describing the psychological and mental actions, states, and processes can help the target readers to understand the textual information. A1 is defined as general and abstract words.

Table 2. Differences between easy and difficult medical texts derived by the Mann-Whitney *U* test.

Variables	Easy texts, mean (SD) score	Difficult texts, mean (SD) score	Mann-Whitney *U*	*P* value
A1	14.09 (14.52)	10.10 (13.13)	97905.00	<.001
X1	0.42 (3.89)	0.18 (1.49)	120325.50	.02
X2	10.41 (11.26)	6.57 (9.14)	89487.50	<.001
X7	3.24 (5.41)	1.79 (3.12)	103350.50	<.001
B1	17.10 (31.14)	15.69 (21.78)	117882.50	.12
B2	15.04 (21.53)	24.68 (34.04)	99536.50	<.001
B3	9.25 (14.30)	12.80 (18.02)	103338.00	<.001
N1	5.74 (8.54)	5.42 (6.51)	123009.00	.66
N2	0.21 (0.70)	0.21 (0.70)	123284.50	.52
N3	5.73 (9.38)	4.77 (5.60)	120978.50	.38
Z5	133.63 (118.93)	122.77 (119.05)	108744.00	<.001
Z6	4.13 (5.28)	3.01 (5.01)	100719.00	<.001
Z7	4.22 (4.62)	2.10 (4.03)	81063.50	<.001

Table 3 shows some of the words annotated as A1 in a typical health text classified as difficult. These general and abstract words were not typical medical and health terms. They were classified and tagged in the corpus study as general English terms. However, the statistically significant *P* value attributed to this word category as shown in Table 2 indicated that they can be used as a discriminating feature to separate easy versus difficult health educational materials for international students in tertiary education. Regarding health domain knowledge, the result shows that the mean of B2 (health and disease) of easy health texts (15.04) was significantly lower than that of difficult texts (24.68). In terms of numeracy demand, the 2 sets of health texts did not different significantly, suggesting that for international students in tertiary education, the use of numbers and quantitative measures in health educational texts did not represent an important barrier. Lastly, the logical sequence of English health texts can be improved using functional words (Z5, Z6, Z7), as the mean scores of these 3 linguistic features in easy health educational resources proved to be significantly higher than those of difficult texts: Z5 (133.63 easy, 122.77 difficult), Z6 (4.13 easy, 3.01 difficult) and Z7 (4.22 easy, 2.10 difficult).

Table 3. A1 in difficult texts.

A1	Keyword concordances
limited to	infections in humans are *limited* to one case of Taï Forest Ebola virus
strains	There are five *strains* that have been identified: Zaire, Sudan, Bundibugyo, Taï Forest, and Reston.
containment	Previous outbreaks had been limited to remote areas allowing initial *containment* efforts to be more effective.
combined	This outbreak was unprecedented in scale, being larger than all other outbreaks *combined.*
spread	The virus *spread* across multiple international boundaries.
boundaries	The virus spread across multiple international *boundaries.*
isolated	Seven other countries had minor outbreaks with nonsustained transmission or *isolated* cases.
events	This article aims to summarize the *events* by country in chronological order.

Methods

Machine Learning Algorithms

The 5 machine learning methods used in this study were extreme gradient boosting (XGBoost) tree, random forest, deep neural networks, and C5.0 decision tree. Logistic regression was used as the baseline model for the evaluation of the performance of the 5 machine learning models. Both XGBoost and random forest are ensemble learning techniques that can be used for both classification and regression issues. Ensemble learning can boost the predictive performance of a single learning algorithm, which is merely better than random guesses. Random forest uses bagging or bootstrap to combine base learners to significantly improve the prediction of the model. XGBoost uses gradient boosting to combine decision trees as base learners. The C5.0 decision tree is a typical tree-based machine learning algorithm. XGBoost, random forest, and C5.0 can be used to learn any patterns underlying the training data without implicit assumptions of the data profiles, such as distribution normality, nonlinearity, multi-linearity, or higher order interactions between the variables. The type of neural networks used in this study is multilayer perceptron, which is a class of feedforward artificial neural network. This technique has been used to provide a nonlinear mapping between the input vector and the output vector. Between the input and output layers, there could be an arbitrary number of hidden layers, which perform complex computations. The strength of multilayer perceptron is to map nonlinear relations between input features and outcomes. The major uses of multilayer perceptron are pattern classification, recognition, prediction, and approximation. The research work of this paper can be seen as a text classification task. Random forest is suitable for analyzing data of high dimensions, as the algorithm builds separate trees and uses bootstrapping to combine these tree-based single learners trained on random subsets of input features. Like random forest, gradient boosting tree is a type of supervised learning algorithm known for its high prediction accuracy.

