JMIR Medical Informatics

Impact Factor (2023): 3.1
Volume 8 (2020), Issue 12 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Original Papers

Original Paper

Development of Phenotyping Algorithms for the Identification of Organ Transplant Recipients: Cohort Study

Lee Wheless[1], MD, MSCR, PhD; Laura Baker[1], BS; LaVar Edwards[1], MS; Nimay Anand[2], BS; Kelly Birdwell[3], MD, MSCI; Allison Hanlon[1], MD, PhD; Mary-Margaret Chren[1], MD

[1]Department of Dermatology, Vanderbilt University Medical Center, Nashville, TN, United States

[2]Meharry Medical College, Nashville, TN, United States

[3]Division of Nephrology and Hypertension, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States

Corresponding Author:
Lee Wheless, MD, MSCR, PhD
Department of Dermatology
Vanderbilt University Medical Center
719 Thompson Lane
Suite 26300
Nashville, TN, 37204
United States
Phone: 1 6153226485
Email: lee.e.wheless@vumc.org

Abstract

Background: Studies involving organ transplant recipients (OTRs) are often limited to the variables collected in the national Scientific Registry of Transplant Recipients database. Electronic health records contain additional variables that can augment this data source if OTRs can be identified accurately.

Objective: The aim of this study was to develop phenotyping algorithms to identify OTRs from electronic health records.

Methods: We used Vanderbilt's deidentified version of its electronic health record database, which contains nearly 3 million subjects, to develop algorithms to identify OTRs. We identified all 19,817 individuals with at least one International Classification of Diseases (ICD) or Current Procedural Terminology (CPT) code for organ transplantation. We performed a chart review on 1350 randomly selected individuals to determine the transplant status. We constructed machine learning models to calculate positive predictive values and sensitivity for combinations of codes by using classification and regression trees, random forest, and extreme gradient boosting algorithms.

Results: Of the 1350 reviewed patient charts, 827 were organ transplant recipients while 511 had no record of a transplant, and 12 were equivocal. Most patients with only 1 or 2 transplant codes did not have a transplant. The most common reasons for being labeled a nontransplant patient were the lack of data (229/511, 44.8%) or the patient being evaluated for an organ transplant (174/511, 34.1%). All 3 machine learning algorithms identified OTRs with overall >90% positive predictive value and >88% sensitivity.

Conclusions: Electronic health records linked to biobanks are increasingly used to conduct large-scale studies but have not been well-utilized in organ transplantation research. We present rigorously evaluated methods for phenotyping OTRs from electronic health records that will enable the use of the full spectrum of clinical data in transplant research. Using several different machine learning algorithms, we were able to identify transplant cases with high accuracy by using only ICD and CPT codes.

(JMIR Med Inform 2020;8(12):e18001) doi:10.2196/18001

KEYWORDS

phenotyping; electronic health record; organ transplant recipients

Introduction

The Scientific Registry for Transplant Recipients (SRTR) is an outstanding resource for studies of organ transplant recipients (OTRs). The SRTR has incomplete data on important variables such as cancers in transplant patients and lacks a common data model [1-3]. Linking records to cancer registries has greatly aided in the collection of these data, but not all outcomes can

be measured in this way [4]. Moreover, the regulations regarding linking these identified data sets to DNA biobanks can be burdensome and limit the scale of genetic studies that can be conducted in OTRs. To address these limitations, other resources that contain a more robust record of patients' health, such as the electronic health record (EHR), can be used [5]. The use of different types of data contained in the EHR to phenotype disease states has gained broad acceptance [6-8]. Most studies seeking broader data have attempted to link EHR data to the SRTR [9-11]. This approach can be problematic because to protect patient privacy according to the Health Insurance Portability and Accountability Act, the linkage is done by the SRTR management team, with new identifiers returned to the investigator. These new identifiers preclude linkage back for updating or correcting records or linking to deidentified genetic databases.

To avoid this issue, several studies have used the presence of an International Classification of Diseases (ICD)-9 or ICD-10 code or Current Procedural Terminology (CPT) code for transplantation to identify transplant patients, although this practice is known to have poor performance [9-11]. ICD codes are used as a means of providing distinct diagnoses for billing purposes. ICD version 9 was first used in 1979 and it ran until October 1, 2014 in the United States, at which time ICD-10 was adopted. Patients whose records span this timepoint thus can contain both ICD-9 and ICD-10 codes in their records, whereas patients seen only prior to then would have exclusively ICD-9 codes. CPT codes designate specific surgeries and procedures. A thorough investigation of the accuracy of using ICD and CPT codes to phenotype OTRs has not been performed nor have formal phenotyping algorithms for identifying transplant patients from the EHR been developed. We therefore conducted this study to develop rigorously evaluated phenotyping algorithms for the identification of transplant patients from EHRs.

Methods

Cohort Assembly

This study used deidentified patient-level data and was designated as an exempt nonhuman subjects research study by the institutional review board at the Vanderbilt University Medical Center (VUMC). We identified all possible OTRs from the Synthetic Derivative [12]. The Synthetic Derivative contains over 2.9 million subjects with deidentified clinical data from the EHR collected longitudinally over several decades since VUMC began using an EHR. The Synthetic Derivative is linked to a large DNA biobank called BioVU [12]. Similar to the entire patient population seen at VUMC, patients are predominantly Caucasian, and there are approximately equal numbers of males and females. The Synthetic Derivative includes all information available in the EHR, incorporating diagnostic codes (ICD-9 and ICD-10), CPT codes, demographics, text from inpatient and outpatient notes (including both subspecialty and primary care), laboratory values, radiology reports, and medication orders. However, records scanned into the EHR are not available in the Synthetic Derivative. Users can perform text-based searches of the entire clinical record within seconds to increase the efficiency and accuracy of data extraction. To identify possible OTRs within the Synthetic Derivative, we used ICD-9 and ICD-10 codes as well as CPT codes specific to each organ (Table 1). We excluded codes for bone, cornea, and skin transplants, as these are uncommon. Although bone marrow and stem cell transplants are not included in SRTR, we included these, given the large number of transplants performed every year and the need to be able to identify these patients.

We randomly selected 1350 patients for chart review to confirm organ transplant status and to serve as training and testing sets (Figure 1). A preliminary analysis of the first 750 charts showed difficulty in the models correctly identifying OTRs with a low number of codes. Overall, there was a bimodal distribution of code count frequencies, with high numbers of patients having only 1 or 2 and >50% having 10 or more codes (Figure 2). Therefore, we reviewed an additional 500 charts with oversampling of those with 1 or 2 codes. There were only 31 lung transplant cases included in the initial sample; therefore, we reviewed an additional 100 charts that had at least one code for lung transplant to increase the sample size. Chart review was performed by 3 authors (LW, LXW, NA) with 20% overlap to determine interrater reliability. Disagreements were settled by reviewers examining the record in question together to make a final determination. The time of possible transplant was defined as the date of the first CPT code for transplant or the earliest transplant code in the chart. Transplant patients were defined as those with any definitive evidence of having a transplant (eg, transplant procedure note, transplant biopsy pathology report, documentation in the chart of having a transplant). Equivocal cases were defined as those with an absence of definitive evidence but with factors potentially related to transplantation (eg, subsequent immunosuppressant use, laboratories measuring tacrolimus levels, multiple cytomegalovirus titers). Patients without documentation of a transplant were defined as those with definitive evidence of having not received a transplant (eg, organ donation, denied listing for transplantation). Patients whose charts contained only ICD and CPT codes but lacking any documentation of notes, pathology records, radiology records, laboratory records, or medications were classified as not having evidence of a transplant unless there were multiple transplant codes at different time points.

Table 1. List of the International Classification of Diseases and Current Procedural Terminology codes used to identify possible organ transplant recipients from the electronic health record.

Transplanted organ	ICD[a]-9 codes	ICD-10 codes	Current Procedural Terminology codes
Heart	V42.1, 996.83, 37.51	Z94.1, Z94.3, T86.2, T86.3, 02YA0Z[b]	33935, 33945
Lung	V42.6, 996.84	Z94.2, Z94.3, T86.3, T86.81, 0BY[b]	32851, 32852, 32853, 32854
Kidney	V42.0, 996.81	Z94.0, T86.1, 0TY[b]	50340, 50370, 50380, 50360, 50365
Liver	V42.7, 996.82	Z94.4, T86.4, 0FY00[b]	47135, 47136
Bone marrow or stem cell	V42.81, V42.82, 996.85, 996.88, 41.0, 41.00, 41.01, 41.02, 41.03, 41.04, 41.05, 41.06, 41.07, 41.08, 41.09	Z94.81, Z94.84, T86.0, T86.5, 30230A[b], 30230G[b], 30230X[b], 30230Y[b], 30233A[b], 30233G[b], 30233X[b], 30233Y[b], 30240A[b], 30240G[b], 30240X[b], 30240Y[b], 30243A[b], 30243G[b], 30243X[b], 30243Y[b], 30250G[b], 30250X[b], 30250Y[b], 30253G[b], 30253X[b], 30253Y[b], 30260G[b], 30260X[b], 30260Y[b], 30263G[b], 30263X[b], 30263Y[b]	38242, 38240, 38241, 38243
Pancreas, intestine, or other	V42.83, V42.84, V42.89, V42.8, V42.83, V42.9, 996.86, 996.87, 996.89, 996.80	Z94.82, Z94.83, Z94.89, Z94.9, T86.85, T86.89, T86.90, T86.91, T86.92, T86.93, T86.99, 0FYG0Z[b]	48554, 48556

[a]International Classification of Diseases.

[b]Means all values under this subheading, eg, "0FYG0Z*" includes 0FYG0Z0, 0FYG0Z1, and 0FYG0Z2.

Figure 1. Selection of patients. From the full electronic health record, we identified 19,817 individuals with at least one transplant code, and from these, we selected a random sample of 1350 individuals for chart review and model building. EHR: electronic health record.

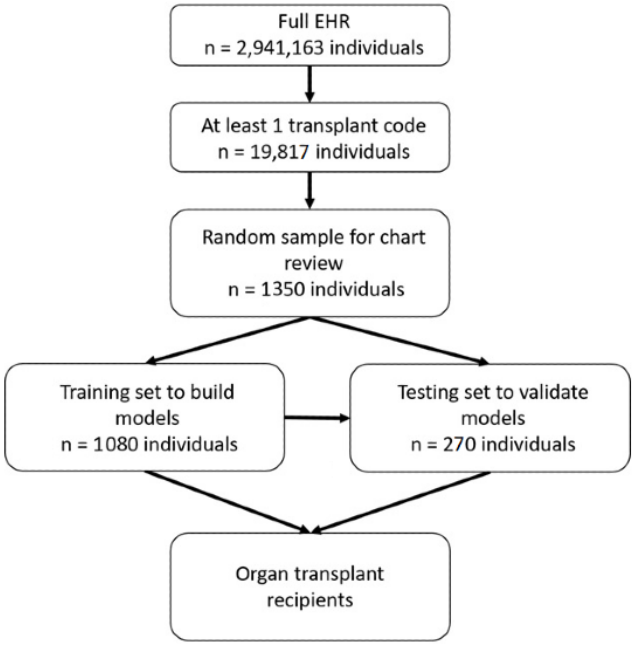
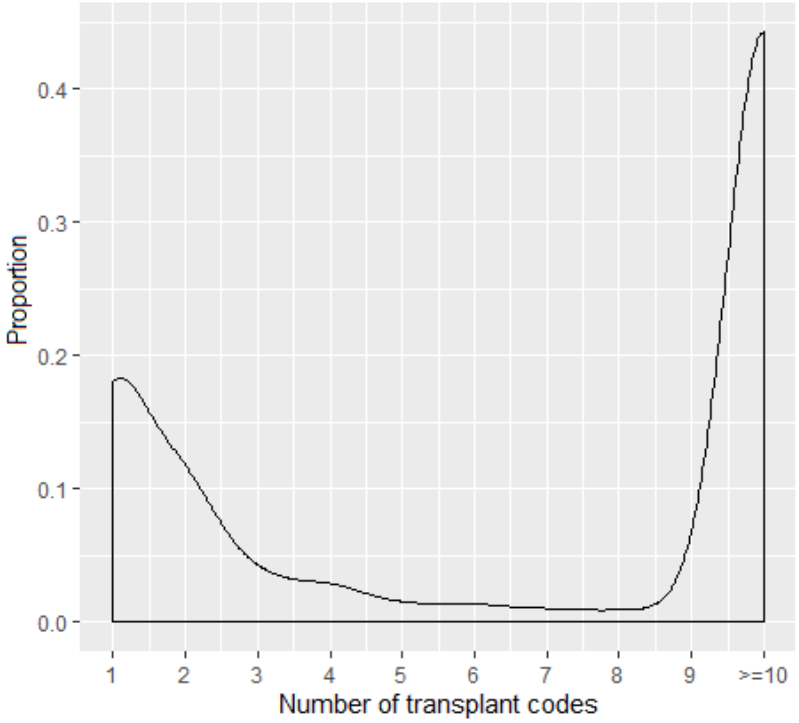


Figure 2. Frequencies of total transplant code counts among those 19,817 individuals with at least one transplant code.



Algorithm Development

We split the population of 1350 into a training set of 1080 individuals (80.0%) and a testing set of the remaining 270 individuals (20.0%). We calculated the positive predictive value (PPV), sensitivity, and F-score at each sequential cut point from each sequential cut point (>1, >2, >3…>10) of the total ICD-9, ICD-10, and CPT transplant codes, labeling those below the cut point as nontransplant patients and those above the cut point as transplant patients. We selected the cut point with the highest F-score in the training set and calculated these values in the test set by using the same cut point. We considered several different models, starting with classification and regression trees (CART), which is perhaps the most approachable to clinicians without any formal training in bioinformatics and then expanding to ensemble methods of random forest (RF) and extreme gradient boosting (XGB). The variables used in the models included age at transplant, race, gender, year of transplant, duration of follow-up, vital status, the codes listed in Table 1, total number of transplant codes, total number of transplant status codes, total number of transplant procedure codes, total number of transplant complications codes, and total number of transplant aftercare codes. Machine learning models were constructed using the training set with 5-fold cross validation and were tuned using the caret package in R 3.5.1 [13,14]. The final tuning parameters for each model are presented in Table S1 of Multimedia Appendix 1. The rpart package was used for CART models [15], the ranger package was used for RF models [16], and the xgboost package was used for XGB models by using method = “xgbTree” in the caret framework [17]. Sensitivity was defined as the number of those predicted as having a transplant divided by the total number of transplant patients. PPV was the number of transplant patients correctly predicted to have a transplant divided by the total number of patients predicted to be transplant patients. Sensitivity and PPV were calculated overall and for each organ type. All models were compared using the F-score, which is calculated as 2*(sensitivity*PPV)/(sensitivity + PPV). An F-score of 1.0 represents perfect classification. Because all charts were selected based on the presence of a transplant code, specificity could not be calculated.

Alternative Search Strategies

Preliminary models suggested difficulty in discriminating between transplant recipients and nontransplant recipients with fewer than 4 transplant codes. We therefore considered the addition of medication and laboratory data. However, among these subjects with few codes, we found that nearly all of them had data for only ICD and CPT codes and not medications; therefore, this strategy was abandoned. We also considered the addition of natural language processing (NLP) methods to augment the search algorithms. While this 2-step process has shown better performance than using codes alone, we observed that the model had excellent performance in patients with unstructured data sources and poor performance in those without unstructured data [18]. As such, the addition of NLP would have improved our classification only minimally, while greatly increasing the complexity of the algorithm. All the algorithms were therefore constructed using the structured data only.

Results

Cohort Assembly

Among patients in the Synthetic Derivative with at least one transplant code, there were 7751 potential renal transplant patients, 3240 potential cardiac transplant patients, 1506 potential lung transplant patients, 3648 potential liver transplant patients, 6401 potential stem cell or bone marrow transplant patients, and 3845 patients potentially with a transplanted

pancreas, small intestine, or other organs besides skin, bone, or eye. Accounting for patients with codes for multiple transplanted organs, there were 19,817 unique individuals.

The mean number of codes per individual was 52.6 and the median count was 6. Many of the individuals had only 1 (4439/19,817, 22.3%) or 2 (2243/19,817, 11.3%) transplant codes (Figure 2). A chart review of 1350 subjects revealed 827 (61.3%) transplant patients, 12 (0.9%) equivocal cases, and 511 (37.9%) patients without documentation of a transplant. Individuals with a greater number of codes were more likely to be OTRs (Table 2). Interrater reliability was extremely high (247/250, 98.8% concordance), and all 3 discrepancies involved patients being labeled as OTRs versus equivocal. The most common reasons for being labeled as not having documentation of a transplant were the lack of adequate data (229/511, 44.8%) or the patient currently or formerly being evaluated for an organ transplant (174/511, 34.1%). Other reasons included coding errors identified during the chart review, such as the patient receiving blood products or tagged red blood cell scans. In preliminary analyses, we considered models excluding the 12 equivocal cases or categorizing them as OTRs or non-OTRs. There were no material differences among the models; therefore, these 12 were labeled as cases in the final models presented.

Table 2. Frequencies, positive predictive value, sensitivity, and F-score by code counts of organ transplant recipients and nonorgan transplant recipients.