Hyperparameters of Machine Learning Algorithms

In this study, hyperparameter tuning of XGBoost involved the following steps. The maximum tree depth for base learners (max_depth) controls the depth of the tree. The larger the depth, the more complex is the model, and the higher are the chances of model overfitting. There is no standard value for max_depth. Larger data sets require deep trees to learn the rules from a complex data set. The value ranges between 0 and infinite. In the cross-validation process, we set max_depth to the default value 8. The number of estimators or boosted trees was set to the default value 20. The minimum sum of instance weight needed in a child node (min_child_weight) is another effective overfitting prevention method. It is calculated by second-order partial derivatives and ranges between 0 and infinite. The larger the value, the more conservative the algorithm is. This was set to the default value of 1 in this study. The maximum delta step (max_delta_step) specifies the maximum step size that a leaf node can take. It ranges between 0 and infinite. Increasing the positive value will make the update step more conservative. The learning objective was set to binary logistic regression, as the target variable has 2 outcome categories, that is, easy versus difficult health education texts. Subsample refers to the subsample ratio of the training instance. For example, setting a subsample to 0.5 means that the algorithm randomly collects half of the entire data set to build the tree model. The value of the subsample was set to the default value 1. Eta refers to the machine learning rate at which the algorithm learns the latent patterns and structures in the training data set. Smaller eta leads to slower computation and thus prevents overfitting. Smaller etas can be compensated by increasing the number of boosted trees or estimators; 0.6 was set as the value in this study. The hyperparameter colsample_bytree controls the number of features or variables supplied to a tree model. It was set to 1. Lastly, alpha and lambda values, which control L1 and L2 regularization, respectively, were set to 1 and 0 to prevent overfitting. Random forest is another powerful ensemble learning technique that outperforms single learning algorithms in machine learning model development. In random forest, decision trees are used as the base learner and bootstrapping aggregation combines these decision trees together to achieve high prediction accuracy. The minimum number of samples and training data required to be at a leaf node (min_samples_leaf) was set to 1. The maximum depth was set to 10. The number of features to use for splitting was set to auto. In the model construction process, the ensemble learning methods selected to increase the prediction accuracy included bootstrapping, bagging, and extremely randomized trees. In the process of hyperparameter optimization, on each iteration, the algorithm will choose a different combination of the features. The maximum number of iterations was set to 1000, and the maximum evaluations were set to 300. The neural networks model used in this study is multilayer perceptron. Only one hidden layer was configured, which contained 13 nodes as the input features (Table 1). The overfitting prevention rate was set to 30%.

Results

Predictive Performance Evaluation

The predictive performance of the 5 machine learning algorithms is shown in Figure 1 and Table 4, and the results of the pairwise corrected resampled two-tailed *t* test are shown in Table 5. The mean scores and their standard deviations of area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and accuracy were obtained through five-fold cross-validation. The cross-validation divided the entire data set into 5 folds of equal size. In each iteration, 5 folds were used as the training data and the remaining fold as the testing data. As a result, on completion of the five-fold cross-validation, each fold was used as the testing data exactly once. We used the pairwise corrected resampled *t* test to counteract the issue of multiple comparisons. The significance level was adjusted to .005 using Bonferroni correction.

Figure 1. Mean receiver operating characteristic curve for the 5 machine learning algorithms. C5: C5 decision tree; LR: logistic regression; MLP: multilayer perceptron; ROC: receiver operating characteristic; RF: random forest; XGB: extreme gradient boosting.

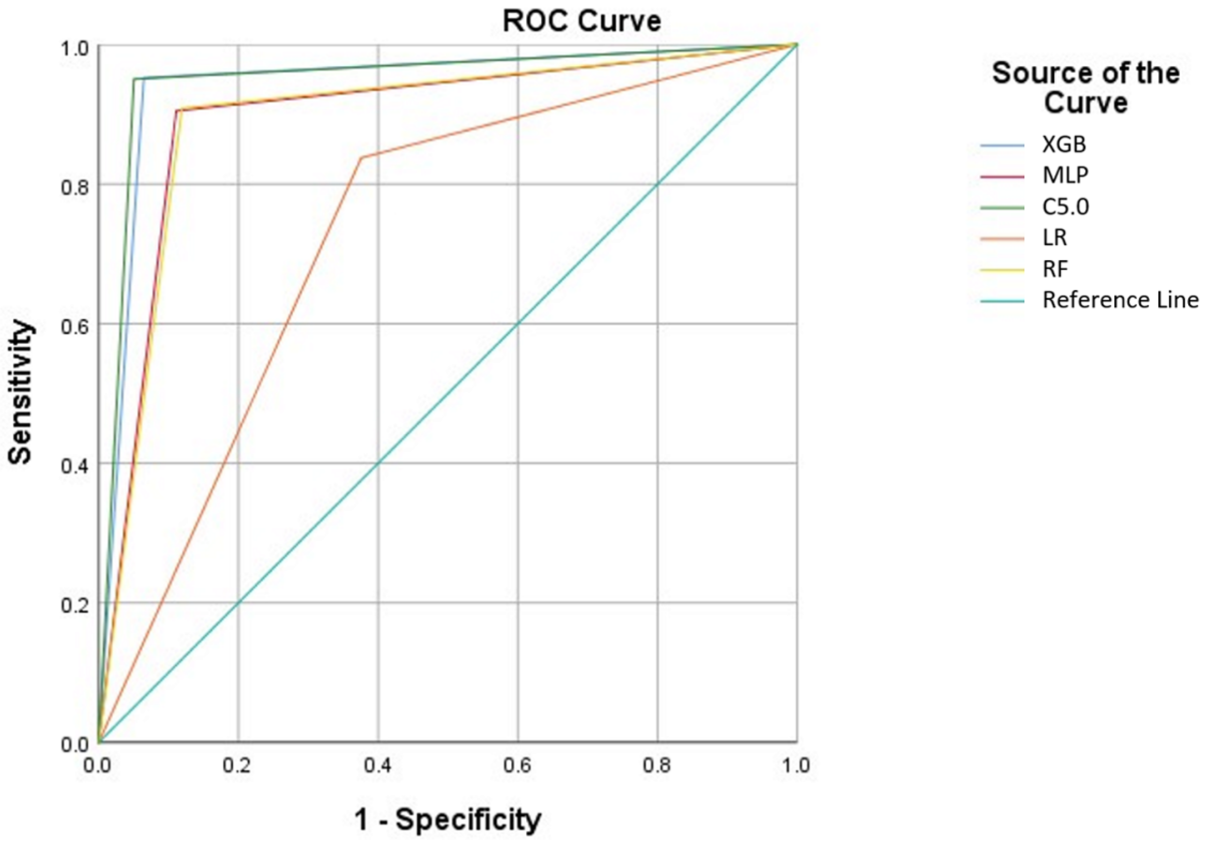


Table 4. Performance of the 5 machine learning models on predicting language understandability of the health texts for international students in tertiary education.

Algorithm	Area under the receiver operating characteristic curve, mean (SD)	Sensitivity, mean (SD)	Specificity, mean (SD)	Accuracy, mean (SD)
Extreme gradient boosting	0.979 (0.006)	0.947 (0.011)	0.944 (0.011)	0.945 (0.01)
Random forest	0.967 (0.033)	0.924 (0.034)	0.885 (0.094)	0.904 (0.064)
Multilayer perceptron	0.946 (0.006)	0.897 (0.006)	0.893 (0.014)	0.895 (0.008)
C5.0 decision tree	0.981 (0.005)	0.95 (0.009)	0.941 (0.023)	0.945 (0.014)
Logistic regression	0.804 (0.002)	0.837 (0.009)	0.627 (0.016)	0.732 (0.004)

The 5 machine learning models (ie, XGBoost, random forest, multilayer perceptron, and C5.0 decision tree) achieved significantly higher AUCs than the linear logistic regression algorithm: XGBoost (*P*<.001), random forest (*P*<.001), C5.0 (*P*<.001), multilayer perceptron (*P*<.001) (Table 4). To be more specific, C5.0 decision tree had a mean score of 0.981 in terms of AUC, followed by XGBoost (0.979), random forest (0.967), neural networks (0.946), and logistic regression (0.804).

XGBoost and C5.0 had significantly higher AUC (Table 5) than multilayer perceptron (XGBoost vs MLP, *P*<.001; MLP vs C5.0, *P*=.001), whereas no significant differences were found between the mean AUCs of XGBoost, random forest, and C5.0 decision tree (XGBoost vs RF, *P*=.44; XGBoost vs C5.0, *P*=.66; RF vs C5.0, *P*=.34). Similarly, all 5 machine learning algorithms had significantly higher mean sensitivity scores than the baseline logistic regression (*P*=.005). C5.0 had the highest mean sensitivity score (0.95) followed by XGBoost (0.947), random forest (0.924), neural networks (0.897), and logistic regression (0.837). XGBoost and C5.0 achieved significantly higher sensitivity scores than multilayer perceptron (XGBoost vs MLP, *P*<.001; MLP vs C5.0, *P*<.001), whereas no significant differences were found between the mean sensitivity scores of XGBoost, C5.0 decision tree, and random forest (XGBoost vs C5.0, *P*=.40; RF vs C5.0, *P*=.15; XGBoost vs RF, *P*=.21). With regards to specificity, that is, the ability of the models to accurately identify health texts classified as easy health education resources, the 5 machine learning models outperformed logistic regression (XGBoost vs LR, *P*<.001; RF vs LR, *P*=.003; MLP vs LR, *P*<.001; C5.0 vs LR, *P*<.001). Again, the mean specificity score of multilayer perceptron was significantly lower than that of XGBoost tree (XGBoost vs MLP, *P*=.001), but not significantly lower than C5.0 (MLP vs C5.0, *P*=.01) and random forest (RF vs MLP, *P*=.86) at the adjusted .005 significance level using Bonferroni correction. Lastly, in terms of overall accuracy, XGBoost and C5.0 achieved the highest mean scores of 0.945, followed by random forest (0.904) and neural networks (0.895). These scores were significantly higher than the mean overall accuracy of logistic regression (0.732) (XGBoost vs LR, *P*<.001; RF vs LR, *P*=.003; MLP vs LR, *P*<.001; C5.0 vs LR, *P*<.001). Again, the differences in the model accuracy were insignificant among XGBoost, C5.0, and random forest (XGBoost vs C5.0, *P*>.99; XGBoost vs RF, *P*=.21; RF vs C5.0, *P*=.17), but significant between the 2 best performing models (XGBoost vs MLP, *P*<.001; MLP vs C5.0, *P*=.002).

Table 5. Results of the pairwise comparison of the model predictive performance by two-tailed *t* test.

Pair number	Comparison	AUC[a] difference		Sensitivity difference		Specificity difference		Accuracy difference	
		Mean (SD)	*P* value	Mean (SD)	*P* value	Mean (SD)	*P* value	Mean (SD)	*P* value
Pair 1	XGB[b] vs RF[c]	0.013 (0.034)	.44	0.023 (0.034)	.21	0.059 (0.089)	.21	0.041 (0.062)	.21
Pair 2	XGB vs MLP[d]	0.034 (0.007)	<.001[e]	0.049 (0.009)	<.001[e]	0.051 (0.013)	.001[e]	0.050 (0.010)	<.001[e]
Pair 3	XGB vs C5.0	–0.001 (0.006)	.66	–0.003 (0.008)	.40	0.003 (0.022)	.76	0.000 (0.015)	>.99
Pair 4	XGB vs LR[f]	0.175 (0.004)	<.001[e]	0.109 (0.006)	<.001[e]	0.317 (0.021)	<.001[e]	0.213 (0.012)	<.001[e]
Pair 5	RF vs MLP	0.021 (0.036)	.27	0.026 (0.037)	.19	–0.008 (0.095)	.86	0.009 (0.066)	.77
Pair 6	RF vs C5.0	–0.014 (0.029)	.34	–0.026 (0.032)	.15	–0.056 (0.076)	.18	–0.041 (0.054)	.17
Pair 7	RF vs LR	0.163 (0.033)	<.001[e]	0.086 (0.034)	.005[e]	0.258 (0.090)	.003[e]	0.172 (0.062)	.003[e]
Pair 8	MLP vs C5.0	–0.035 (0.008)	.001[e]	–0.052 (0.011)	<.001[e]	–0.048 (0.023)	.01	–0.050 (0.016)	.002[e]
Pair 9	MLP vs LR	0.142 (0.005)	<.001[e]	0.060 (0.007)	<.001[e]	0.266 (0.015)	<.001[e]	0.163 (0.008)	<.001[e]
Pair 10	C5.0 vs LR	0.177 (0.004)	<.001[e]	0.112 (0.010)	<.001[e]	0.314 (0.020)	<.001[e]	0.213 (0.014)	<.001[e]

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

[b] XGB: extreme gradient boosting.