Transplant codes	1	2	3	4	5	6	7	8	9	10
Non-OTR[a], n	269	173	27	12	8	8	2	1	2	9
OTR, n	51	95	24	21	8	7	6	4	16	607
PPV[b]	0.621	0.765	0.909	0.941	0.956	0.967	0.978	0.981	0.983	0.985
Sensitivity	1.000	0.939	0.879	0.797	0.772	0.763	0.754	0.747	0.743	0.723
F-score	0.767	0.843	0.894	0.863	0.854	0.853	0.852	0.848	0.846	0.834

[a]OTR: organ transplant recipient.

[b]PPV: positive predictive value.

Models for Overall Transplant Status

Using 3 or more codes as the cut point for calling a patient a transplant recipient had the highest F-score (Table 2). The sensitivity and PPV of the code counts and the CART, RF, and XGB models for identifying OTRs are shown in Table 3. CART, RF, and XGB all performed comparably, with RF having the highest F-score in the testing set. Applying the overall RF model to the full study population yielded a final sample size of 13,445 OTRs. For comparison, VUMC has performed 7671 solid organ transplants between January 1, 1988 and February 28, 2019, and 1323 bone marrow and stem cell transplants from 2015 to 2018 [19,20].

Table 3. Positive predictive value, sensitivity, and F-scores for each model to identify individuals with any organ transplant in the training and testing sets.

Model	Training set			Testing set		
	PPV[a]	Sensitivity	F-score	PPV	Sensitivity	F-score
>3 codes	0.909	0.876	0.892	0.911	0.911	0.911
CART[b]	0.911	0.872	0.891	0.903	0.892	0.898
RF[c]	0.909	0.887	0.898	0.909	0.909	0.909
XGB[d]	0.925	0.882	0.903	0.846	0.892	0.868

[a]PPV: positive predictive value.

[b]CART: classification and regression tree.

[c]RF: random forest.

[d]XGB: extreme gradient boosting.

Organ-Specific Models

Many patients had codes for >1 organ type; therefore, we included all of the codes in organ-specific models. The 2 most important variables in these models in all 3 algorithms included codes for either the correct organ transplant status (V42 and Z94 codes, with decimals specifying organ type), complications of the correct transplanted organ (996 or T86 codes, with decimals specifying organ type), or procedural codes specifying the correct organ type (Table S2 of Multimedia Appendix 1). The PPV, sensitivity, and F-scores for the training and testing sets for each organ type are presented in Table 4.

Table 4. Positive predictive value, sensitivity, and F-score for each machine learning model to identify individuals with specific organ transplant types in the training and testing sets.

Organ, model	Training set			Testing set		
	PPV[a]	Sensitivity	F-score	PPV	Sensitivity	F-score
Heart						
>5 codes	0.974	0.8	0.879	1	0.8	0.889
CART[b]	0.94	0.732	0.824	0.75	0.923	0.828
RF[c]	0.972	0.814	0.886	0.875	1	0.933
XGB[d]	0.972	0.814	0.886	0.875	0.875	0.875
Lung						
>4 codes	0.919	0.872	0.895	1	1	1
CART	0.868	0.78	0.821	0.864	1	0.927
RF	1	0.915	0.956	0.864	1	0.927
XGB	0.981	0.898	0.938	0.864	1	0.927
Kidney						
>4 codes	0.918	0.789	0.849	0.947	0.818	0.878
CART	0.824	0.84	0.832	0.887	0.94	0.913
RF	0.901	0.84	0.869	0.943	0.893	0.917
XGB	0.888	0.85	0.868	0.906	0.906	0.906
Liver						
>6 codes	0.963	0.89	0.925	1	1	1
CART	0.928	0.865	0.896	0.95	0.792	0.864
RF	0.979	0.894	0.935	1	1	1
XGB	0.979	0.904	0.94	1	0.952	0.976
Bone marrow						
>6 codes	0.918	0.69	0.788	0.933	0.875	0.903
CART	0.862	0.884	0.873	0.882	0.789	0.833
RF	0.932	0.828	0.877	0.863	0.898	0.88
XGB	0.909	0.859	0.883	0.863	0.846	0.854

[a]PPV: positive predictive value.

[b]CART: classification and regression tree.

[c]RF: random forest.

[d]XGB: extreme gradient boosting.

Sensitivity Analyses

The United States transitioned from ICD-9 to ICD-10 coding on October 1, 2014. We examined if the model performance differed before or after this time point and found good stability overall. For example, the XGB model for overall transplant status had an F-score of 0.92 before and 0.89 after October 1, 2014. We also noted that the majority of our cases underwent a transplant after the year 2000. We examined model performance before and after January 1, 2000 and found very stable F-scores (0.91 before and 0.92 after in the XGB model for overall transplant status), suggesting little impact on the model based on this imbalance.

Discussion

In this study, we developed and validated phenotyping algorithms for identifying OTRs from the EHR. Using several different rule-based and machine learning methods, we were able to identify OTRs overall with 90% PPV and sensitivity and higher values for several individual organ types. The algorithms all performed comparably well, although RF tended to be the most consistent. The development of these phenotyping algorithms was necessary as the PPV for using at least one transplant code to identify OTRs was only 60%, indicating that studies based on the presence of only one of these codes may have biased results.

The SRTR of the United Network for Organ Sharing and the Organ Procurement and Transplant Network is the primary national database for transplant recipient outcomes research. Because the SRTR is not linked directly to patient records in EHRs, it is not possible to collect data on additional variables not captured by the data entry forms. As a result, many important variables and outcomes are completely omitted. Indeed, a recent study of cardiac transplants using SRTR data found that advanced machine learning methods did not outperform the more traditional prediction models for 1-year survival, with the authors concluding that the methods were hindered by limited data in the registry [21]. By developing validated algorithms to identify OTRs from the EHR, a broader range of studies can be conducted using the data in the full clinical record.

Large reviews of the accuracy of diagnostic and procedural codes show <90% concordance with true diagnoses in inpatient and outpatient settings, both in the United States and other countries [9,22,23]. In a study from Canada, the use of ICD codes alone to identify kidney donors had only 60% sensitivity and 78% PPV, which were similar to our findings for transplant recipients [9]. While the primary diagnosis for a visit is less likely to be missed, secondary diagnoses were more likely to be omitted from the coding. In the United States, up to 12 diagnoses can be entered for an encounter, though only 4 are allowed to be linked to an individual service, with the codes generating the highest reimbursements being prioritized by the medical coders. As a result, transplant patients seen for critical illnesses or procedures may have been less likely to have a transplant code listed.

Many of the charts we reviewed contained only 1 or 2 transplant codes. In addition, these charts often had only ICD and CPT codes but no documents, medications, or laboratory data. Two possible explanations for this lack of data are that handwritten notes and outside records are not scanned into the Synthetic Derivative, and patients with sparse data that could make them potentially identifiable are redacted more often than those with deeper coverage of their records. Regardless of the reason for lack of data, these patients were all called nontransplant patients, and therefore, our algorithm might underestimate the PPV for those with few codes. We attempted to improve our accuracy in classifying these individuals with few transplant codes. First, after identifying this problem in our preliminary analyses, we increased our initial sample by 67% with oversampling of those with only 1 or 2 codes to provide the models with more data points with which to learn to classify them. We also considered adding medications to our algorithms as well as applying NLP to the documents in the EHR. Although these strategies might have augmented the PPV and sensitivity, the gains would have been minimal as those individuals with data besides ICD and CPT codes tended to have a higher number of transplant codes, and therefore, the algorithms had more accurate classification of these patients without the extra data. Moreover, classifying individuals with sparse data as non-OTRs eliminates even those true OTRs who would be excluded from later analyses due to missing data. The true transplant cases that were misclassified were almost exclusively those who had only a single presentation to VUMC with no additional follow-up. Thus, they tended to have only 1 or 2 diagnostic or procedural codes. From a broader standpoint, these were patients who also had little data to contribute to any downstream analyses of the cohort. Therefore, while the models excluded some cases, the overall information loss was low.

There was notable variation in the model metrics both within and between organ types. The reason for the different performance was likely 2-fold. First, there were low numbers for lung transplant recipients (n=81) compared to kidney transplant recipients (n=259); therefore, it is not surprising that the kidney models performed better. Second, the number of different codes contributing to a specific organ type also played a role. For example, although there were 249 stem cell or bone marrow transplant patients, there were 50 different ICD and CPT codes for this type of transplant. Therefore, it is not surprising that the bone marrow models tended to perform worse than the other organ types that had far fewer codes associated, as there were likely subsets within the cross-validations that did not include certain codes. Each code is used in different clinical settings and can be subject to individual coding preferences; therefore, this variability would be expected across institutions.

This study had several limitations. All the data were from a single medical center and coding practices may differ among institutions. Any center wishing to use this approach would need to perform a validation step to confirm the models' performance, although EHR algorithms have been shown to have good portability between populations [24]. VUMC is a high-volume transplant center, and as a result, many patients are seen there for either transplant surgery alone or for follow-up after receiving a transplant elsewhere. This fragmentation of care can limit the available data. Our models consistently predicted slightly greater numbers of OTRs than the number of transplant procedures that have been performed at VUMC. These numbers suggest that we are in fact correctly labeling the majority of those transplants performed at VUMC, while also capturing those whose transplants were performed elsewhere but have been seen in follow-up at VUMC. More than half of the possible OTRs in our EHR had >10 transplant codes, indicating high-density data for these individuals. If we had used >10 transplant codes as our cutoff for OTR determination, the PPV would be 98.5% and the sensitivity would still be 72.3%. Conversely, a large proportion of our cohort had low numbers of transplant codes, which can correlate with the duration of the follow-up. Although the cases identified with low numbers of codes could have easily been excluded a priori by requiring a set number of total codes, doing so would falsely inflate our sensitivity measures, as many true cases would not have been investigated and confirmed on chart review. Our goal was to provide accurate estimates of the algorithm's overall performance, even if many of the identified cases would ultimately be excluded due to missing data in subsequent analyses. Many patients had no available text data from notes. This deficiency likely was the outcome of handwritten notes not being included in the Synthetic Derivative. Thus, we were not able to add NLP to our algorithms, which potentially could have improved our models. EHRs can be a powerful tool for investigating outcomes not captured by large registries.

In this study, we have validated algorithms for identifying OTR overall and OTRs receiving specific organs by using only ICD

and CPT codes. Single variable phenotyping algorithms based on code counts alone perform well but can be improved by using RFs. These algorithms can be used to construct EHR-based cohorts to broaden the range of clinical and translational studies conducted on organ transplants.

Acknowledgments

We would like to specially thank Dr. Josh Denny for the helpful discussions and suggestions. Dr. Wheless is supported by grants from the Skin Cancer Foundation and the Dermatology Foundation. This project was supported by the National Center for Research Resources, Grant UL1 RR024975-01, and is now at the National Center for Advancing Translational Sciences, Grant 2 UL1 TR000445-06. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Authors' Contributions

LW designed the study. LW, LXW, NA, and LE performed the research and analyzed the data. All authors were involved in writing and revising the manuscript and have approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Tuning parameters and variable importance for final models.
[XLSX File (Microsoft Excel File), 18 KB - medinform_v8i12e18001_app1.xlsx]

References

1. Yanik EL, Nogueira LM, Koch L, Copeland G, Lynch CF, Pawlish KS, et al. Comparison of Cancer Diagnoses Between the US Solid Organ Transplant Registry and Linked Central Cancer Registries. Am J Transplant 2016 Oct;16(10):2986-2993 [FREE Full text] [doi: 10.1111/ajt.13818] [Medline: 27062091]
2. Engels EA, Pfeiffer RM, Fraumeni JF, Kasiske BL, Israni AK, Snyder JJ, et al. Spectrum of cancer risk among US solid organ transplant recipients. JAMA 2011 Nov 02;306(17):1891-1901 [FREE Full text] [doi: 10.1001/jama.2011.1592] [Medline: 22045767]
3. Cho S, Mohan S, Husain SA, Natarajan K. Expanding transplant outcomes research opportunities through the use of a common data model. Am J Transplant 2018 Jun;18(6):1321-1327 [FREE Full text] [doi: 10.1111/ajt.14892] [Medline: 29687963]
4. Garrett GL, Yuan JT, Shin TM, Arron ST, Transplant Skin Cancer Network (TSCN). Validity of skin cancer malignancy reporting to the Organ Procurement Transplant Network: A cohort study. J Am Acad Dermatol 2018 Feb;78(2):264-269. [doi: 10.1016/j.jaad.2017.09.003] [Medline: 29031659]
5. Srinivas TR, Taber DJ, Su Z, Zhang J, Mour G, Northrup D, et al. Big Data, Predictive Analytics, and Quality Improvement in Kidney Transplantation: A Proof of Concept. Am J Transplant 2017 Mar;17(3):671-681 [FREE Full text] [doi: 10.1111/ajt.14099] [Medline: 27804279]
6. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc 2013 Jun;20(e1):e147-e154 [FREE Full text] [doi: 10.1136/amiajnl-2012-000896] [Medline: 23531748]
7. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc 2016 Nov;23(6):1046-1052 [FREE Full text] [doi: 10.1093/jamia/ocv202] [Medline: 27026615]
8. What is the Phenotype KnowledgeBase? Phenotype KnowledgeBase. URL: https://phekb.org [accessed 2020-09-16]
9. Lam NN, Lentine KL, Klarenbach S, Sood MM, Kuwornu PJ, Naylor KL, et al. Validation of Living Donor Nephrectomy Codes. Can J Kidney Health Dis 2018;5:2054358118760833 [FREE Full text] [doi: 10.1177/2054358118760833] [Medline: 29662679]
10. Santos CAQ, Brennan DC, Olsen MA. Accuracy of Inpatient International Classification of Diseases, Ninth Revision, Clinical Modification Coding for Cytomegalovirus After Kidney Transplantation. Transplant Proc 2015;47(6):1772-1776 [FREE Full text] [doi: 10.1016/j.transproceed.2015.04.087] [Medline: 26293049]
11. Dhakal S, Burwen DR, Polakowski LL, Zinderman CE, Wise RP. Assessment of tissue allograft safety monitoring with administrative healthcare databases: a pilot project using Medicare data. Cell Tissue Bank 2014 Mar;15(1):75-84. [doi: 10.1007/s10561-013-9376-y] [Medline: 23824508]

12. Roden D, Pulley J, Basford M, Bernard G, Clayton E, Balser J, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther 2008 Sep;84(3):362-369 [FREE Full text] [doi: 10.1038/clpt.2008.89] [Medline: 18500243]
13. Kuhn M. Building Predictive Models in R Using the Caret Package. J Stat Soft 2008;28(5). [doi: 10.18637/jss.v028.i05]
14. The R Project for Statistical Computing. Vienna, Austria URL: https://www.R-project.org/ [accessed 2020-09-16]
15. rpart: Recursive Partitioning and Regression Trees Internet. Therneau T, Atkinson B. 2018. URL: https://CRAN.R-project.org/package=rpart [accessed 2020-09-16]
16. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Soft 2017;77(1). [doi: 10.18637/jss.v077.i01]
17. xgboost: Extreme Gradient Boosting. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T, Li M, Xie J, Lin M, Geng Y, Li Y. 2019. URL: https://CRAN.R-project.org/package=xgboost [accessed 2020-09-16]
18. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc 2016 Sep;23(5):1007-1015 [FREE Full text] [doi: 10.1093/jamia/ocv180] [Medline: 26911811]
19. OPTN: Organ Procurement and Transplant Network. URL: https://optn.transplant.hrsa.gov [accessed 2019-04-22]
20. Center for International Blood and Marrow Transplant Research. URL: https://www.cibmtr.org [accessed 2019-05-06]
21. Miller PE, Pawar S, Vaccaro B, McCullough M, Rao P, Ghosh R, et al. Predictive Abilities of Machine Learning Techniques May Be Limited by Dataset Characteristics: Insights From the UNOS Database. J Card Fail 2019 Jun;25(6):479-483. [doi: 10.1016/j.cardfail.2019.01.018] [Medline: 30738152]
22. Burns EM, Rigby E, Mamidanna R, Bottle A, Aylin P, Ziprin P, et al. Systematic review of discharge coding accuracy. J Public Health (Oxf) 2012 Mar;34(1):138-148 [FREE Full text] [doi: 10.1093/pubmed/fdr054] [Medline: 21795302]
23. Horsky J, Drucker EA, Ramelson HZ. Accuracy and Completeness of Clinical Coding Using ICD-10 for Ambulatory Visits. AMIA Annu Symp Proc 2017;2017:912-920 [FREE Full text] [Medline: 29854158]
24. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc 2012;19(2):212-218 [FREE Full text] [doi: 10.1136/amiajnl-2011-000439] [Medline: 22101970]

Abbreviations

CART: classification and regression tree
CPT: current procedural terminology
EHR: electronic health record
ICD: International Classification of Diseases
NLP: natural language processing
OTR: organ transplant recipient
PPV: positive predictive value
RF: random forest
SRTR: scientific registry for transplant recipients
VUMC: Vanderbilt University Medical Center
XGB: extreme gradient boosting

Edited by G Eysenbach; submitted 28.01.20; peer-reviewed by KM Kuo, V Castillo, Z Predmore, S Purkayastha, P Rane, N Onyeakusi; comments to author 12.06.20; revised version received 21.07.20; accepted 31.10.20; published 10.12.20.