[c] RF: random forest.

[d] MLP: multilayer perceptron.

[e] Significant at the adjusted .005 significance level using Bonferroni correction.

[f] LR: logistic regression.

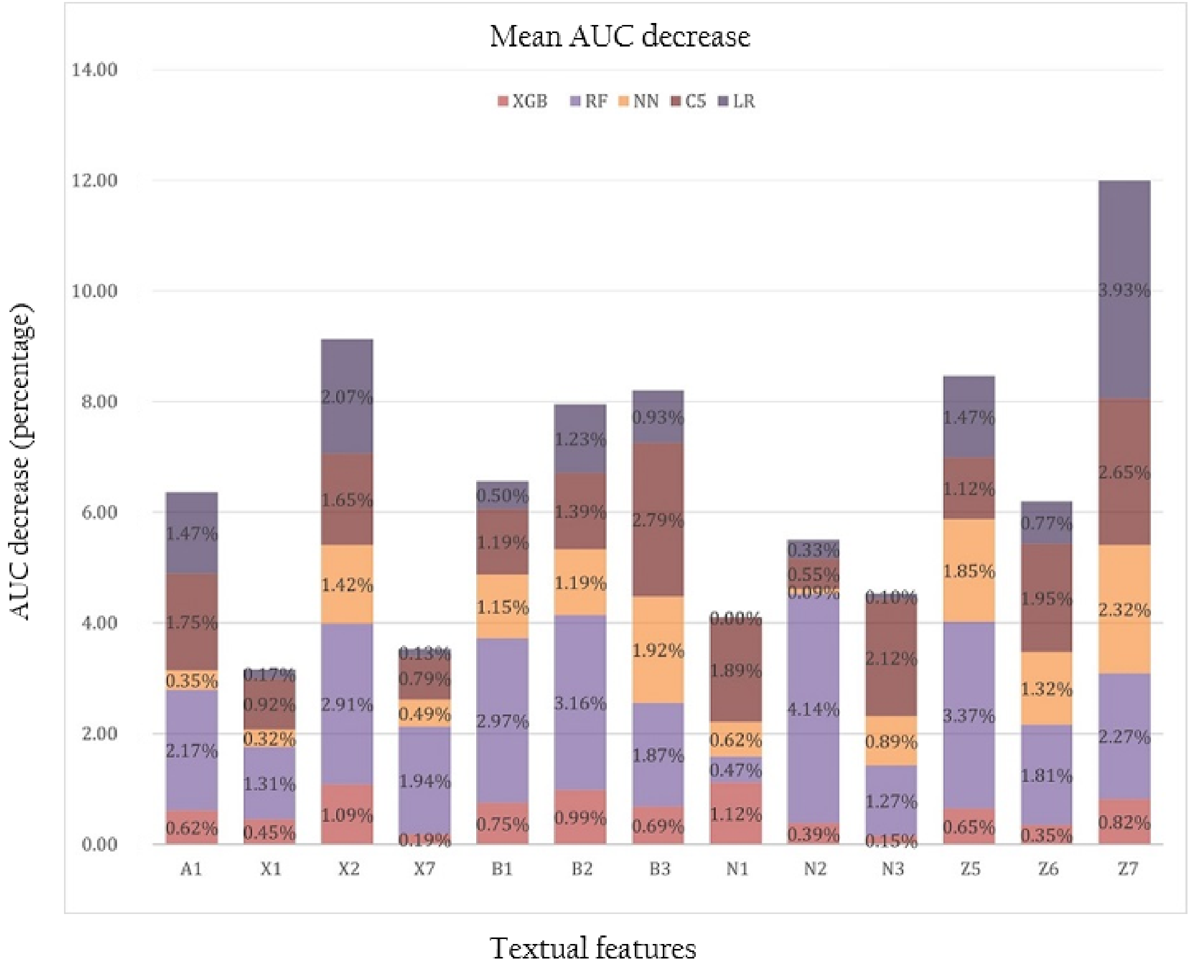
Variable Ranking

To have a deeper understanding of the 5 machine learning algorithms, including the baseline logistic regression, we ranked the impact of the 13 predictor variables on the mean AUCs of the 5 algorithms. This was achieved through the successive permutation of the values of the input linguistic features. To ensure the stability and reliability of the experimental models, five-fold cross-validation was repeated with each permutation exercise. As a result, we obtained the mean decease (in percentage) in the AUCs of the 5 machine learning algorithms as shown in Table 6 and Figure 2.

Table 6. Mean decrease in the area under the receiver operating characteristic curve of the 5 machine learning algorithms.

Feature	Predictor variable	Extreme gradient boosting (%)	Random forest (%)	Deep neural networks (%)	C5.0 decision tree (%)	Logistic regression (%)
General and abstract terms	A1	0.62	2.17	0.35	1.75	1.47
Psychological actions, states, processes	X1	0.45	1.31	0.32	0.92	0.17
Mental actions and processes	X2	1.09	2.91	1.42	1.65	2.07
Wanting, planning, and choosing	X7	0.19	1.94	0.49	0.79	0.13
Anatomy and physiology	B1	0.75	2.97	1.15	1.19	0.50
Health and disease	B2	0.99	3.16	1.19	1.39	1.23
Medicines and medical treatment	B3	0.69	1.87	1.92	2.79	0.93
Numbers	N1	1.12	0.47	0.62	1.89	0.00
Mathematics	N2	0.39	4.14	0.09	0.55	0.33
Measurement	N3	0.15	1.27	0.89	2.12	0.10
Grammatical bin	Z5	0.65	3.37	1.85	1.12	1.47
Negative	Z6	0.35	1.81	1.32	1.95	0.77
If	Z7	0.82	2.27	2.32	2.65	3.93

Figure 2. The impact of different linguistic features on the machine learning algorithms. AUC: area under the receiver operating characteristic curve; C5: C5 decision tree; LR: logistic regression; NN: neural networks; RF: random forest; XGB: extreme gradient boosting. A1: general and abstract terms; X1: psychological actions, states, and processes; X2: mental actions and processes; X7: wanting, planning, and choosing; B1: anatomy and physiology; B2: health and disease; B3: medicines and medical treatment; N1: numbers; N2: mathematics; N3: measurement; Z5: grammatical bin; Z6: negative; Z7: if.



The results showed that for the best performing algorithm, that is the XGBoost tree, linguistic features that had relatively larger impact on the mean model AUC were N1 (numbers, 1.12%), X2 (mental actions and processes, 1.09%), B2 (health and disease, 0.99%), and Z7 (if, conditional, 0.82%). Textual features that had relatively large impact on C5.0 decision tree were B3 (medicines and medical treatment, 2.79%), Z7 (if, conditional, 2.65%), N3 (measurements, 2.12%), Z6 (negative, 1.95%), N1 (numbers, 1.89%), A1 (general and abstract terms, 1.75%), X2 (mental actions and processes, 1.65%), B2 (health and disease, 1.39%), B1 (anatomy and physiology, 1.19%), and Z5 (grammatical bin, 1.12%). For random forest, most linguistic variables had impact on the decrease of the mean AUC larger than 1% and the only exception was N1 (numbers), which reduced the AUC by 0.47%. For the baseline logistic regression, 5 linguistic features reduced the model AUC by more than 1.0%: Z7 (if, conditional, 3.93%), X2 (mental actions/processes, 2.07%), A1 (general and abstract terms, 1.47%), Z5 (grammatical bin, 1.47%), and B2 (health and disease, 1.23%).

AUC Impact of Individual Textual Features

It is worth noting that the AUC impact of these linguistic features on each of the 5 machine learning algorithms did not correlate with their significance to discriminate between easy and difficult health texts. For example, Table 2 shows that there were no statistically significant differences between easy and difficult health texts in their means of B1 (anatomy and physiology, *P*=.12), N1 (numbers, *P*=.66), N2 (mathematics, *P*=.52), and N3 (measurement, *P*=.38). As a result, these features had limited impact on the mean AUC of logistic regression. By contrast, B1 had large impact on the mean AUC of random forest (2.97% AUC decrease), C5.0 (1.19% AUC decrease), and neural networks (1.15% AUC decrease); N1 had large impact on the AUC of XGBoost (1.12% AUC decrease) and C5.0 (1.89% AUC decrease); N2 had large impact on random forest (4.14% AUC decrease) and N3 had large impact on random forest (1.27% AUC decrease) and C5.0 (2.12% AUC decrease). It became clear that XGBoost was the most parsimonious model that achieved the highest mean AUC with less textual features as large predictor variables. The 4 linguistic features with large impact on the AUC of XGBoost, N1, X2, B2, Z7 suggest that 5 evaluation dimensions were critical to the quantitative analysis of the understandability of health education resources.

Impact of the 5 Evaluation Dimensions on the Algorithm Performance (AUCs)

As shown in Table 7, for XGBoost, the evaluation dimension that had the largest impact on the AUC of the algorithm was information evidentness (2.35%), followed by information in logical sequence (1.82%), numeracy skills (1.66%), and the relevance of health information for educational purposes (1.44%). Medical domain knowledge was ranked as the dimension with the least AUC impact (0.99%). Similar patterns were found with random forest. Health information evidentness (8.33%) was ranked as the most impactful dimension, followed by textual logical sequence (7.45%), numeracy skills (5.88%), and the relevance of health information for educational purposes (4.84%). Again, medical knowledge (3.16%) had the smallest impact on the AUC of random forest. C5.0 decision tree differs from XGBoost tree and random forest in that logical sequence (5.72%) replaced information evidentness (5.11%) as the dimension with the largest impact on the C5.0 tree model. Neural networks identified logical sequence (5.49%), relevance to health educational purposes (3.07%), and information evidentness (3.07%) as the 3 evaluation dimensions with the largest impact on the model performance, followed by numeracy skills (1.6%) and domain knowledge (1.19%). Similar to the first 4 machine learning algorithms, logistic regression also identified logical sequence (6.17%) as the most impactful dimension on the model performance, followed by information evidentness (3.84%), educational relevance (1.43%), domain knowledge (1.23%), and numeracy skills (0.43%). It is useful to note that for all models, logical sequence, information evidentness, and educational purpose relevance were identified as the most important dimensions with the largest impact on the model prediction accuracy, whereas medical domain knowledge was ranked as the dimension with the least impact on the algorithm performance.

Table 7. Impact of the different dimensions on the area under the curves of the algorithms.