Please cite as:
Wheless L, Baker L, Edwards L, Anand N, Birdwell K, Hanlon A, Chren MM
Development of Phenotyping Algorithms for the Identification of Organ Transplant Recipients: Cohort Study
JMIR Med Inform 2020;8(12):e18001
URL: http://medinform.jmir.org/2020/12/e18001/
doi:10.2196/18001
PMID:33156808

Original Paper

Family History Information Extraction With Neural Attention and an Enhanced Relation-Side Scheme: Algorithm Development and Validation

Hong-Jie Dai[1,2,3], PhD; You-Qian Lee[1], BEng; Chandini Nekkantti[4], BTech; Jitendra Jonnagaddala[5], PhD

[1]College of Electrical Engineering and Computer Science, Department of Electrical Engineering, National Kaohsiung University of Science and Technology, Kaohsiung City, Taiwan

[2]School of Post-Baccalaureate Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan

[3]National Institute of Cancer Research, National Health Research Institutes, Tainan, Taiwan

[4]CGD Health Pty Ltd, Belconnen, Australia

[5]School of Public Health and Community Medicine, University of New South Wales, Sydney, Australia

Corresponding Author:
Hong-Jie Dai, PhD
College of Electrical Engineering and Computer Science
Department of Electrical Engineering
National Kaohsiung University of Science and Technology
No. 415, Jiangong Road, Sanmin District
Kaohsiung City, 807
Taiwan
Phone: 886 7 3814526 ext 15510
Email: hjdai@nkust.edu.tw

Abstract

Background: Identifying and extracting family history information (FHI) from clinical reports are significant for recognizing disease susceptibility. However, FHI is usually described in a narrative manner within patients' electronic health records, which requires the application of natural language processing technologies to automatically extract such information to provide more comprehensive patient-centered information to physicians.

Objective: This study aimed to overcome the 2 main challenges observed in previous research focusing on FHI extraction. One is the requirement to develop postprocessing rules to infer the member and side information of family mentions. The other is to efficiently utilize intrasentence and intersentence information to assist FHI extraction.

Methods: We formulated the task as a sequential labeling problem and propose an enhanced relation-side scheme that encodes the required family member properties to not only eliminate the need for postprocessing rules but also relieve the insufficient training instance issues. Moreover, an attention-based neural network structure was proposed to exploit cross-sentence information to identify FHI and its attributes requiring cross-sentence inference.

Results: The dataset released by the 2019 n2c2/OHNLP family history extraction task was used to evaluate the performance of the proposed methods. We started by comparing the performance of the traditional neural sequence models with the ordinary scheme and enhanced scheme. Next, we studied the effectiveness of the proposed attention-enhanced neural networks by comparing their performance with that of the traditional networks. It was observed that, with the enhanced scheme, the recall of the neural network can be improved, leading to an increase in the F score of 0.024. The proposed neural attention mechanism enhanced both the recall and precision and resulted in an improved F score of 0.807, which was ranked fourth in the shared task.

Conclusions: We presented an attention-based neural network along with an enhanced tag scheme that enables the neural network model to learn and interpret the implicit relationship and side information of the recognized family members across sentences without relying on heuristic rules.

(JMIR Med Inform 2020;8(12):e21750) doi:10.2196/21750

KEYWORDS

family history information; natural language processing; deep learning; electronic health record

Introduction

Family history information (FHI), such as a patient's family members and their corresponding side of the family (ie, maternal or paternal), health-related problems like medical histories and current disorders, and habits of substance use, is not only an essential risk factor for many chronic and hereditary diseases such as cardiovascular diseases, diabetes, and cancers [1] but also an important clue for individualized disease diagnosis, treatment, prediction, and prevention [2-6]. FHI is usually described in an unstructured free-text format within a patient's electronic health record, and its content depends on pieces of information provided by patients about the health situation of their relatives during clinical visits. Therefore, it will be beneficial if natural language processing (NLP) can be employed to identify FHI to provide a more comprehensive view of patient-centered information for physicians.

In general, FHI consists of 3 essential factors, including the relationship between family members, side of the members, and associated observations. Early studies working on the identification of FHI [7,8] relied on the Unified Medical Language System to extract FHI and applied rules to associate the relations. The release of available FHI training corpora such as the BioCreative/OHNLP challenge 2018 [9] and the 2019 n2c2/OHNLP shared tasks prompted the advancement of NLP for automatically extracting FHI. Researchers currently apply a variety of approaches to tackle the task of FHI extraction. For example, Dai [10] introduced 3 inside, outside, beginning (IOB)2-based tag sets that can be utilized to identify family members and their observations along with the bidirectional long short-term memory (BiLSTM)-conditional random field (CRF) model. The first was the standard IOB-2 scheme, which only captures the spans of the mentioned family members and observations. Therefore, 5 tags including B/I-FM, B/I-Ob, and O were used. The second scheme further encodes the family side information in the tag set for family members. For example, "Mother" is not associated with any family side values, so its mention is assigned with the B/I-FM-NA tag, while other tag sets include the B/I-FM-Paternal and B/I-FM-Maternal tags. The relation-side scheme was the last proposed tag scheme in which both the type and side properties are encoded. Consequently, all possible combinations of the 2 properties that appeared in the training set were represented by the tag scheme.

Without encoding both the side and relationship information in tag sets like the relation-side scheme for model training, previous work had to develop sophisticated postprocessing rules that relied on commonsense knowledge and surrounding text to infer the 2 properties of family members and integrate handcrafted rules with deep learning models in a pipeline structure. In addition to the challenge of optimizing both submodules separately, there are at least two other known limitations of applying postprocessing rules. One is the inability to determine cases like indirect relatives as pointed out by Dai [10] and Shi et al [11], and the other is the general ability to classify FHIs represented in different writing styles. Unfortunately, although the aforementioned relation-side scheme is expected to facilitate the development of a single end-to-end model to conquer the task of FHI extraction, the experiment results by Dai [10] revealed issues of insufficient and imbalanced training instances. In light of these constraints, we eliminated the postprocessing rules and managed the issue of training instances by proposing an enhanced relation-side tag scheme. Moreover, we introduced the attention-based neural network structure to better exploit intrasentence and intersentence information to determine the FHIs requiring cross-sentence inference.

Methods

We preprocessed medical notes to generate sentences and the corresponding tokens associated with their part-of-speech information via our clinical toolkit [12]. By formulating the FHI extraction task as a sequential labelling problem, we applied the proposed tag scheme to encode the gold annotations to generate the datasets for training the proposed network models. In the following subsections, we first introduce the relation-side scheme proposed by Dai [10] and the enhanced version proposed in this work, followed by descriptions of the architecture of the developed model that can utilize cross-sentence information via the sentence-level and document-level neural attentions.

Tag Scheme Design

In order to exclude the need for postprocessing steps, Dai [10] presented the relation-side scheme in which both the side and family relationship properties are encoded within the IOB tag sets for family member entities. Table 1 displays an example of the encoded annotations. Taking the first family member mention "two paternal aunts" as an example, we included the side and relationship information ("paternal" and "aunt," respectively, in this case) in the tag set. Since both side and relationship attributes were encoded and later learned by the machine learning model, it is not necessary to apply postprocessing algorithms to infer the 2 properties.

Table 1. An example sentence encoded with the relation-side scheme and enhanced version: "The patient has two paternal aunts and one paternal half–brother, all were diagnosed with type-2 diabetes."

Word	Relation-side scheme	Enhanced relation-side scheme
has	O	O
two	B-Aunt-Paternal	I-FM
paternal	I-Aunt-Paternal	I-FM
aunts	I-Aunt-Paternal	E-Aunt-Paternal
and	O	O
one	B-Brother-NA	I-FM
paternal	I-Brother-NA	I-FM
half-brother	I-Brother-NA	E-Brother-NA
,	O	O
…	…	…
type-2	B-OB	B-OB
diabetes	I-OB	I-OB

The drawback of the relation-side scheme is that the tag scheme combines all required information in its encoding, which is too specific and may result in problems of insufficient training instances. Take the annotations of the n2c2/OHNLP shared task as an example. In their annotations, the first-degree relatives, which include 8 types of family members (ie, Father, Mother, Parent, Sister, Brother, Daughter, Son, and Child), do not have the value of the family side property (refer to the tags ending with "NA" in Table 1). However, annotations of the other 7 family members (ie, Grandmother, Grandfather, Grandparent, Cousin, Sibling, Aunt, and Uncle) contain both properties. Therefore, we have at most 8 x 2 x 1 + 7 x 2 x 3 = 58 tags for family members. Consequently, we proposed the enhanced relation-side scheme in which only the I (inner) and E (end) tags were used and the relationship and side properties were only encoded in the E tag. For example, in Table 1, we can see that the word "paternal" of the 2 family member mentions was encoded by I-FM, which implies that the word is a part of a family mention. The annotations for the last words of the 2 mentions were encoded by including their relationship and side information. The number of possible tags was reduced to 1 + 8 x 1 + 7 x 3 = 30. On the other hand, for observations like "type-2 diabetes" in Table 1, both schemes used the ordinary IOB tag set to encode the annotations. The enhanced tag scheme is preferred because it greatly reduced the size of the tag sets and transition matrix used later in the CRF layer of the developed model.

Baseline Network Architecture

We used the network architecture developed by Dai [10] as a baseline. The network architecture is very similar to the entity recognition part of the network developed by Shi et al [11], with the major difference being that the latter further extended the network with an additional BiLSTM to create a joint learning model. Both were top-ranked systems in the BioCreative/OHNLP challenge.

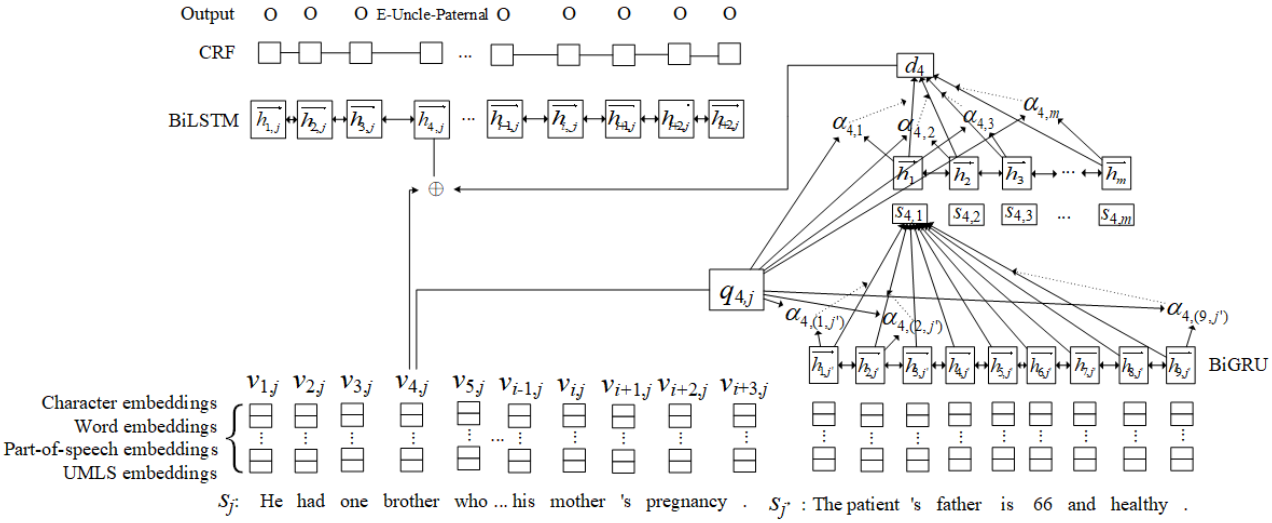
In our implementation, the baseline architecture consists of 2 core parts, with the first being the representation layer in which the sequence of tokens $\mathbf{t} = \{t_1,t_2,\ldots,t_n\}$ was represented as a vector by concatenating the character-level representation based on convolutional neural networks, pre-trained word representations, the randomly initialized part-of-speech embedding, and the pre-trained Unified Medical Language System embedding [13]. Based on the investigation by Dai [10] on the effectiveness of applying different pretrained word embeddings to the task of FHI extraction and the effectiveness of the recent advancement of contextualized word representations, global vectors for word representation (GloVe) [14] and the embeddings from language models (ELMo) [15] were used to represent the tokens. The concatenated representation was then inputted to a BiLSTM network with CRF as the output layer to infer predictions for each token.

The BiLSTM CRF networks have been shown to be able to efficiently model contextual information and label dependencies [16] and is currently a strong baseline. However, one major constraint is that the networks can only exploit contexts within individual sequences but cannot digest cross-sentence information. To overcome this limitation, we enhanced the baseline model by introducing the neural attentions described in the next subsection.

Attention-Enhanced BiLSTM-CRF Network Architecture

Figure 1 illustrates the network architecture of the proposed attention-enhanced network. In the network, for each token $t_{i,j}$ in a given sentence s_j, we applied the attention mechanism to make it attend to certain tokens among all sentences $\{s_1,s_2,\ldots,s_m\}$ of the document **d** to allow the model to determine the type and the attributes of the token $t_{i,j}$ by considering information at the sentence and document levels. Each sentence s_j in the input document **d** is expressed as $\mathbf{t_j} = \{t_{1,j},t_{2,j},\ldots,t_{n,j}\}$ where n is the number of tokens in s_j.

Figure 1. Proposed attention-enhanced bidirectional long short-term memory (BiLSTM)-conditional random field (CRF) network architecture. ⊕ indicates a concatenation of two vectors. BiGRU: bidirectional gated recurrent unit; UMLS: Unified Medical Language System.



Like our baseline model, each token $t_{i,j}$ in the sequence of tokens $\mathbf{t_j}$ was represented as a vector $v_{i,j}$ by concatenating the embeddings described in the previous subsection. Before sending the vector to the BiLSTM-CRF layer as an input, a hierarchical attention layer is introduced to enrich the vector to enable the model in utilizing cross-sentence information. In the attention layer, the attention score, which conveys the associations between the current token's representation $v_{i,j}$ and all tokens' representations in **d**, was hierarchically calculated using the following content-based function adapted from Luong et al [17] where $\mathbf{W_t}$ and $\mathbf{W_{t'}}$ are learned parameters and $h_{i',j'}$ is the hidden state of the bidirectional gated recurrent unit at the token $t_{i',j'}$ from another sentence:

$$s_{j'}: q(v_{i,j}) = \mathbf{W_t}v_{i,j} + b_q \textbf{(1)}$$

$$t\_w(h_{i',j'}) = \tanh(\mathbf{W_{t'}}h_{i',j'} + b_{t\_s}) \textbf{(2)}$$

The score was calculated sentence-wise for the token $t_{i,j}$ to derive its attention weight $\alpha_{i,(i',j')}$ for the token $t_{i',j'}$ in the sentence $s_{j'}$:

$$\text{score}(v_{i,j},h_{i',j'}) = q(v_{i,j})^{\mathrm{T}} t\_w(h_{i',j'}) \textbf{(3)}$$

The aggregated score $s_{i,j'}$ for all tokens in $s_{j'}$ was calculated as follows:

Given the aggregated sentence scores $\mathbf{s}_i = \{s_{i,1},s_{i,2},\ldots s_{i,m}\}$ for the token $t_{i,j}$, we derived a document vector d_i in a similar way to summarize the information from all sentences. First, a bidirectional gated recurrent unit was used to encode $\mathbf{s}_i$, which can generate the hidden state h_k for the k^{th} vector in $\mathbf{s}_i$. Analogous to the hierarchical attention networks proposed by Yang et al [18], we rewarded sentences that provide clues to infer the type and attribute information of the target token $t_{i,j}$ using the following attention mechanism:

$$t\_s(h_k) = \tanh(\mathbf{W}_s h_k + b_{t\_s}) \textbf{(6)}$$

$$\text{score}(v_{i,j},h_k) = q(v_{i,j})^{\mathrm{T}} t\_s(h_k) \textbf{(7)}$$

The output of the hierarchical attention layer d_i can be considered as a document-level vector that summarizes information across sentences in **d** for token $t_{i,j}$, which provides clues for determining FHI. Finally, the document vector was treated as an additional feature vector and concatenated with the original token representations to form the input of the BiLSTM-CRF model.

Experiment Configurations

The dataset released by the 2019 n2c2/OHNLP shared task was used to evaluate the performance of the proposed network architecture along with the designed tag scheme. The training and test sets consist of 99 and 117 unstructured clinical notes, respectively. We randomly selected 83 of the 99 notes as the final training set, with the remaining 16 notes as the validation set in the training process. The validation set was not used in training but was used to determine the optimum parameters without overfitting the training set. We configured 3 runs for the participation of the n2c2/OHNLP family history extraction track. Both the first and second configurations were based on the proposed neural attention network along with the enhanced relation-side scheme. The only difference is that when processing a given sentence, the first configuration took all sentences in the note into consideration, while the second only examined sentences before the current one. The last run was based on the baseline BiLSTM-CRF network described in the previous subsection.

In addition to the submitted runs, we studied the effectiveness of the proposed tag scheme by training the baseline and attention-enhanced networks with different schemas and reported their performance on the test set. Table 2 summarizes all the configurations studied in this work. All the networks were implemented using CUDA 10.1 and PyTorch libraries

trained on machines equipped with NVIDIA Tesla P100 graphics cards. The mini-batch gradient descent along with Adam [19] was used for optimizing the parameters. The epoch was set to 200, and the early stopping strategy (a patience value of 50) was used if no improvement in the F score or loss was observed or the loss became zero on the validation set. The same set of hyperparameters and a fixed random seed were used to train all the configurations shown in Table 2.

Table 2. Summary of the configurations studied in this work.