Evaluation dimensions	Understandability	Predictor variable	Extreme gradient boosting (%)	Random forest (%)	Deep neural networks (%)	C5.0 decision tree (%)	Logistic regression (%)
Dimension 1	Information evidentness	A1, X1, X2, X7	2.35	8.33	2.58	5.11	3.84
Dimension 2	Relevance to education purpose	B1, B3	1.44	4.84	3.07	3.98	1.43
Dimension 3	Domain knowledge	B2	0.99	3.16	1.19	1.39	1.23
Dimension 4	Numeracy demand	N1, N2, N3	1.66	5.88	1.60	4.56	0.43
Dimension 5	Logical sequence	Z5, Z6, Z7	1.82	7.45	5.49	5.72	6.17

Discussion

Principal Findings

The study of the readability of health educational resources has, for long, relied on medical readability calculators among which the Flesch Reading Ease Score [34], Gunning Fog [35], Flesch-Kincaid Grade Level Readability [34], Coleman-Liau Index [36], Simple Measure of Gobbledygook Index [37], Automated Readability Index [38], and Lensear Write Formula [39] are some of the most influential and widely used ones. However, this medical formula–based approach to linguistic readability evaluation, despite being convenient and fast, has known limitations, including interformula inconsistency and reported lack of flexibility and adaptability with populations with diverse language, cultural backgrounds, as well as cognitive abilities. Furthermore, these evaluation tools were originally designed for readers from native English-speaking backgrounds, assuming the health educators who developed the health resources and the target readers have similar knowledge and understanding of the general English vocabulary, logical organization of health materials, and communication of the intentions and purposes of health educational materials. These assumptions, which underlined the design of existing medical readability formula, were increasingly challenged by applications of these tools with diverse populations and communities with limited exposure to the health care systems of English-speaking countries [40-42]. The limitation of the existing medical readability tools also reflects in their exclusive focus on the morphological, syntactic complexity, using low-frequency polysyllabic words, medical terminology, and sentence lengths as the main textual complexity measures.

The more recent patient-oriented health resource evaluation guidelines such as PEMAT has greatly enriched the dimensions of readability evaluation, expanding the evaluation criteria from medical domain knowledge (using familiar, everyday language) to encompass dimensions such as health information relevance, purposefulness to the target readers (information classified as distractor or key information), numeracy demand, and the logical sequence of health texts. Despite the wide adoption of these more comprehensive and user-adaptive evaluation guidelines, no quantitative tools have been developed to implement the multidimensional evaluation in a cost-effective, instant manner. This represents a critical research gap in current health material evaluation, as there are growing demands from both clinical and research settings for automated evaluation tools of the understandability of written health materials. Advances in computational methods such as machine learning algorithms can help address the increasing gap between the practical needs for more cost-effective, integrated quantitative tools that are able to deal with health texts in large quantities and the known limitations of medical readability formulas and expert-led evaluation guidelines, which are slow and time-consuming to implement and update.

Our study developed the first quantitative tool for the evaluation of written health education materials based on the PEMAT guidelines. We developed and compared 5 machine learning algorithms by using logistic regression as the baseline model. The results showed that all 5 models (XGBoost, C5.0, random forest, multilayer perceptron) outperformed logistic regression in terms of AUC, sensitivity, specificity, and overall accuracy. We found that in the evaluation of health information understandability, information evidentness, educational relevance, and logical sequence were ranked consistently more important than numeracy skills and medical domain knowledge. This ranking of the importance of these evaluation dimensions may be explained by the demographical profiles of the target readership: international students in tertiary education with adequate English skills (International English Language Testing System mean score 6.5) and high health literacy (mean score 16.5 in the Short Assessment of Health Literacy-English test). These results challenged the traditional view that lack of medical knowledge and numeracy skills caused the lack of health information understandability. Improving the writing style and health information organization can significantly improve the understandability of health information for non-English speakers, especially for those of higher educational attainment and health literacy levels and with distinct language and cultural backgrounds.

Limitations and Future Research

The textual linguistic features used in the model development were limited. In future research, we will increase the features to be studied in the evaluation of health material understandability, by adding, for example, syntactic and morphological features of texts. The underlying evaluation framework we used was PEMAT. There are, however, other studies that explored health information accessibility from cognitive and psychological experiments. These studies may help expand the current scope of PEMAT, which is intended for the evaluation of written health resources for readers with average cognitive skills, rather than those with cognitive impairments caused by physical or mental health issues. The new quantitative tools have the potential to be further adapted for different readerships as well as written health materials in languages other than English.

Conclusions

An important contribution of this paper lies in its efforts to bridge the gap between the 2 distinct approaches to health information evaluation. This was achieved via the translation of clinically developed patient health education materials assessment guidelines to quantitative evaluation models, that is, machine learning algorithms by using a limited number of semantic features to accurately predict the readability (binary outcome) of health educational resources for international students in tertiary education with adequate English proficiency and health literacy but distinct language and cultural backgrounds.

Authors' Contributions

MJ contributed to the research design, statistical design, analysis, and machine learning algorithms. YL contributed to the research design, health information collection, data collection, and analysis. MZ and ZL collected the health information. BZ, XL, YL, and YZ contributed to data annotation and collection.

Conflicts of Interest

None declared.