Configuration	Description	Notation
Baseline + relation-side scheme	BiLSTM-CRF[a] with relation-side scheme	B-RS
Baseline + enhanced relation-side scheme	BiLSTM-CRF with enhanced relation-side scheme	B-ERS
Attention + relation-side scheme	Attention-enhanced BiLSTM-CRF with relation-side scheme	A-RS
Attention + enhanced relation-side scheme	Attention-enhanced BiLSTM-CRF with enhanced relation-side scheme paying attention to limited sentences	A-ERS
Attention + enhanced relation-side scheme (+)	Attention-enhanced BiLSTM-CRF with enhanced relation-side scheme paying attention to all sentences	A-ERS+

[a]BiLSTM-CRF: bidirectional long short-term memory-conditional random field.

The official evaluation script [20] released by the organizers was used to report the performance of the developed models. The performance for the recognized family member mentions including their family side attributes and observations were reported in terms of the standard precision (P), recall (R), and F1-measure (F) defined as follows at the article level:

$$\text{Precision} = TP/TP + FP \quad \textbf{(10)}$$

$$\text{Recall} = TP/TP + FN \quad \textbf{(11)}$$

$$F_1 = 2 \times P \times R/(P + R) \quad \textbf{(12)}$$

For each recognized family member mention, the 15 types of relatives described in the previous subsections were considered for evaluation. For each correctly recognized family member mention, its side of the family (ie, paternal, maternal, or not available) must also be correctly classified so that a true positive can be counted, else both the false positive and false negative are increased by one.

Results

In the following subsections, we first compare the performance of the baseline model with the enhanced relation-side scheme to that of the model with the original scheme. Subsequently, we investigate the effect of the proposed attention-enhanced network architectures.

Effect of the Enhanced Relation-Side Scheme

Table 3 outlines the performance of the baseline models with the original relation-side scheme (B-RS) and the proposed enhanced version (B-ERS). The last column of the table also shows the F scores for both models on the validation set and the number of executed epochs before terminating. With the early stopping strategy described in the previous section, both models terminated their training phase in advance and achieved F scores larger than 0.94 on the training set. The B-ERS model generally outperformed the B-RS model on the validation and test sets. It can be observed that the recalls of the B-ERS model for both family member mention and observation were better than those of the B-RS model by 0.061 and 0.117, respectively, which led to an increase in the overall F score of 0.024. These results demonstrated that the proposed enhanced scheme provides a better representation and facilitates a better learning process for the model.

Table 3. Effect of the proposed enhanced relation-side scheme on the test and validation sets.

Configuration	Family member			Observation			Overall	F on the validation set	Number of epochs
	P[a]	R[b]	F	P	R	F	F		
B-RS[c]	0.896	0.658	0.759	0.718	0.813	0.762	0.761	0.795	88
B-ERS[d]	0.882	0.719	0.792	0.674	0.928	0.781	0.785	0.822	124

[a]P: precision.

[b]R: recall.

[c]B-RS: bidirectional long short-term memory-conditional random field with relation-side scheme.

[d]B-ERS: bidirectional long short-term memory-conditional random field with enhanced relation-side scheme.

Effect of the Cross-Sentence Attention

Table 4 provides the results of the comparative evaluation in accordance with the P, R, and F scores of the B-RS model. All proposed attention-enhanced BiLSTM-CRF models obtained better P, R, and F scores than those of the baseline model (B-RS). Among them, A-ERS+, our best submitted run during the 2019 n2c2/OHNLP shared task, had the best performance with improvements of 0.034, 0.058, and 0.046 in terms of P, R, and F scores, respectively. It is noted that the proposed attention mechanism apparently improved the recall of family member mention for all 3 models. In particular, the recall of A-ERS+

can be boosted by 0.118, resulting in a better F score of 0.807. Furthermore, the F scores of observations among the attention-enhanced models were also improved by at least 0.022.

Table 4. Comparison of the performance of the different attention-enhanced bidirectional long short-term memory-conditional random field (BiLSTM-CRF) models.

Performance measures	A-RS[a]	A-ERS[b]	A-ERS+[c]
Family member			
Precision	–0.031	–0.008	–0.046
Recall	+0.053	+0.092	+0.118
F score	+0.022	+0.054	+0.052
Observation			
Precision	–0.031	+0.011	+0.061
Recall	+0.053	+0.074	+0.018
F score	+0.022	+0.038	+0.042
Overall F score	+0.007	+0.044	+0.046

[a]A-RS: attention-enhanced BiLSTM-CRF with relation-side scheme.

[b]A-ERS: attention-enhanced BiLSTM-CRF with enhanced relation-side scheme paying attention to limited sentences.

[c]A-ERS+: attention-enhanced BiLSTM-CRF with enhanced relation-side scheme paying attention to all sentences.

Discussion

Principal Findings

Dai [10] provided an intensive analysis of the effectiveness of applying different tag schemes to the task of FHI extraction. In short, the advantage of applying the relation-side scheme is that we can eliminate the creation of heuristic rules for determining the relationship and side information of the recognized family member mentions, which is a major issue experienced by using standard tag schemes. Nevertheless, Dai [10] also pointed out that employing the scheme could lead to sparse and imbalanced training instances if the released dataset was small, which hinders the construction of a reliable model for identifying the desired properties of recognized mentions.

In this study, we addressed these issues by developing an enhanced relation-side scheme that achieved promising results, as shown in Table 4. We believe that the performance gain comes from the refined tag set distribution, where the enhanced scheme has significantly fewer tag types (30 vs 66). The tag with the highest distribution in the enhanced scheme is I-FM, which indicates that 35% of family member mentions in the training set consist of more than 1 token after tokenization, followed by E-FM-Mother-Na (7%), E-FM-Sister-NA (6%), E-FM-Father-NA (6%), E-FM-Brother-NA (6%), E-FM-Aunt-Maternal (5%), E-FM-Son-NA (4%), E-FM-Aunt-Paternal (4%), E-FM-Daughter-NA (3%), and E-FM-Uncle-Paternal (3%; Multimedia Appendix 1).

By contrast, no tags occupied more than 10% of the overall distribution in the original relation-side scheme. The top 10 tag types are as follows: B-FM-Mother-NA (7%), B-FM-Father-NA (6%), B-FM-Sister-NA (6%), B-FM-Brother-NA (5%), B-FM-Aunt-Maternal (5%), I-FM-Aunt-Maternal (4%), B-FM-Son-NA (4%), B-FM-Aunt-Paternal (4%), B-FM-Daughter-NA (4%), and I-FM-Grandmother-Maternal (3%; Multimedia Appendix 1). It is also worth noting that some family member types possessed frequent inner tags. For example, there are more instances of the inner tag for “Aunt-Maternal” (I-FM-Aunt-Maternal) than other members such as son and daughter, and the inner tag of “Grandmother-Maternal” (I-FM-Grandmother-Maternal) appears more frequently than its beginning tag. A scrutiny of the example shown in Table 1 revealed that the use of the tag scheme increased the degree of lexical ambiguity. For instance, the word “paternal” in Table 1 is assigned with 2 different tags (“I-Brother-NA” and “I-Aunt-Paternal”) although it is just a hint for the mention of family members. This observation also leads to the issue of imbalanced training samples because the word “paternal” could be a beginning or inner word for several types of family members. However, the distribution of those member types is skewed in the training set.

On the other hand, the enhanced relation-side scheme uses I-FM to capture clues that enable the model to learn and make final classifications based on the word with the most informative representation, which is usually the last word in terms of the family member entities. The scheme also resolves the problem of insufficient training samples. By considering Table 1 as an example, the traditional IOB2 scheme encodes all properties in its tag set. As a result, the token “aunts” can be associated with 6 different kinds of tags (B/I-Aunt-Paternal/Maternal/NA). With respect to the enhanced scheme, the token can only be associated with one of the E-Aunt-Paternal/Materal/NA tags, regardless of it being a single or compound noun. Examination of this problem from a different perspective is displayed in Table 5, which shows an evidently higher level of ambiguity in the relation scheme against the enhanced version. It was also found that even with the final CRF layer, the model with the original relation-side scheme could generate illegal tag sequences in the decoding phase, for instance a B-Aunt-Paternal followed by an I-Brother-Paternal, which was not observed in the model with the enhanced scheme.

Table 5. Comparison of the degrees of ambiguity between the relation-side scheme and enhanced relation-side scheme. Note that the tokens that were only associated with the "O" tag were excluded.

Scheme type	Number of possible tags associated with a token											
	1	2	3	4	5	6	7	8	9	10	17	20
Relation-side scheme	535	174	41	3	3	8	5	1	5	1	1	1
Enhanced relation-side scheme	535	188	38	11	5	1	0	0	0	0	0	0

Another challenge that was brought up in Dai [10] is that the perception of the member type and its side property may require cross-sentence inference. In light of this issue, we proposed using the attention mechanism to enhance the ability of the model for identifying these 2 properties. As shown in Table 4, the F scores of not only the family members but also the observations were improved by implementing the attention mechanism, with the improvement particularly due to a boost in the recall. After comparing the results of the models with and without the attention mechanism, we confirmed that the attention-enhanced networks can better exploit the intrasentence and intersentence information to successfully determine the type and side information of family member mentions in which the traditional model failed. Take the following 2 sentences as an example:

> *The **father** of the baby has a **maternal uncle** with a repaired cleft lip. His **uncle** is otherwise said to be healthy.*

The attention-enhanced model can correctly assign the side attribute (ie, maternal) for the "uncle" mentioned in the second sentence, while this could not be accomplished by the baseline model. We identified several similar cases on the test set, although these correct assignments could not be captured by the applied article level evaluation metrics.

Furthermore, we observed that the enhanced model can learn better from the implicit dispersed second-degree relative descriptions without interfering with rules created based on human knowledge. Some examples that can be correctly inferred are as follows.

The enhanced model can correctly assign the "Cousin_Paternal" tag to the children of the patient's aunt even when the mentions are dispersed away from each other:

> *The **paternal aunt** died in her late 57s due to heart complications. She had five **children**. One of these children is a **daughter** who was diagnosed with breast cancer at the age of 42...*

Another similar example would be the sentence, where the enhanced model can correctly determine the side and member type of the mention "son":

> *Mrs. Lucas has another **paternal uncle** who has a **son** with mental retardation of unknown cause.*

For the following sentence, the mentions "sisters" and "brother" within the sentence located in the later part of the document can be correctly recognized by the enhanced model as "Aunt_Paternal" and "Uncle_Paternal," respectively:

> *Ms. James AJ Benjamin's **father**, 55s, is reportedly in good health. ... He has two **sisters** and a **brother**, 63s–71s, who are reportedly in good health.*

In the following description, the second mention of "mother" is successfully assigned with "Grandmother_Maternal":

> *She is 5 feet 6-8 inches tall and the patient's **mother** resembles her own **mother** in facial appearance.*

For the following narrative, the model learned to assign the mention "daughter" with "Sister_NA":

> *The **father** has a 9-year-old **daughter** with another partner who is healthy.*

We also noted that the enhanced networks can acknowledge negative clues and avoid false positive cases of observations:

> *She has **no** history of joint hypermobility, easy bruising, or problems with healing.*
>
> *They do not look different than other members of the family, and **do not have** any major internal birth defects.*

Error Analysis

Although models with neural attentions learned to infer implicit relationships among recognized family member mentions by interpreting the contextual expressions with weighted attentions, ambiguity of the context can still occasionally confuse the model in making incorrect classifications. Some examples as such are listed.

In the following example, while the patient is Mrs. William, the attention-enhanced model focused on the terms "He," "sister," and "his father" and mistakenly assigned the mention "son" with the "Cousin_Paternal" tag:

> *... William's husband is healthy at age 38 with a history of melanoma ... He also has a 39-year-old **sister** who is healthy with a healthy 10-year-old **son**. ... His **father** is alive at age 59 with coronary disease, ...*

In the following example, even with the proposed methods, the developed models could not recognize "mother's mother's brothers" in the second sentence as a family mention. Nevertheless, the attention-enhanced model was able to classify the first mention "brother" as the patient's uncle and the mention "children" as the patient's cousin. On the contrary, the baseline model classified the first and the second mentions as "brother" and "son," respectively:

> *A **brother** is the father of two **children**, a male with **mental retardation** and a daughter with bicuspid mitral valve stenosis and aortic stenosis. Another of*

> *Benjamin's **mother's mother's brothers** is the father of two girls, one of whom ...*

Based on the description, the attention-enhanced model incorrectly considered the mention "father" to be referring to the father of the patient (ie, Mrs. Henrietta):

> *Mrs. Henrietta is of Indian descent. The **father** of the baby is of Indonesian descent.*

For the following sentence, the attention-enhanced model failed to ignore the in-law relationships:

> *Her husband has an identical twin **brother** who is healthy with fraternal twin **daughters**, ...*

Some annotation errors or biases in the corpus were identified during the error analysis. First, we found that not all instances of the same family member in a given electronic health record were annotated, which means that some mentions may only be annotated once even if they refer to the same entity. In general, more cases as such occurred in the annotation of first-degree relatives rather than those of the second-degree relatives (0.586 vs 0.839) based on our estimation on the training set. One conspicuous example of this error can be found in the sentence "*The patient's **mother** is 54 now,*" where the mention "mother" was not annotated. We also noticed that the spans of some family member annotations were incorrect, which may lead to a decrease in performance. For instance, the two annotations in the sentences "*His only [**child,**] a daughter ...*" and "*This aunt has five healthy sons and one [**daughter,**] age 67, ...*" will instruct the models to accept commas to be the last token of a family mention.

Comparison With Prior Work

Several research projects have previously worked on the FHI extraction task. Shi et al [11] developed a neural network model based on BiLSTM networks for joint learning of FHIs and the relations among them. Zhan et al [21] fine-tuned the bidirectional encoder representations from transformers [22] by including an additional Biaffine classifier adapted from the dependency parsing to extract FHIs. Most researchers considered the extraction of FHIs as a sequential labelling task and exploited sequential labelling models to address it. For instance, Kim et al [23] established an ensemble of 10 BiLSTM-CRF models along with ELMo representations to identify FHIs. Later, Wu and Verspoor [24] and Ambalavanan and Devarakonda [25] implemented similar strategies to encode the side information in their tag sets. The former applied a BiLSTM model with ELMo and a tag set that allow the model to recognize mentions of family members and determine their side information at the same time, while the latter further contained family relationship information in their tag set. Similar to this work, the attempt of these 2 works is to eliminate the application of postprocessing rules to infer the required properties of family members.

Conclusions

In this paper, we considered the problem of FHI extraction as a sequential labelling task and presented an attention-based neural network approach to handle this problem. The main contribution of our work is that we presented an improved tag scheme that enables the model to learn and interpret the implicit relationships and side information of the recognized family members without relying on heuristic rules. Moreover, a network structure with neural attentions was proposed to exploit intrasentence and intersentence information to determine the family member mentions and side attributes requiring cross-sentence inference. The feasibility of the proposed method was assessed on the dataset released by the 2019 n2c2/OHNLP shared task on family history extraction and was officially ranked 4th among 17 teams. Although the proposed methods addressed the limitations raised, our error analysis revealed challenges including annotation bias and the requirement of common-sense reasoning, which leave room for further improvement in the future.

Acknowledgments

The authors gratefully acknowledge funding from the Ministry of Science and Technology of Taiwan: grant numbers MOST-106-2221-E-143-007-MY3 and grant numbers MOST 109-2221-E-992-074-MY3. We also thank Dr. Feichen Shen and the other organizers of the n2c2/OHNLP track on family history extraction for their effect in organizing the challenge and releasing the annotated data.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Comparison of the tag set distributions on the training set between the relation-side scheme and its enhanced version. Only the tag names within the top 10 of the distribution are shown in the figure.

[PNG File , 282 KB - medinform_v8i12e21750_app1.png]

References

1. Yoon PW, Scheuner MT, Peterson-Oehlke KL, Gwinn M, Faucett A, Khoury MJ. Can family history be used as a tool for public health and preventive medicine? Genet Med 2002 Aug;4(4):304-310. [doi: 10.1097/00125817-200207000-00009] [Medline: 12172397]

2. Claassen L, Henneman L, Janssens ACJ, Wijdenes-Pijl M, Qureshi N, Walter FM, et al. Using family history information to promote healthy lifestyles and prevent diseases; a discussion of the evidence. BMC Public Health 2010 May 13;10(1):248 [FREE Full text] [doi: 10.1186/1471-2458-10-248] [Medline: 20465810]
3. Guttmacher A, Collins FS, Carmona RH. The family history--more important than ever. N Engl J Med 2004 Nov 25;351(22):2333-2336. [doi: 10.1056/NEJMsb042979] [Medline: 15564550]
4. Murff H, Byrne D, Syngal S. Cancer Risk Assessment: Quality and Impact of the Family History Interview. American Journal of Preventive Medicine 2004 Oct;27(3):239-245. [doi: 10.1016/s0749-3797(04)00119-9]
5. Williams RR, Hunt SC, Heiss G, Province MA, Bensen JT, Higgins M, et al. Usefulness of cardiovascular family history data for population-based preventive medicine and medical research (The Health Family Tree Study and the NHLBI Family Heart Study). The American Journal of Cardiology 2001 Jan;87(2):129-135. [doi: 10.1016/s0002-9149(00)01303-5]
6. Wood ME, Kadlubek P, Pham TH, Wollins DS, Lu KH, Weitzel JN, et al. Quality of Cancer Family History and Referral for Genetic Counseling and Testing Among Oncology Practices: A Pilot Test of Quality Measures As Part of the American Society of Clinical Oncology Quality Oncology Practice Initiative. JCO 2014 Mar 10;32(8):824-829. [doi: 10.1200/jco.2013.51.4661]
7. Friedlin J, McDonald CJ. Using a natural language processing system to extract and code family history data from admission reports. AMIA Annu Symp Proc 2006:925 [FREE Full text] [Medline: 17238544]
8. Goryachev S, Kim H, Zeng-Treitler Q. Identification and extraction of family history information from clinical reports. AMIA Annu Symp Proc 2008 Nov 06:247-251 [FREE Full text] [Medline: 18999129]
9. Liu S, Wang Y, Liu H. Selected articles from the BioCreative/OHNLP challenge 2018. BMC Med Inform Decis Mak 2019 Dec 27;19(Suppl 10):262 [FREE Full text] [doi: 10.1186/s12911-019-0994-6] [Medline: 31882003]
10. Dai H. Family member information extraction via neural sequence labeling models with different tag schemes. BMC Med Inform Decis Mak 2019 Dec 27;19(Suppl 10):257 [FREE Full text] [doi: 10.1186/s12911-019-0996-4] [Medline: 31881965]
11. Shi X, Jiang D, Huang Y, Wang X, Chen Q, Yan J, et al. Family history information extraction via deep joint learning. BMC Med Inform Decis Mak 2019 Dec 27;19(Suppl 10):277 [FREE Full text] [doi: 10.1186/s12911-019-0995-5] [Medline: 31881967]
12. Dai H, Syed-Abdul S, Chen C, Wu C. Recognition and Evaluation of Clinical Section Headings in Clinical Documents Using Token-Based Formulation with Conditional Random Fields. Biomed Res Int 2015;2015:873012-873010 [FREE Full text] [doi: 10.1155/2015/873012] [Medline: 26380302]
13. De Vine L, Zuccon G, Koopman B, Sitbon L, Bruza P. Medical semantic similarity with a neural language model. 2014 Presented at: 23rd ACM International Conference on Conference on Information and Knowledge Management; November 3-7, 2014; Shanghai, China p. 1819-1822. [doi: 10.1145/2661829.2661974]
14. Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. 2014 Presented at: Empirical Methods in Natural Language Processing (EMNLP); October 2014; Doha, Qatar p. 1532-1543. [doi: 10.3115/v1/d14-1162]
15. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. 2018 Presented at: 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT); June 1-6, 2018; New Orleans, LA. [doi: 10.18653/v1/n18-1202]
16. Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. 2016 Presented at: 54th Annual Meeting of the Association for Computational Linguistics; August 7–12, 2016; Berlin, Germany p. 1064-1074. [doi: 10.18653/v1/p16-1101]
17. Luong MT, Pham H, Manning CD. Effective Approaches to Attention-based Neural Machine Translation. 2015 Presented at: Conference on Empirical Methods in Natural Language Processing; September 17-21, 2015; Lisbon, Portugal. [doi: 10.18653/v1/d15-1166]
18. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. 2016 Presented at: 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 12-17, 2016; San Diego, CA. [doi: 10.18653/v1/n16-1174]
19. Loshchilov I, Hutter F. Decoupled weight decay regularization. 2019 Presented at: ICLR 2019; May 6-9, 2019; New Orleans, LA p. 5101.
20. Liu S, Mojarad MR, Wang Y, Wang L, Shen F, Fu S, et al. Overview of the BioCreative/OHNLP Family History Extraction Task. 2018 Presented at: BioCreative/OHNLP Challenge 2018; August 29, 2018; Washington DC. [doi: 10.1145/3233547.3233672]
21. Zhan K, Xiong Y, Fu H, Jiang D, Tang B, Chen Q, et al. Family History Extraction Using Deep Biaffine Attention. 2019 Presented at: n2c2/OHNLP Shared Task and Workshop; November 15, 2019; Washington DC. [doi: 10.2196/preprints.23587]
22. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT); June 2-7, 2019; Minneapolis, MN p. 4171-4186.
23. Kim Y, Heider M, Lally IRH, Meystre SM. A Hybrid Model for Entity Identification and Relation Classification of Family History Information. 2019 Presented at: n2c2/OHNLP Shared Task and Workshop; November 15, 2019; Washington DC.
24. Wu T, Verspoor K. Deep Neural Networks for Family History Information Extraction. 2019 Presented at: n2c2/OHNLP Shared Task and Workshop; November 15, 2019; Washington DC.

25. Ambalavanan AK, Devarakonda M. Named Entity Recognition for Family History Extraction. 2019 Presented at: n2c2/OHNLP Shared Task and Workshop; November 15, 2019; Washington DC.

Abbreviations

A-ERS: attention-enhanced bidirectional long short-term memory-conditional random field with enhanced relation-side scheme paying attention to limited sentences
A-ERS+: attention-enhanced bidirectional long short-term memory-conditional random field with enhanced relation-side scheme paying attention to all sentences
A-RS: attention-enhanced bidirectional long short-term memory-conditional random field with relation-side scheme
B-ERS: bidirectional long short-term memory-conditional random field with enhanced relation-side scheme
BiLSTM: bidirectional long short-term memory
B-RS: bidirectional long short-term memory-conditional random field with relation-side scheme
CRF: conditional random field
ELMo: embeddings from language models
F: F score
FHI: family history information
GloVe: global vectors for word representation
IOB: inside, outside, beginning
NLP: natural language processing
P: precision
R: recall

Edited by C Lovis; submitted 20.07.20; peer-reviewed by S Kim, R Abeysinghe; comments to author 26.09.20; revised version received 10.10.20; accepted 18.10.20; published 01.12.20.

Please cite as:
Dai HJ, Lee YQ, Nekkantti C, Jonnagaddala J
Family History Information Extraction With Neural Attention and an Enhanced Relation-Side Scheme: Algorithm Development and Validation
JMIR Med Inform 2020;8(12):e21750
URL: https://medinform.jmir.org/2020/12/e21750
doi:10.2196/21750
PMID:33258777

Original Paper

The Impact of Pretrained Language Models on Negation and Speculation Detection in Cross-Lingual Medical Text: Comparative Study

Renzo Rivera Zavala[1,2*], MSc; Paloma Martinez[1*], PhD

[1]Department of Computer Science and Engineering, Carlos III University of Madrid, Madrid, Spain

[2]Department of Computer Science and Engineering, Universidad Católica de Santa Maria, Arequipa, Peru

[*]all authors contributed equally

Corresponding Author:
Renzo Rivera Zavala, MSc
Department of Computer Science and Engineering
Carlos III University of Madrid
Avda. Universidad, 30
Leganes
Madrid, 28911
Spain
Phone: 34 916249433
Email: renzomauricio.rivera@alumnos.uc3m.es

Abstract

Background: Negation and speculation are critical elements in natural language processing (NLP)-related tasks, such as information extraction, as these phenomena change the truth value of a proposition. In the clinical narrative that is informal, these linguistic facts are used extensively with the objective of indicating hypotheses, impressions, or negative findings. Previous state-of-the-art approaches addressed negation and speculation detection tasks using rule-based methods, but in the last few years, models based on machine learning and deep learning exploiting morphological, syntactic, and semantic features represented as spare and dense vectors have emerged. However, although such methods of named entity recognition (NER) employ a broad set of features, they are limited to existing pretrained models for a specific domain or language.

Objective: As a fundamental subsystem of any information extraction pipeline, a system for cross-lingual and domain-independent negation and speculation detection was introduced with special focus on the biomedical scientific literature and clinical narrative. In this work, detection of negation and speculation was considered as a sequence-labeling task where cues and the scopes of both phenomena are recognized as a sequence of nested labels recognized in a single step.

Methods: We proposed the following two approaches for negation and speculation detection: (1) bidirectional long short-term memory (Bi-LSTM) and conditional random field using character, word, and sense embeddings to deal with the extraction of semantic, syntactic, and contextual patterns and (2) bidirectional encoder representations for transformers (BERT) with fine tuning for NER.

Results: The approach was evaluated for English and Spanish languages on biomedical and review text, particularly with the BioScope corpus, IULA corpus, and SFU Spanish Review corpus, with F-measures of 86.6%, 85.0%, and 88.1%, respectively, for NeuroNER and 86.4%, 80.8%, and 91.7%, respectively, for BERT.

Conclusions: These results show that these architectures perform considerably better than the previous rule-based and conventional machine learning–based systems. Moreover, our analysis results show that pretrained word embedding and particularly contextualized embedding for biomedical corpora help to understand complexities inherent to biomedical text.

(JMIR Med Inform 2020;8(12):e18953) doi:10.2196/18953

KEYWORDS

natural language processing; clinical text; deep learning; long short-term memory; contextual information

Introduction

A part of clinical data is often described in unstructured free text, such as that recorded in electronic health records (EHRs), medical records, and clinical narrative, which is not analyzed. Besides, scientific literature databases collect valuable publications necessary to extract biomedical data, such as drug or protein interactions, adverse drug effects, disabilities, diseases, treatments, detection of cancer symptoms, and suicide prevention. Biomedical experts and clinicians need to access information and knowledge in their different research areas, convert research results into clinical practice, accelerate biomedical research, provide clinical decision support, and generate data and information in a structured way for downstream processing and applications, such as those specified previously [1]. However, identifying all the data in unstructured documents and translating these data to structured data can be a complex and time-consuming task. It is impossible for experts to process all the documents without tools that filter, classify, and extract information. That is why new techniques are necessary for the extraction of useful knowledge in a precise and efficient way.

One of the main tools currently used for text mining is natural language processing (NLP) and specifically an information extraction system. Information extraction is devoted to processing text and detecting relevant information about specific subjects (for instance, a disease of a patient in a clinical note or a carcinoma in a radiologic report). In information extraction, we can identify low-level tasks and high-level tasks (Figure 1). Low-level tasks are more feasible and affordable processing tasks, such as sentence segmentation, tokenization, and word decomposition. High-level tasks are more complex tasks because they require semantic and contextual knowledge that is provided by domain-specific resources, such as ontologies, and they involve disambiguating terms (such as abbreviations that are highly ambiguous terms) and making inferences with the extracted knowledge. These high-level tasks are named entity recognition (NER), relation extraction, and negation and speculation detection, among others (Tables 1 and 2). For example, extracting a patient's current diagnostic information involves NER, disambiguation, negation and speculation detection, relation extraction, and temporal inference. Figure 2 provides an example of an annotation generated by a medical information extraction system [2].

Figure 1. Typical information extraction pipeline. NLP: natural language processing; PoS: part of speech.

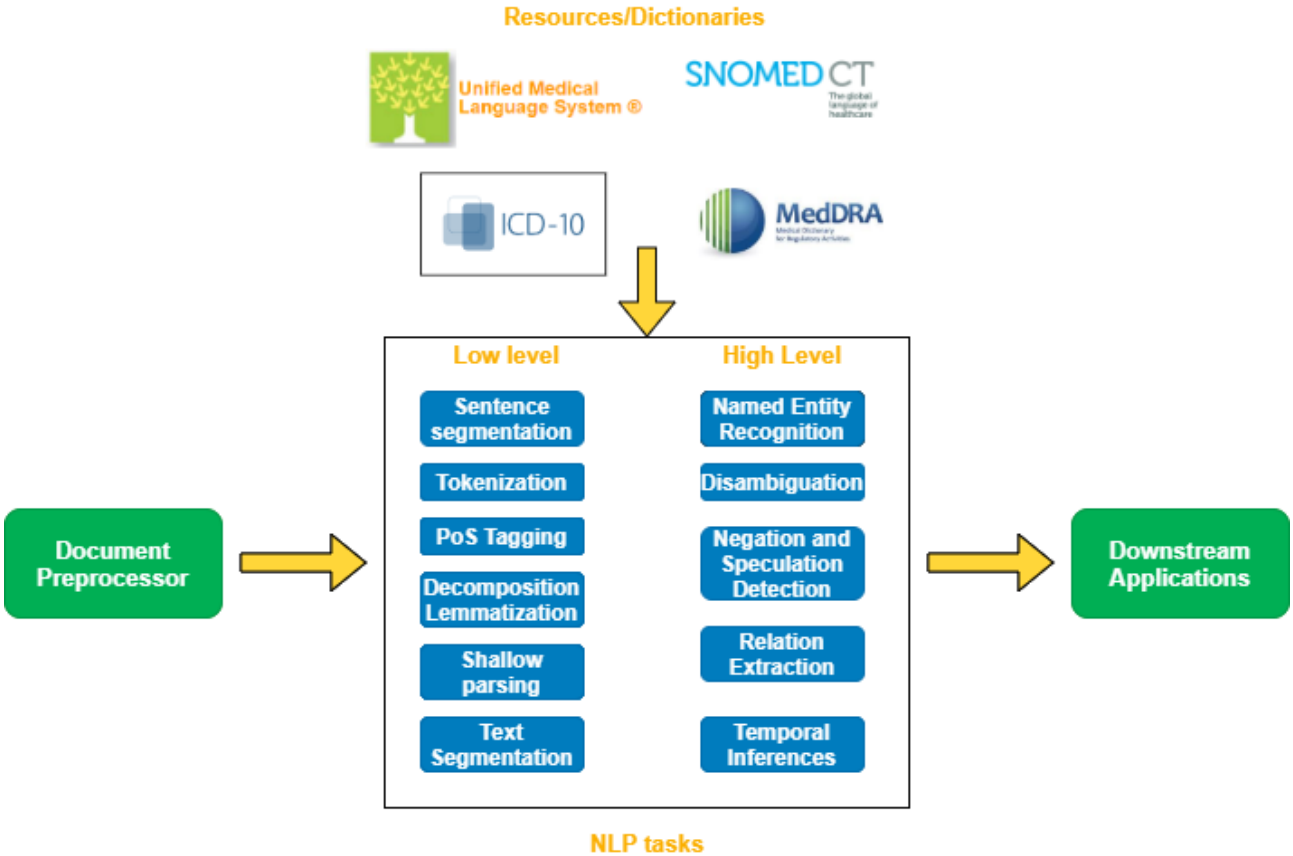


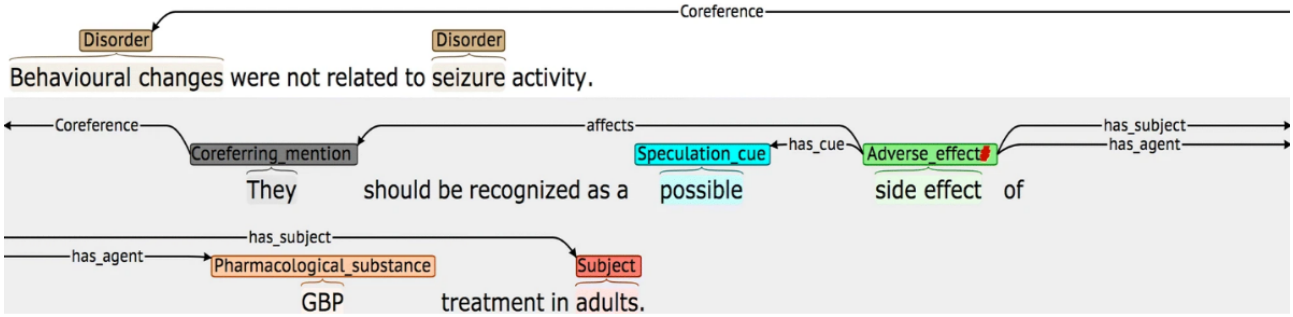
Table 1. Natural language processing low-level tasks.

Task	Objective	Challenge
Sentence segmentation	Detection limit of a sentence.	High use of abbreviations and titles such as “mg” and “Dr” makes this task difficult.
Tokenization	Detection of words and punctuation marks.	Terms combining different types of alphanumeric characters and other signs, such as hyphens, slash, and separators (“10 mg/day” and “N-acetylcysteine”).
Part-of-speech (PoS) tagging	Assigns a PoS tag to a term.	Use of homographs and gerunds.
Decomposition/lemmatization	Word stemming by removing suffixes. Very important for concept normalization.	Many medical terms, such as “nasogastric,” need decomposition to understand the meaning of the term.
Shallow parsing	Identification of the phrases of a sentence.	Inherent complexities from the language (for instance, prepositional attachment).
Text segmentation	Division of the text into relevant parts, such as paragraphs, sections, and others.	In a clinical report, identify sections, such as patient’s history, diagnosis, treatment, etc.

Table 2. Natural language processing high-level tasks.

Task	Objective	Challenge
Named entity recognition	Identification and classification of concepts of interest, such as diseases, drugs, and genes.	Multitoken concepts (“acute rhinovirus bronchitis”) and short concepts (“mg”).
Disambiguation	Identification of the correct sense of a term given a specific context.	A considerable number of abbreviations with several senses, such as Pt (patient/physiotherapy) and LFT (liver function test/lung function test).
Negation and speculation detection	Inferring whether a named entity is present or absent.	They are commonly marked in the clinical narrative by words such as “not” and “without.”
Relation extraction	Identification of relationships between concepts.	Relation between a particular disease and a specific symptom or drug-drug interaction. For example, pharmacodynamic interaction between aspirin and ibuprofen (antagonistic interaction).
Temporal inferences	Given temporal expressions or temporal relationships, inferences are made about probable events in another temporal space.	The most complex task in information extraction. For example, “asbestos exposure and smoking until a particular genetic mutation occurs causes lung cancer in 1-3 years with a probability of 0.2.”

Figure 2. Information extraction pipeline annotation result [2].