References

1. World Health Organization. WHO strategic communications framework for effective communications. World Health Organization. 2017. URL: https://www.who.int/mediacentre/communication-framework.pdf [accessed 2021-04-14]
2. Beaunoyer E, Arsenault M, Lomanowska AM, Guitton MJ. Understanding online health information: Evaluation, tools, and strategies. Patient Educ Couns 2017 Feb;100(2):183-189. [doi: 10.1016/j.pec.2016.08.028] [Medline: 27595436]
3. Moult B, Franck LS, Brady H. Ensuring quality information for patients: development and preliminary validation of a new instrument to improve the quality of written health care information. Health Expect 2004 Jun;7(2):165-175 [FREE Full text] [doi: 10.1111/j.1369-7625.2004.00273.x] [Medline: 15117391]
4. Birru MS, Monaco VM, Charles L, Drew H, Njie V, Bierria T, et al. Internet usage by low-literacy adults seeking health information: an observational analysis. J Med Internet Res 2004 Sep 03;6(3):e25 [FREE Full text] [doi: 10.2196/jmir.6.3.e25] [Medline: 15471751]
5. Helitzer D, Hollis C, Cotner J, Oestreicher N. Health literacy demands of written health information materials: an assessment of cervical cancer prevention materials. Cancer Control 2009 Jan;16(1):70-78 [FREE Full text] [doi: 10.1177/107327480901600111] [Medline: 19078933]
6. Blake SC, McMorris K, Jacobson KL, Gazmararian JA, Kripalani S. A qualitative evaluation of a health literacy intervention to improve medication adherence for underserved pharmacy patients. J Health Care Poor Underserved 2010 May;21(2):559-567. [doi: 10.1353/hpu.0.0283] [Medline: 20453356]
7. Sidhu MS, Gale NK, Gill P, Marshall T, Jolly K. A critique of the design, implementation, and delivery of a culturally-tailored self-management education intervention: a qualitative evaluation. BMC Health Serv Res 2015 Feb 07;15:54 [FREE Full text] [doi: 10.1186/s12913-015-0712-8] [Medline: 25890256]
8. Ammenwerth E, Brender J, Nykänen P, Prokosch H, Rigby M, Talmon J, HIS-EVAL Workshop Participants. Visions and strategies to improve evaluation of health information systems. Reflections and lessons based on the HIS-EVAL workshop in Innsbruck. Int J Med Inform 2004 Jun 30;73(6):479-491. [doi: 10.1016/j.ijmedinf.2004.04.004] [Medline: 15171977]
9. Kim H, Park S, Bozeman I. Online health information search and evaluation: observations and semi-structured interviews with college students and maternal health experts. Health Info Libr J 2011 Sep;28(3):188-199 [FREE Full text] [doi: 10.1111/j.1471-1842.2011.00948.x] [Medline: 21831218]
10. Balyan R, Crossley SA, Brown W, Karter AJ, McNamara DS, Liu JY, et al. Using natural language processing and machine learning to classify health literacy from secure messages: The ECLIPPSE study. PLoS One 2019;14(2):e0212488 [FREE Full text] [doi: 10.1371/journal.pone.0212488] [Medline: 30794616]
11. D'Alessandro DM, Kingsley P, Johnson-West J. The readability of pediatric patient education materials on the World Wide Web. Arch Pediatr Adolesc Med 2001 Jul;155(7):807-812. [doi: 10.1001/archpedi.155.7.807] [Medline: 11434848]
12. Taylor HE, Bramley DEP. An Analysis of the Readability of Patient Information and Consent forms used in Research Studies in Anaesthesia in Australia and New Zealand. Anaesthesia and Intensive Care 2019 Jan 16;40(6):995-998. [doi: 10.1177/0310057x1204000610]
13. Crossley SA, Balyan R, Liu J, Karter AJ, McNamara D, Schillinger D. Developing and Testing Automatic Models of Patient Communicative Health Literacy Using Linguistic Features: Findings from the ECLIPPSE study. Health Commun 2020 Mar 02:1-11. [doi: 10.1080/10410236.2020.1731781] [Medline: 32114833]
14. Meade CD, Smith CF. Readability formulas: Cautions and criteria. Patient Education and Counseling 1991 Apr;17(2):153-158. [doi: 10.1016/0738-3991(91)90017-y]
15. Mumford M. A descriptive study of the readability of patient information leaflets designed by nurses. J Adv Nurs 1997 Nov;26(5):985-991. [doi: 10.1046/j.1365-2648.1997.00455.x] [Medline: 9372404]
16. Friedman DB, Hoffman-Goetz L. A systematic review of readability and comprehension instruments used for print and web-based cancer information. Health Educ Behav 2006 Jun;33(3):352-373. [doi: 10.1177/1090198105277329] [Medline: 16699125]
17. Karačić J, Dondio P, Buljan I, Hren D, Marušić A. Languages for different health information readers: multitrait-multimethod content analysis of Cochrane systematic reviews textual summary formats. BMC Med Res Methodol 2019 Apr 05;19(1):75 [FREE Full text] [doi: 10.1186/s12874-019-0716-x] [Medline: 30953453]
18. Cline R, Haynes K. Consumer health information seeking on the Internet: the state of the art. Health Educ Res 2001 Dec;16(6):671-692. [doi: 10.1093/her/16.6.671] [Medline: 11780707]
19. Collins-Thompson K. Computational assessment of text readability. International Journal of Applied Linguistics. 2015 Jan 23. URL: http://www-personal.umich.edu/~kevynct/pubs/ITL-readability-invited-article-v10-camera.pdf [accessed 2021-04-05]
20. Benjamin RG. Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty. Educ Psychol Rev 2011 Oct 4;24(1):63-88. [doi: 10.1007/s10648-011-9181-8]
21. Feng L, Jansche M, Huenerfauth M, Elhadad N. A comparison of features for automatic readability assessment. URL: https://www.aclweb.org/anthology/C10-2032.pdf [accessed 2021-04-05]
22. Si L, Callan J. A statistical model for scientific readability. 2001 Presented at: Proceedings of the tenth international conference on Information and knowledge management; October; NY, USA. [doi: 10.1145/502585.502695]