Consequently, information extraction tools must address many inherent natural language challenges, such as ambiguity, spelling variations, abbreviations, speculation, and negation. In this work, we address the negation and speculation problems. Negation and speculation expressions are extensively used both in spoken and written communications. Negation converts a proposition represented by a linguistic unit (sentence, phrase, or word) into its opposite, for instance, the existence or absence of medical conditions in a clinical narrative. It is marked by words (such as “not” and “without”), suffixes (such as “less”), or prefixes (such as “a”). Around 10% of the sentences in MEDLINE abstracts include negation phenomena [3]. The BioScope corpus contains more than 20,000 sentences, among which almost 2000 (11.4%) are negated or uncertain sentences [4]. In the general domain, the SFU ReviewSP-NEG corpus is composed of approximately 9455 sentences, among which nearly a third are negated or uncertain sentences [5]. Different works have shown the importance of dealing with negations, for instance, during the analysis of EHRs [1] or in information retrieval tasks on rare disease patient records related to Crohn disease, lupus, and NPHP1 from a clinical data warehouse [6]. In relation to speculation (or modality), both are referred to as expressing facts that are not known with certainty (such as hypotheses and conjectures). There are different types of

expressions that have speculation meanings as follows: modal auxiliaries (must/should/might/may/could be), judgment verbs (suggest), evidential verbs (appear), deductive verbs (conclude), adjectives (likely), adverbs (perhaps), nouns (there is a possibility), conditional words, etc.

These phenomena have a scope, that is, affect a part of the text denoted by the presence of negation or speculation cues. Cues usually occur in the context of some assumption, which works to deny or counteract that assumption. These cues can be single words, simple phrases, or complex verb phrases, which may precede or succeed the words that are within their scope [7]. According to grammar, the scope of the negation or speculation corresponds to the totality of words affected by it. In NLP, negation or speculation cues act as operators that can change the meaning of the words in their scope. Thus, they establish what is a fact and what is not, owing to the ability to affect the truth value of a phrase or sentence [8]. However, negation detection is a complex task owing to the multiple forms in which it can appear as follows: (1) syntactic (ie, negation in sentences, clauses, and phrases that include words expressing negation, such as no/not, never/ever, and nothing), (2) lexical negation (eg, "lack of"), and (3) morphological negation (eg, illegal and impossible) [5].

Negation processing can be divided into two phases. First, keywords/cues indicating negation or speculation are detected, and second, definition of the linguistic scope of these cues is made at the sentence level. In English, negation and speculation detection is a well-studied phenomenon. However, in other languages, such as Spanish, it is an underaddressed and even more complicated task owing to the limited number of annotated corpora and the inherent complexities of the language, such as double negation (eg, the hospital will not allow no more visitors). NegEx [9], one of the most popular rule-based algorithms for negation detection in English, is a simple regular expression-based algorithm that uses negation cue words without considering the semantics of a sentence. Some recent works also exploit this algorithm for negation detection in other languages, such as French, German, and Swedish [10], Swedish [11], and Spanish [12]. Machine learning methods have been applied to cope with the negation detection task, using mainly a conditional random field (CRF) algorithm with dense vector features, such as character or word embedding [13,14]. More recently, deep learning approaches using recurrent neural networks (RNNs), convolutional neuronal networks (CNNs), and encoder-decoder models have also been exploited to solve this task [15-17].

In this work, we addressed the negation and speculation detection tasks as named entity recognition (NER) tasks that solve the identification of cues and scope of this phenomena in a single step. We present two deep learning approaches. First, we implemented two bidirectional long short-term memory (Bi-LSTM) layers with a CRF layer based on the NeuroNER model proposed previously [18]. Specifically, we extended NeuroNER by adding context information to the character and word-level information, such as part-of-speech (PoS) tags and information about overlapping or nested entities. Moreover, in this work, we used several pretrained word-embedding models as follows: (1) word2vec model (Spanish Billion Word Embeddings [19]), which was trained on the 2014 dump of Wikipedia, (2) pretrained word2vec model of word embedding trained with PubMed and PubMed Central articles [20], and (3) sense-disambiguation embedding model [21], where different word senses are represented with different sense vectors. To the best of our knowledge, no previous work has exploited a sense embedding model for the negation detection task. Finally, we implemented the bidirectional encoder representations for transformers (BERT) model with fine tuning using a BERT multilingual pretrained model.

Since the health care system has started adopting cutting-edge technologies, there is a vast amount of data collected mainly in unstructured formats, such as clinical narratives, electronic reports, and EHRs. Therefore, there is a high amount of unstructured data. All of these data involve relevant challenges for information extraction and utilization in the health care domain through various applications of NLP in health care, such as clinical trial matching [22], automated registry reporting, clinical decision support [23], and predicting health care utilization [24]. However, all these applications must deal with inherent NLP challenges, with negation and speculation detection being highly crucial owing to the abuse of negation and speculation particles in the clinical narrative and clinical records.

Work in negation detection has focused on the following two subtasks: (1) cue detection to identify negation terms and (2) scope resolution to determine the coverage of a cue in a phrase or sentence. However, in previous research, negation detection has focused on the straight detection of negated entities [17]. Early negation detection work has relied on rule-based approaches. Rule-based approaches have been shown to be effective in NLP challenges. They use hand-crafted rules based on grammatical patterns and keyword matching. Some token-based systems are NegEx [25], NegFinder [26], NegHunter [27], and NegExpander [28]. DepNeg [29] uses syntactic parsing. Among rule-based approaches, the most used negation detection tool in English is NegEx [13], which employs an exact match to a list of medical entities and negation triggers (eg, "NO history of exposure" and "DENIES any nausea"). NegEx was adapted to address negation detection for other languages, such as Swedish [11], French [30], German [12], and Spanish [31]. Light et al [3] used a hand-crafted list of negation cues to identify speculation sentences in MEDLINE abstracts. Likewise, several biomedical NLP studies have used rules to identify the speculation of extracted information [32-35]. An analysis of a set of Spanish clinical notes from a hospital [36] reported some statistics of several groups of patterns considering the groups defined in the NegEx algorithm [25] as follows: morphologically negates, adverbs, prenegative phrases, postnegative phrases, and pseudonegative phrases. These patterns were applied to the data set, and only the more frequent patterns were inspected (about 100 contexts per pattern). Figure 3 shows the frequencies of the set of negation patterns in the studied corpus, where negation patterns using adverbs ("no," "ni," and "sin") are the more productive patterns, followed by adverbs together with evidential and perception verbs (eg, "no se evidencia" + symptom). There are other negation words, such

as “nadie” (nobody) and “negative” (negative), which do not appear in the data set.

Figure 3. Statistics of the set of negation patterns [30].

Negation pattern	# Contexts
no	736,440
ni	195,144
sin	53,475
no (evidencia|evidencias|sugerencia) de...	3,391
no (se)? (aprecia|revela|siente|ve)...	2,934
nunca	287
tampoco	167
(libre|libres) de	104
(incapacidad|imposibilidad) (de|para)	100
descartando	51
(descartado|descartada|descartar|...|excluido|excluida|excluir) (de|del|por|para)?	42
sin (ningún|ninguna|ningunos|ningunas)	20
nada	16
(imposible|imposibles|impracticable|impracticables|irrealizable|irrealizables)	15
ausencia (completa|total)? de	6
exceptuando	4
desaparición (completa|total)? de	2
omisión de	2
(anulación|anulaciones) de	2
no (solo|necesariamente)	1

Approaches to speculation and negation detection that exploit semisupervised or supervised machine learning models require manually labeled corpora. Medlock [37] used spare word representation features as inputs to classify sentences from biological articles (included in the molecular biology database FlyBase) as certain or uncertain based on semiautomatically collected training examples. Vincze et al [4] extended this approach [37] incorporating n-gram features and a semisupervised selection of keyword features. Morante and Daelemans [38] created a negation cue and scope detection system in biomedical text. This system identifies negation cues using the compressed decision tree (IGTREE) algorithm. It uses a meta-learner based on memory-based learning, a support vector machine, and conditional random fields (CRFs) for determining the scope of the negation. The system was evaluated on the BioScope data set [4], with an F-measure of 98.74% for cue detection and 89.15% for scope determination. Cruz et al [39] focused on negation cue detection in the BioScope corpus using the C4.5 and naive bayes algorithms, with the top F-measure of 86.8% for biomedical articles. Other studies have incorporated POS tag information [40] or different classifiers [41] that followed the two-step approach. Zou et al [42] proposed a tree kernel–based method for scope identification, based on structured syntactic parse features. The system was evaluated on the BioScope corpus, achieving a valuable improvement compared with the state-of-the-art approach, with an F-measure of 92.8% for negation detection.

In previous years, negation and speculation detection was being addressed as a sequence-labeling task. One of the most used algorithms for negation detection is CRF. White et al [43] proposed a CRF-based model with a set of lexical, structural, and syntactic features for scope detection. Kang et al [14] incorporated character-level and word-level dense representations (embeddings) in a CRF algorithm. The best F-measure was 99% for cue detection and 94% for scope detection in Chinese text, and it was concluded that embedding features can help to achieve better performance. Santiso et al [13] proposed a similar system using spare and dense word feature representations and a CRF algorithm to detect only negated entities in Spanish clinical text. The system obtained F-measures of 45.8% and 81.2% for the IxaMed-GS corpus [44] and the IULA corpus [45], respectively.

However, more recently, deep learning approaches are getting more attention, specifically RNNs and CNNs. Lazib et al [46] proposed a hybrid RNN and CNN system with a feature set of word embedding and a syntactic path (the shortest syntactic path from the candidate token to the cue in both constituency and dependency parse trees) to treat this task, and it proved to be very powerful in capturing the potential relationship between the token and the cue. Later, Lazib et al [47] proposed various RNN models to automatically find the part of the sentence affected by a negation cue. They used an automatically extracted word embedding representation of the terms as the only feature. Their Bi-LSTM model achieved an F-measure of 89.38% for the SFU review corpus [48], outperforming all previous hand-encoded feature-based approaches.

Similarly, Fancellu et al [49] used a Bi-LSTM model to solve the task of negation scope detection, and it outperformed the best result of Sem shared task 2012 [50]. Some approaches were proposed to rely on syntactic parse information to automatically extract the most relevant features [51]. Qian et al [15] designed a CNN-based model with probabilistic weighted average pooling to address speculation and negation scope detection. Evaluation of the BioScope corpus showed that their approach achieved substantial improvement. Finally, Bathia et al [17] proposed an end-to-end neural model to jointly extract entities and negations based on the hierarchical encoder-decoder NER model. The

system was evaluated on the 2010 i2b2/VA challenge data set, obtaining an F-score of 90.5% for negation detection.

Motivated by the recent success of machine learning and deep learning approaches in solving various NLP issues, in this paper, we proposed the following two methods: (1) a machine and deep learning model combining two Bi-LSTM networks and a last CRF network, and (2) a BERT model with fine tuning to solve negation and speculation detection issues in multidomain text in both English and Spanish. Negation processing in the Spanish clinical narrative has been little addressed in previous years. Moreover, to the best of our knowledge, sense or context embedding has not been exploited for the negation detection task.

Methods

Overview

We addressed the task of negation and speculation detection as a sequence-labeling task, where we classified each token in a sentence as being part of the negation or speculation cue or negation scope. We have presented the data sets used for training, validating, and evaluating our systems. We have presented a deep network with a preprocessing step, a learning transfer phase, two recurrent neural network layers, and the last layer with a CRF classifier. Moreover, to compare our system performance, we used a baseline model based on a multilayer bidirectional transformer encoder.

NER Architecture

We have address the NER task as a sequence-labeling task. In order to train our model, first, text must be preprocessed to create the input for the deep network. Sentences were split and tokenized using Spacy [52], an open-source library for advanced NLP with support for 26 languages. The output from the previous process was formatted to BRAT format [53]. BRAT is a standoff format where each line represents an annotation (such as entity, relation, and event). We used the information from the BRAT format (example in Figure 4) to annotate each token in a sentence using BMEWO-V extended tag encoding (entity tags used in Table 3), which allowed us to capture information about the sequence of tokens in the sentence.

Figure 4. Examples of annotations in BRAT format over a sentence extracted from the IULA Spanish Clinical Record corpus (translation to English: soft, depressible abdomen, no masses or megalias, not painful).

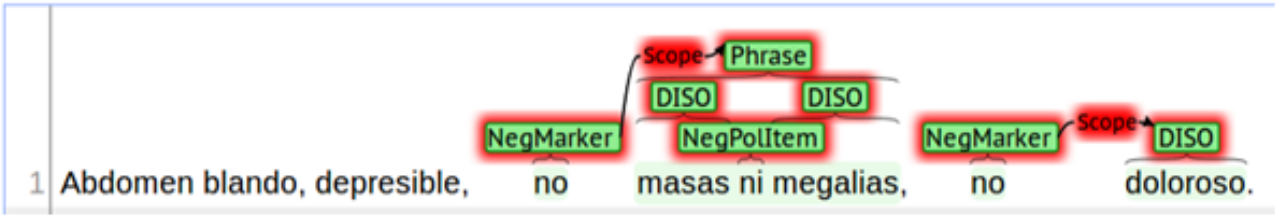


Table 3. Entity tags for BMEWO-V tag encoding in the IULA Spanish Clinical Record corpus.

Entity	Tags
NegMarker[a]	B/M/E/W/V-NegMarker
NegPolItem[b]	B/M/E/W/V-NegPolItem
NegPredMarker[c]	B/M/E/W/V-NegPredMarker
PROC[d]	B/M/E/W/V-PROC
DISO[e]	B/M/E/W/V-DISO
PHRASE[f]	B/M/E/W/V-PHRASE
BODY[g]	B/M/E/W/V-BODY
SUBS[h]	B/M/E/W/V-SUBS
Others	O

[a]NegMarker: no, tampoco, sin [4].

[b]NegPolItem: ni, ninguno, ... [4].

[c]NegPredMarker: negative verbs, nouns, and adjectives [4].

[d]PROC: procedure.

[e]DISO: clinical finding.

[f]PHRASE: nonmedical text spans.

[g]BODY: body structure.

[h]SUBS: substance pharmacological/biological product.

In BMEWO-V encoding, the B tag indicates the start of an entity, the M tag represents the continuity of an entity, the E tag indicates the end of an entity, the W tag indicates a single entity, and the O tag represents other tokens that do not belong to any entity. The V tag allows representation of overlapping entities. BMEWO-V is similar to other previous encodings [54]; however, it also allows the representation of discontinuous entities and overlapping or nested entities. As a result, we obtained the sentences annotated in CoNLL-2003 format (Table 4).

Table 4. Tokens annotated in the ConLL-2003 format.

Token	File	Start offset	End offset	Tag	Tag
Abdomen	negation_iac_3_corr	0	7	O[a]	O
blando	negation_iac_3_corr	8	14	O	O
,	negation_iac_3_corr	14	15	O	O
depresible	negation_iac_3_corr	16	26	O	O
,	negation_iac_3_corr	26	27	O	O
no	negation_iac_3_corr	28	30	W-NegMarker[b]	W-NegMarker
masas	negation_iac_3_corr	31	36	V-Phrase[c]	W-DISO[d]
ni	negation_iac_3_corr	37	39	V-Phrase	W-NegPolIten[e]
megalias	negation_iac_3_corr	40	48	V-Phrase	W-DISO
,	negation_iac_3_corr	48	49	O	O
no	negation_iac_3_corr	50	52	W-NegMarker	W-NegMarker
doloroso	negation_iac_3_corr	53	61	W-DISO	W-DISO
.	negation_iac_3_corr	61	62	O	O

[a]O: other (no entity annotation).

[b]NegMarker: no, tampoco, sin [4].

[c]Phrase: nonmedical text spans.

[d]DISO: clinical finding.

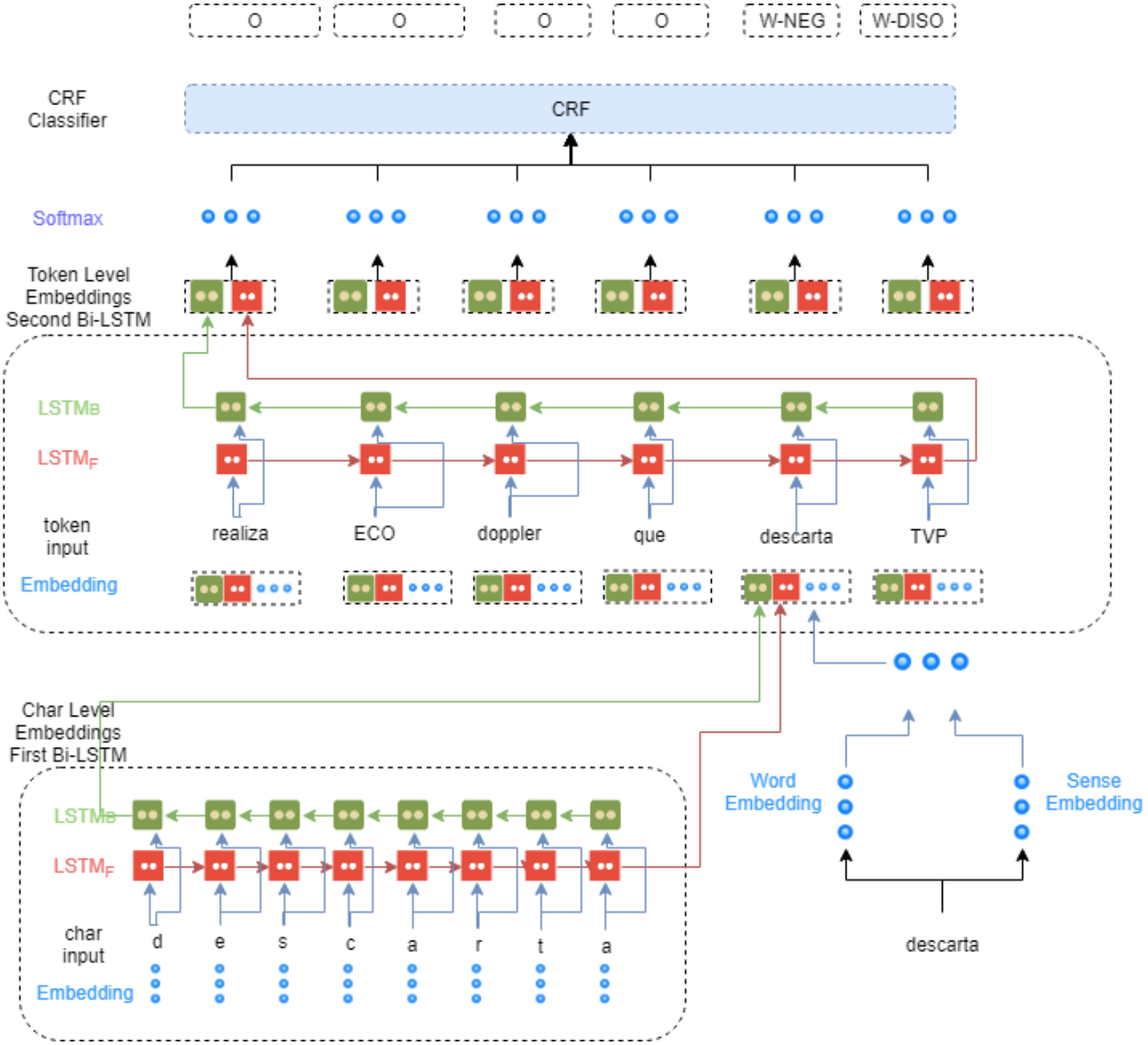
[e]NegPolItem: ni, ninguno, ... [4].