23. Brown M, Bussell J. Medication adherence: WHO cares? Mayo Clin Proc 2011 Apr;86(4):304-314. [doi: 10.4065/mcp.2010.0575] [Medline: 21389250]
24. Ley P. Satisfaction, compliance and communication. Br J Clin Psychol 1982 Nov;21 (Pt 4):241-254. [doi: 10.1111/j.2044-8260.1982.tb00562.x] [Medline: 7171877]
25. Winker MA, Flanagin A, Chi-Lum B, White J, Andrews K, Kennett RL, et al. Guidelines for medical and health information sites on the internet: principles governing AMA web sites. American Medical Association. JAMA 2000;283(12):1600-1606. [doi: 10.1001/jama.283.12.1600] [Medline: 10735398]
26. Lutfey KE, Wishner WJ. Beyond "compliance" is "adherence". Improving the prospect of diabetes care. Diabetes Care 1999 Apr;22(4):635-639 [FREE Full text] [doi: 10.2337/diacare.22.4.635] [Medline: 10189544]
27. Allen FCL, Cole JB. Foreign student syndrome: fact or fable? J Am Coll Health 1987 Jan;35(4):182-186. [doi: 10.1080/07448481.1987.9938986] [Medline: 3819191]
28. Cole JB, Allen FC, Green JS. Survey of health problems of overseas students. Social Science & Medicine. Part A: Medical Psychology & Medical Sociology 1980 Dec;14(6):627-631. [doi: 10.1016/S0271-7123(80)80072-1]
29. Burns RB. Study and Stress among First Year Overseas Students in an Australian University. Higher Education Research & Development 1991 Jan;10(1):61-77. [doi: 10.1080/0729436910100106]
30. Lee SD, Bender DE, Ruiz RE, Cho YI. Development of an easy-to-use Spanish Health Literacy test. Health Serv Res 2006 Aug;41(4 Pt 1):1392-1412 [FREE Full text] [doi: 10.1111/j.1475-6773.2006.00532.x] [Medline: 16899014]
31. Lee S, Stucky B, Lee J, Rozier R, Bender D. Short Assessment of Health Literacy-Spanish and English: a comparable test of health literacy for Spanish and English speakers. Health Serv Res 2010 Aug;45(4):1105-1120 [FREE Full text] [doi: 10.1111/j.1475-6773.2010.01119.x] [Medline: 20500222]
32. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014 Sep;96(3):395-403 [FREE Full text] [doi: 10.1016/j.pec.2014.05.027] [Medline: 24973195]
33. Rayson P, Archer D, Piao S, McEnery A. The UCREL semantic analysis system. 2004 Presented at: Proceedings of the beyond named entity recognition semantic labelling for NLP tasks workshop; May; Lisbon, Portugal p. 7-12 URL: https://eprints.lancs.ac.uk/id/eprint/1783/
34. Flesch R. A new readability yardstick. J Appl Psychol 1948 Jun;32(3):221-233. [doi: 10.1037/h0057532] [Medline: 18867058]
35. Gunning R. Readability yardsticks. In: The Technique of Clear Writing. New York: McGraw-Hill; 1968.
36. Coleman M, Liau TL. A computer readability formula designed for machine scoring. Journal of Applied Psychology 1975;60(2):283-284. [doi: 10.1037/h0076540]
37. Mc LG. SMOG grading-a new readability formula. 1969. URL: https://ogg.osu.edu/media/documents/health_lit/WRRSMOG_Readability_Formula_G._Harry_McLaughlin__1969_.pdf [accessed 2021-04-14]
38. Senter R, Smith E. Automated readability index. Defense Technical Information Center. 1967. URL: https://apps.dtic.mil/sti/citations/AD0667273 [accessed 2021-04-05]
39. O'Hayre J. Gobbledygook has gotta go. 1966. URL: https://www.governmentattic.org/15docs/Gobbledygook_Has_Gotta_Go_1966.pdf [accessed 2021-04-14]
40. Kim W, Kim I, Baltimore K, Imtiaz AS, Bhattacharya BS, Lin L. Simple contents and good readability: Improving health literacy for LEP populations. Int J Med Inform 2020 Sep;141:104230. [doi: 10.1016/j.ijmedinf.2020.104230] [Medline: 32688291]
41. Schur C, Lucado J, Feldman J. Local public health capacities to address the needs of culturally and linguistically diverse populations. Journal of Public Health Management and Practice 2011;17(2):177-186. [doi: 10.1097/phh.0b013e3181fb0037]
42. Andrus MR, Roth MT. Health literacy: a review. Pharmacotherapy 2002 Mar;22(3):282-302. [doi: 10.1592/phco.22.5.282.33191] [Medline: 11898888]

Abbreviations

AUC: area under the receiver operating characteristic curve
PEMAT: Patient Education Materials Assessment Tool
USAS: UCREL English Semantic Analysis System
XGBoost: extreme gradient boosting

Edited by T Hao; submitted 04.03.21; peer-reviewed by A Pagano, S Laviosa, D Li; comments to author 16.03.21; revised version received 17.03.21; accepted 30.03.21; published 06.05.21.

Please cite as:
Ji M, Liu Y, Zhao M, Lyu Z, Zhang B, Luo X, Li Y, Zhong Y
Use of Machine Learning Algorithms to Predict the Understandability of Health Education Materials: Development and Evaluation Study
JMIR Med Inform 2021;9(5):e28413
URL: https://medinform.jmir.org/2021/5/e28413
doi:10.2196/28413
PMID:33955834

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org