Unlike other detection approaches that detect negation or speculation cues in the first stage and recognize the scope of both of them in the second stage (two-stage system), we proposed a one-stage approach (threaten cue entities within scope entities as nested entities, recognizing both entities [cues and scopes] in a single stage).

Bi-LSTM CRF Model: NeuroNER Extended

Our proposal involves the adaption of a state-of-the-art NER model named NeuroNER [18] based on deep learning to identify entities as negation and speculation. The architecture of our model consists of an initial Bi-LSTM layer for character embedding. In the second layer, we concatenate the output of the first layer with word embedding and sense-disambiguate embedding for the second Bi-LSTM layer. Finally, the last layer uses a CRF to obtain the most suitable labels for each token. An overview of the system architecture can be seen in Figure 5.

Figure 5. The architecture of the hybrid Bi-LSTM CRF model for negation and speculation recognition. Bi-LSTM: bidirectional long short-term memory; CRF: conditional random field.



To facilitate training of our model, we first performed a learning transfer step. Learning transfer aims to perform a task on a data set using knowledge learned from a previous data set [55]. As is shown in many studies, speech recognition [56], sentence classification [57], and NER [58] learning transfer improves generalization of the model, reduces training time on the target data set, and reduces the amount of labeled data needed to obtain high performance. We propose learning transfer as input for our model using the following two different pretrained embedding models: (1) word embedding and (2) sense-disambiguation embedding. Word embedding is an approach to represent words as vectors of real numbers, which has gained much popularity among the NLP community because it is able to capture syntactic and semantic information among words.

Although word embedding models are able to capture syntactic and semantic information, other linguistic information, such as morphological information, orthographic transcription, and POS tags, are not exploited in these models. According to a previous report [59], the use of character embedding improves learning for specific domains and is useful for morphologically rich languages (as is the case of the Spanish language). For this reason, we decided to consider the character embedding representation in our system to obtain morphological and orthographic information from words. We used a 25-feature vector to represent each character. In this way, tokens in sentences are represented by their corresponding character embeddings, which are the inputs for our Bi-LSTM network.

We used the Spanish Billion Words model [19], which is a pretrained model of word embedding trained on different text corpora written in Spanish (such as Ancora Corpus [60] and Wikipedia). Furthermore, we used a pretrained word embedding model induced from PubMed and PubMed Central texts and their combination using the word2vec tool [20]. PubMed text considers abstracts of scientific articles as of the end of September 2013, with a total of 22 million records. PubMed Central text considers full-text articles as of the end of September 2013 and constitutes a total of 600,000 articles. These resources were derived from the combination of abstracts from PubMed and full-text documents from the PubMed Central

Open Access subset written in English. We also experimented with Google word2vec embedding [61] trained on 100 billion words from Google News [62].

We also integrated the sense2vec [21] model, which provides multiple embeddings for each word based on the sense of the word. This model is able to analyze the context of a word and then assign a more adequate vector for the meaning of the word. In particular, we used the Reddit Vector, a pretrained model of sense-disambiguation representation vectors introduced previously [21]. This model was trained on a collection of comments published on Reddit (corresponding to the year 2015). The details of pretrained embedding models are shown in Table 5.

Table 5. Details of the pretrained embedding models.

Detail	Spanish Billion Words	Google News	PubMed and PubMed Central	Reddit
Language	Spanish	English	English	Multilingual
Corpus size	1.5 billion	100 billion	6 trillion	2 billion
Vocab size	1 million	3 million	2 million	1 million
Array size	300	300	200	128
Algorithm	Skip-gram BOW	Skip-gram BOW	Skip-gram BOW	Sense2Vec

The output of the first layer was concatenated with word embedding and sense-disambiguation embedding obtained from pretrained models for each token in a given input sentence. This concatenation of features was the input for the second Bi-LSTM layer. The goal of the second layer was to obtain a sequence of probabilities corresponding to each label of the BMEWO-V encoding format. In this way, for each input token, this layer returned six probabilities (one for each tag in BMEWO-V). The final tag should be with the highest probability for each token.

To improve the accuracy of predictions, we also used a CRF [63] model, which takes as input the label probability for each independent token from the previous layer and obtains the most probable sequence of predicted labels based on the correlations between labels and their context. Handling independent labels for each word shows sequence limitations. For example, considering the drug sequence-labeling problem, an "I-NEGATION" tag cannot be found before a "B-NEGATION" tag or an "I- NEGATION" tag cannot be found after a "B-NEGATION" tag. Finally, once tokens have been annotated with their corresponding labels in the BMEWO-V encoding format, the entity mentions must be transformed into the BRAT format. V tags, which identify nested or overlapping entities, are generated as new annotations within the scope of other mentions.

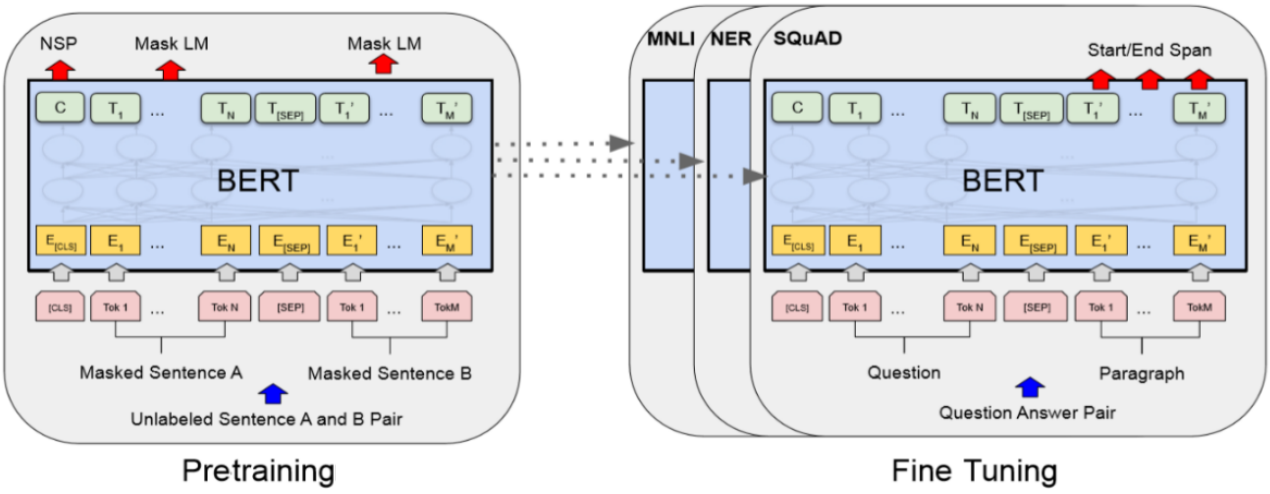
Multilayer Bidirectional Transformer Encoder: BERT

The use of word representations from pretrained unsupervised methods is a crucial step in NER pipelines. Previous models, such as word2vec [62], Glove [64], and FastText [65], focused on context-independent word representations or word embedding. However, in the last few years, models have focused on learning context-dependent word representations, such as ELMo [66], CoVe [67], and the state-of-the-art BERT model [68], and then fine tuning these pretrained models on downstream tasks.

BERT is a context-dependent word representation model that is based on a masked language model and is pretrained using the transformer architecture [69]. BERT replaces the sequential nature of language modeling. Previous models, such as RNN (LSTM & GRU), combine two unidirectional layers (ie, Bi-LSTM), and as a replacement for the sequential approach, the BERT model employs a much faster attention-based approach. BERT is pretrained in the following two unsupervised tasks: (1) masked language modeling that predicts randomly masked words in a sequence and hence can be used for learning bidirectional representations by jointly conditioning both left and right contexts in all layers and (2) next sentence prediction to train a model that understands sentence relationships. A previous report [70] provides a detailed description of BERT.

Owing to the benefits of the BERT model, we adopted a pretrained BERT model with 12 transformer layers (12 layers, 768 hidden, 12 heads, 110 million parameters) and an output layer with SoftMax to perform the NER task. The transformer layer has the following two sublayers: a multihead self-attention mechanism, and a position-wise, fully connected, feed-forward network, followed by a normalization layer. An overview of the BERT architecture is presented in Figure 6.

Figure 6. BERT pretraining and fine-tuning architecture overview [62]. BERT: bidirectional encoder representations from transformers.



Data Sets

The proposed systems are evaluated for the following three data sets: (1) the BioScope corpus introduced in the CoNLL-2010 Shared Task [7] for the detection of speculation cues and their linguistic scope [4], (2) the SFU ReviewSP-NEG corpus used in Task 2 in the 2018 edition of the Workshop on Negation in Spanish (NEGES 2018) [71], and (3) the IULA Spanish Clinical Record corpus [72]. Therefore, we evaluated the proposed system in two different languages (English and Spanish) and different text types (clinical narrative, biomedical literature, and user reviews). Spanish, contrary to other languages such as English, does not have enough corpora, data sets, pretrained models, and resources. Furthermore, research on Spanish negation and speculation detection is insufficient, and this is even more in the biomedical domain. Being aware of this setback, in this particular study, we used the scarce Spanish resources available.

The BioScope corpus is a widely used and freely available resource consisting of medical and biological texts written in English annotated with speculative and negative cues and their scopes. BioScope includes the following three different subcorpora: (1) clinical free texts (clinical radiology records), (2) full biological papers from Flybase and the BMC Bioinformatics website, and (3) biological abstracts from the GENIA corpus [73]. The corpus statistics are shown in Table 6.

Table 6. BioScope corpus details.

Variable	Abstracts	Full papers	Clinical narratives
Total			
Number of documents	1954	9	1273
Number of sentences	6383	2624	11,872
Speculation			
Number of sentences	2101	519	855
Number of scopes	2659	672	1112
Negation			
Number of sentences	1597	339	865
Number of scopes	1719	376	870

Concerning negation and speculation, the CoNNLL-2010 Shared Tasks divide the BioScope data set into three subtasks. The first two subtasks are as follows: (1) Task 1B sentence speculation detection for biological abstracts and full articles and (2) Task 1W sentence speculation detection for paragraphs from Wikipedia, possibly containing weasel information. Both tasks consist of a binary classification problem for detecting speculation cues and speculation at the sentence level and the final task (Task 2), which aims the in-sentence hedge scope to distinguish uncertain information from facts in general and biomedical domains. The BioScope corpus includes a different data set for each subtask. Detailed information about these data sets can be seen in Table 7.

Table 7. BioScope subtask data sets.

Task and subset	Number of documents	Number of sentences	Number of cues	Number of scopes
Task 1B				
Training	966	10,806	2540	N/A[a]
Validation	316	3735	836	N/A
Testing	15	5003	N/A	N/A
Task 1W				
Training	1646	8343	2363	N/A
Validation	540	2768	770	N/A
Testing	2346	9634	N/A	N/A
Task 2				
Training	966	11,009	2556	2519
Validation	316	3533	820	808
Testing	15	5003	N/A	N/A

[a]N/A: not applicable.

The IULA Spanish Clinical Record corpus consists of 300 manually annotated and anonymized clinical records from several services of one of the main hospitals in Barcelona. These clinical records are written in Spanish. The corpus contains annotations on syntactic and lexical negation markers and their respective scopes. Morphological negation was excluded. There are 3194 sentences, and of these, 1093 (34.22%) were annotated with negation cues. IULA Spanish Clinical Record corpus details and its entity distribution can be found in Tables 8 and 9, respectively.

Table 8. IULA Spanish Clinical Record corpus details.

Item	Clinical narrative, n
Documents	300
Sentences	3194
Annotated sentences	1093
Negated entities	1456

Table 9. IULA Spanish Clinical Record corpus entity distribution.

Entity	Total, n
NegMarker[a]	1007
NegPredMarker[b]	86
NegPolItem[c]	114
BODY[d]	7
SUBS[e]	14
DISO[f]	1064
PROC[g]	93
Phrase[h]	278

[a]NegMarker: no, tampoco, sin [4].

[b]NegPredMarker: negative verbs, nouns, and adjectives [4].

[c]NegPolItem: ni, ninguno, ... [4].

[d]BODY: body structure.

[e]SUBS: substance pharmacological/biological product.

[f]DISO: clinical finding.

[g]PROC: procedure.

[h]PHRASE: nonmedical text spans.

To the best of our knowledge, the IULA Spanish Clinical Record corpus has not been used in any task or challenge. Therefore, we randomly split the data set into training, validation, and testing data sets. Details about the data sets can be seen in Table 10.

Table 10. IULA Spanish Clinical Record data sets.

Subset	Number of sentences	Number of entities
Training	1774	2839
Validation	701	924
Testing	719	920

The SFU ReviewSP-NEG corpus is the first Spanish corpus that includes event negation as part of the annotation scheme, as well as the annotation of discontinuous negation markers. Moreover, it is the first corpus where the negation scope is annotated. The corpus also includes syntactic negation, scope, and focus. However, neither lexical nor morphological negation is included. Annotations on the event and on how negation affects the polarity of the words within its scope are also included. The Spanish SFU Review corpus consists of 400 reviews from the Ciao website [74] from the following eight different domains: cars, hotels, washing machines, books, phones, music, computers, and movies. It is composed of 9455 sentences, and of these, 3022 (31.97%) contain at least one negation cue. SFU ReviewSP-NEG corpus text distribution can be found in Table 11. The SFU ReviewSP-NEG corpus was used in Task 2 of NEGES 2018 for identifying negation cues in Spanish. The data set was randomly divided into training, validation, and testing data sets. Details about the data sets can be seen in Table 12.

Table 11. SFU ReviewSP-NEG corpus details.

Item	Reviews, n
Comments	400
Sentences	9455
Annotated sentences	3022
Negated entities	3941

Table 12. SFU ReviewSP-NEG data sets.

Subset	Reviews, n	Sentences, n	Negated entities, n
Training	264	1774	606
Validation	56	701	209
Testing	80	719	285

Negation cues and scope are annotated in each corpus (the IULA corpus does not include the subject within the scope). Regarding the negation in coordinated structures, the corpora also show differences. In the SFU ReviewSP-NEG corpus, a distinction is made between the coordinated negative structures. Each negation cue is independent and has its own scope. Moreover, the scopes of those negative structures with discontinuous negation cues consider the whole coordination. The IULA Spanish Clinical Record always includes coordination within the scope. Furthermore, we found that double negation (eg, "No síntoma de disnea NI dolor torácico" [No symptoms of dyspnea or chest pain]) and negation locutions, which are multiword expressions that express negation (eg, "con AUSENCIA DE vasoespasmo" [with absence of vasospasm]) were only addressed in the SFU ReviewSP-NEG corpus. Additionally, speculative expressions and uncertain annotations (eg, "Earths and clays MAY have provided prehistoric peoples") were only addressed in the BioScope corpus.

Results

We evaluated the negation detection system using the training, validation, and testing data sets provided by the task organizers for the CoNLL-2010 Shared Task (BioScope) and for Task 2 of NEGES 2018 (SFU ReviewSP-NEG). The IULA Spanish Clinical Record corpus has not been previously applied to any task or competition. Therefore, we split the corpus randomly into training and testing data sets to evaluate the proposal in the clinical domain.

The Bi-LSTM CRF model was trained using available pretrained word and sense embedding models on general and biomedical domains for Spanish, English, and multilingual texts. We evaluated the use of multidomain and multilanguage pretrained embedding models (general domain word and sense embeddings and multilanguage NLP tools) on the BioScope Task 1W data sets (biomedical domain and English text), with a precision, recall, and F-score of 86.2%, 87%, and 86.6%, respectively. Based on our experiments, we found that the use of specific domain (biomedical) and specific language (English) embeddings highly improved the negation and speculation detection task (Table 13). Moreover, to evaluate the performance impact, we evaluated each of our proposed features and made comparisons with base NeuroNER implementation with PubMed and PubMed Central word embeddings on the BioScope Task 1W test data set. As shown in Table 14, sense feature representation and the BIOES-V tag encoding format improved each token representation, which implies that features play different roles in capturing token-level features for NER tasks, thus making improvements in their combination.

Table 13. Pretrained word embedding model evaluation on the BioScope Task 1W test data set.

Name–embedding	Precision (%)	Recall (%)	F-score (%)
NeuroNER–Google News	78.3	80.4	79.3
NeuroNER–PubMed and PubMed Central	80.8	82.1	81.4
NeuroNER Extended–Google News	80.2	83.2	81.7
NeuroNER Extended–PubMed and PubMed Central	86.2	87.0	86.6

Table 14. Feature evaluation on the BioScope Task 1W test data set.

Name–feature	Precision (%)	Recall (%)	F-score (%)
NeuroNER–Base	78.3	80.4	81.4
NeuroNER–Sense	84.7	86.2	85.4
NeuroNER–BIOES-V	81.7	83.5	82.6
NeuroNER–Sense and BIOES-V	86.2	87.0	86.6

Moreover, we used the pretrained BERT multilingual general domain model with 12 transformer layers (12 layers, 768 hidden, 12 heads, 110 million parameters) trained on the general domain Wikipedia and Bookcorpus corpora, and fine-tuned for NER using a single output layer based on the representations from its last layer to compute only token-level BIOES-V probabilities. BERT directly learns WordPiece embeddings during the pretraining and fine-tuning steps.

Precision, recall, and the F-score were used to evaluate the performance of our system. The parameters of the sets and the hyperparameters for our Bi-LSTM CRF model are summarized in Table 15. The hyperparameters were optimized on each validation data set.

Table 15. NeuroNER system hyperparameters for each task.

Parameter	BioScope	IULA	SFU ReviewSP-NEG
Language	English	Spanish	Spanish
Pretrained word embedding	PubMed and PubMed Central + Reddit	Spanish Billion Words + Reddit	Spanish Billion Words + Reddit
Sense-disambiguation embedding dimension	128	128	128
Word embedding dimension	200	300	300
Character embedding dimension	50	50	50
Hidden layers dimension (for each LSTM)	100	100	100
Learning method	Stochastic gradient descent	Stochastic gradient descent	Stochastic gradient descent
Dropout rate	0.5	0.5	0.5
Learning rate	0.005	0.005	0.005
Epochs	100	100	100

The CoNLL-2010 Shared Task [75] considers two different evaluation criteria. Task 1 is made at the sentence level, and cue annotations in the sentence are not considered. However, it is optionally evaluated. The F-measure of the speculation class is employed as the chief evaluation metric. Task 2 involves the annotation of "cue" + "xcope" tags in sentences. The scope-level F-measure is used as the chief metric where true positives are scopes that match the gold standard clue words and gold standard scope boundaries assigned to the clue words.

Tables 16 to 20 compare the results obtained by the participating systems in the CoNLL-2010 Shared Task and our deep learning approach using pretrained embedding models and the BMEWO-V encoding format. Our extended version of NeuroNER achieved similar results to the best work presented in this task. In particular, our system achieved the highest precision (83.2%), with lower recall.

For subtask 1 (identification speculation at the sentence level and cue annotations), our system obtained the top F-score for speculation and cue detection (see Tables 16 to 18).

Table 16. Task 1B Wikipedia sentence-level speculation detection (BioScope).

Name	Precision (%)	Recall (%)	F-score (%)
Georgescul [76]	72.0	51.7	60.2
Ji et al [77]	62.7	55.3	58.7
Chen et al [78]	68.0	49.7	57.4
BERT	83.7	48.5	61.4
NeuroNER Extended	83.2	41.0	54.9

Table 17. Task 1B Wikipedia cue-level detection (BioScope).

Name	Precision (%)	Recall (%)	F-score (%)
Tang et al [79]	63.0	25.7	36.5
Li et al [80]	76.1	21.6	33.7
Özgür et al [81]	28.9	14.7	19.5
BERT	63.7	33.2	43.6
NeuroNER Extended	63.0	25.7	36.5

Table 18. Task 1W biological sentence-level speculation detection (BioScope).

Name	Precision (%)	Recall (%)	F-score (%)
Tang et al [79]	85.0	87.7	86.4
Zhou et al [82]	86.5	85.1	85.8
Li et al [80]	90.4	81.0	85.4
BERT	85.5	87.3	86.4
NeuroNER Extended	86.2	87.0	86.6

Table 19. Task 1W biological cue-level detection (BioScope).

Name	Precision (%)	Recall (%)	F-score (%)
Tang et al [79]	81.7	81.0	81.3
Zhou et al [82]	83.1	78.8	80.9
Li et al [80]	87.4	73.4	79.8
BERT	80.7	79.5	80.1
NeuroNER Extended	81.4	79.2	80.3

Table 20. Task 2 cue-level detection and scope determination (BioScope).

Name	Precision (%)	Recall (%)	F-score (%)
Morante et al [83]	59.6	55.2	57.3
Rei et al [6]	56.7	54.6	55.6
Velldal et al [84]	56.7	54.0	55.3
BERT	46.1	55.6	50.4
NeuroNER Extended	50.4	40.3	44.8

Table 21 shows the results for the IULA corpus. Furthermore, we compared our results with the work presented previously [85]. We used the evaluation criteria presented in this work; however, the subsets were different. As can be seen, our system outperformed the results obtained previously [85], with a difference of nearly 4 points for the F-measure.

Table 21. Results of cue level and scope detection for the IULA Clinical Record data set.

Name	Precision (%)	Recall (%)	F-score (%)
Santiso et al [85]	79.1	83.5	81.2
BERT	77.8	84.3	80.8
NeuroNER Extended	84.2	85.9	85.0

The NEGES 2018 Task 2 negation cue detection uses the evaluation script proposed in the SEM 2012 Shared Task–Resolving the Scope and Focus of Negation [50]. Table 22 shows the results for the different domains included in the data set. It can be observed that the F-score was always over 80%. We compared our results with the participating systems presented in this task. A detailed description of the evaluation has been provided previously [71]. As can be seen in Table 23, our system outperformed the rest of the participating systems.

Furthermore, we compared NeuroNER Extended and BERT implementations in terms of resources and time consumption on the IULA Clinical Record training and validation subsets. As shown in Table 24, the training time was slightly higher in NeuroNER Extended. However, training implies the generation of character and token level embeddings, unlike the BERT implementation that obtains word vector representations directly from the pretrained model. In terms of hardware resource consumption, we found that BERT implementation had a high use of resources, especially RAM and GPU.

Table 22. NeuroNER Extended results of negation detection for the SFU ReviewSP-NEG data set.

Domain	Precision (%)	Recall (%)	F-score (%)
Cars	87.5	74.47	80.46
Hotels	95.92	77.05	85.46
Washing machines	94.44	75.56	83.95
Books	95.45	87.5	91.3
Phones	97.06	90.83	93.84
Music	92.31	92.31	92.31
Computers	95.45	80.77	87.5
Movies	95.88	84.55	89.86

Table 23. Results of negation cues and scope detection for the SFU ReviewSP-NEG data set.

Name	Precision (%)	Recall (%)	F-score (%)
Fabregat et al [86]	79.5	59.6	68.0
Loharja et al [87]	79.1	83.5	81.2
BERT	92.6	90.8	91.7
NeuroNER Extended	94.3	82.9	88.1

Table 24. Training parameters for the deep learning models.

Training parameter	Specifications	NeuroNER Extended	BERT
CPU	Intel Core i7 7700 at 3.60 GHz	50%	30%
RAM	16 GB DDR4	40%	80%
GPU	GeForce RTX 2060 SUPER 16 RAM	40%	80%
Training time	Minutes	15 min	13 min

Discussion

Principal Findings

We used different pretrained models and investigated their effects on performance. For NeuroNER Extended, we used general and domain-specific pretrained word embedding models, and likewise, we used pretrained multilanguage and language-specific models. We found that the use of specific domain (biomedical) and specific language pretrained models highly improved the negation and speculation detection. Moreover, to the best of our knowledge, there is no pretrained biomedical Spanish model for context-dependent word representations (pretrained BERT). The low performance of the BERT model is mainly attributed to the use of a general domain and multilingual pretrained model. However, the BERT model outperformed the NeuroNER Extended model and other state-of-the-art approaches in general domain data sets, such as SFU ReviewSP-NEG, and the specific domain BioScope (Task 1B data set corpus obtained from Wikipedia text).

Moreover, we presented the analysis of the most frequent false negatives and false positives for negation and speculation cues and scope detection. Negation and speculation cues, such as “would,” “apenas” (“barely”), “ni” (“neither” or “nor”), “except,” “could,” “idea,” “notion,” and “may,” are half of the time labeled as negation and speculation cues. This ambiguity led our system to classify some tokens as false positive or inversely as false negative, causing a drop in performance. Furthermore, some multitoken negation and speculation cues, such as “ni siquiera” (“not even”), “ni tan siquiera” (“not even”), “ni si quiera” (“not even”), and “en ningún momento” (“not at any moment”), are sometimes labeled as a single token word (ie, “ni_siquiera,” “ni_tan_siquiera,” “ni_si_quiera,” and “en_ningún_momento”), and some others are labeled as multitoken cues. Long multitoken negation and speculation cues, such as “remains to be determined” and “raising the intriguing possibility,” are not detected or partially matched. This proves that shorter sentences, with shorter scopes and shorter negation and speculation cues, are easier to process. A longer sentence has a more complex syntactic structure and is tougher to be processed by the system. It should be noted that clinical text is undoubtedly distinct from biomedical text. It is characterized by short sentences (usually phrases) and misspellings, with abuse of negation particles and abbreviations, among other important features.

Furthermore, in the context of real medical applications, negation and speculation detection is a fundamental task in any information extraction system. For instance, in cohort selections for a clinical trial, patients with a specific condition are required, and it is essential to know if a term representing a disease or any other feature is negated or not in a clinical note in order to get the right answer to the query (Is the variable V valid for

patient P?). An additional example would be the detection of adverse drug reactions, that is, the extraction of causal relations between drugs and diseases. It is a crucial step to discard the absence of adverse drug reactions early and thus prevent medical applications from analyzing them or providing wrong information.

Conclusions

In this work, we proposed a system for the detection of negated entities, negation cues, negation scope, and speculation in multidomain text in English and Spanish. We addressed the speculation and negation detection task as a sequence-labeling task. Although previous studies have already applied deep learning to this task, our approach is the first to exploit sense embedding as the input of the deep network. In a sense embedding model, each meaning word is represented with a different vector. Therefore, sense embedding models can help to solve ambiguity, which is one of the most critical challenges in NLP.

Our experiments show that the use of dense representation of words (word-level embedding, character-level embedding, and sense embedding) provides good results in detecting negated entities, negation cues, and negation scope determination. Compared with previous work, our system achieved an F-score performance of over 85%, outperforming most current state-of-the-art methods for negation and speculation detection. Moreover, our work is one of the few that addressed the task for Spanish text and different domains using context-independent and context-dependent pretrained models.

In future work, we plan to test whether other supervised classifiers, such as Markov random fields and optimum path forest, would obtain more benefits from dense vector representation. That is to say, we would use the same continuous representations with the Markov random fields and optimum path forest classifiers. Moreover, we plan to train word context-dependent and independent embeddings obtained from multiple Spanish biomedical corpora to enhance word representations using different models, such as FastText and pretrained BERT. Furthermore, we plan to explore different models for embeddings that combine in a single representation not only words but also semantic information contained in domain-specific resources, such as UMLS [88] and SNOMED-CT [89].

Acknowledgments

This work was supported by the Research Program of the Ministry of Economy and Competitiveness, Government of Spain (DeepEMR Project TIN2017-87548-C2-1-R).

Conflicts of Interest

None declared.

References

1. Dalianis H. Clinical Text Mining. Cham, Switzerland: Springer; 2018.
2. Thompson P, Daikou S, Ueno K, Batista-Navarro R, Tsujii J, Ananiadou S. Annotation and detection of drug effects in text for pharmacovigilance. J Cheminform 2018 Aug 13;10(1):37 [FREE Full text] [doi: 10.1186/s13321-018-0290-y] [Medline: 30105604]
3. Light M, Qiu XY, Srinivasan P. The Language of Bioscience: Facts, Speculations, and Statements In Between. ACL Anthology. 2004. URL: https://www.aclweb.org/anthology/W04-3103/ [accessed 2020-11-22]
4. Vincze V, Szarvas G, Farkas R, Móra G, Csirik J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics 2008 Nov 19;9 Suppl 11:S9 [FREE Full text] [doi: 10.1186/1471-2105-9-S11-S9] [Medline: 19025695]
5. Jiménez-Zafra SM, Morante R, Martin M, Ureña-López LA. A review of Spanish corpora annotated with negation. ACL Anthology. 2018. URL: https://www.aclweb.org/anthology/C18-1078/ [accessed 2020-11-22]
6. Rei M, Briscoe T. Combining Manual Rules and Supervised Learning for Hedge Cue and Scope Detection. ACL Anthology. URL: https://www.aclweb.org/anthology/W10-3008 [accessed 2020-11-22]
7. Farkas R, Vincze V, Móra G, Csirik J, Szarvas G. The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. ACL Anthology. 2010. URL: https://www.aclweb.org/anthology/W10-3001/ [accessed 2020-11-22]
8. Kato Y. A natural history of negation. By LAURENCE R. HORN. Chicago: The University of Chicago Press, 1989. Pp. xxii, 637. EL 1991 Jul 01;8:190-208. [doi: 10.9793/elsj1984.8.190]
9. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001 Oct;34(5):301-310 [FREE Full text] [doi: 10.1006/jbin.2001.1029] [Medline: 12123149]
10. Chapman WW, Hillert D, Velupillai S, Kvist M, Skeppstedt M, Chapman BE, et al. Extending the NegEx lexicon for multiple languages. Stud Health Technol Inform 2013;192:677-681 [FREE Full text] [Medline: 23920642]
11. Skeppstedt M. Negation detection in Swedish clinical text: An adaption of NegEx to Swedish. J Biomed Semantics 2011;2 Suppl 3:S3 [FREE Full text] [doi: 10.1186/2041-1480-2-S3-S3] [Medline: 21992616]

12. Cotik V, Stricker V, Vivaldi J, Rodriguez H. Syntactic methods for negation detection in radiology reports in Spanish. ACL Anthology. 2016. URL: https://www.aclweb.org/anthology/W16-2921/ [accessed 2020-11-22]
13. Santiso S, Casillas A, Pérez A, Oronoz M. Word embeddings for negation detection in health records written in Spanish. Soft Comput 2018 Nov 23;23(21):10969-10975. [doi: 10.1007/s00500-018-3650-7]
14. Kang T, Zhang S, Xu N, Wen D, Zhang X, Lei J. Detecting negation and scope in Chinese clinical notes using character and word embedding. Comput Methods Programs Biomed 2017 Mar;140:53-59. [doi: 10.1016/j.cmpb.2016.11.009] [Medline: 28254090]
15. Qian Z, Li P, Zhu Q, Zhou G, Luo Z, Luo W. Speculation and Negation Scope Detection via Convolutional Neural Networks. ACL Anthology. 2016. URL: https://www.aclweb.org/anthology/D16-1078/ [accessed 2020-11-22]
16. Lazib L, Qin B, Zhao Y, Zhang W, Liu T. A syntactic path-based hybrid neural network for negation scope detection. Front. Comput. Sci 2018 Aug 2;14(1):84-94. [doi: 10.1007/s11704-018-7368-6]
17. Bhatia P, Busra Celikkaya E, Khalilia M. End-to-End Joint Entity Extraction and Negation Detection for Clinical Text. In: Shaban-Nejad A, Michalowski M, editors. Precision Health and Medicine. W3PHAI 2019. Studies in Computational Intelligence, vol 843. Cham: Springer; 2019:139-148.
18. Dernoncourt F, Lee JY, Szolovits P. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. ACL Anthology. 2017. URL: https://www.aclweb.org/anthology/D17-2017/ [accessed 2020-11-22]
19. Cardellino C. Spanish Billion Words Corpus and Embeddings. Cristian Cardellino. 2016 Mar. URL: https://crscardellino.github.io/SBWCE/ [accessed 2020-11-22]
20. Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S. Distributional Semantics Resources for Biomedical Text Processing. In: Proceedings of LBM 2013. 2013 Presented at: 5th International Symposium on Languages in Biology and Medicine; December 12-13, 2013; Tokyo, Japan p. 39-44.
21. Trask A, Michalak P, Liu J. sense2vec - A Fast and Accurate Method for Word Sense Disambiguation In Neural Word Embeddings. arXiv. 2015 Nov 19. URL: https://arxiv.org/abs/1511.06388 [accessed 2020-11-22]
22. Helgeson J, Rammage M, Urman A, Roebuck MC, Coverdill S, Pomerleau K, et al. Clinical performance pilot using cognitive computing for clinical trial matching at Mayo Clinic. JCO 2018 May 20;36(15_suppl):e18598-e18598. [doi: 10.1200/jco.2018.36.15_suppl.e18598]
23. Imler TD, Morea J, Imperiale TF. Clinical decision support with natural language processing facilitates determination of colonoscopy surveillance intervals. Clin Gastroenterol Hepatol 2014 Jul;12(7):1130-1136. [doi: 10.1016/j.cgh.2013.11.025] [Medline: 24316106]
24. Agarwal V, Zhang L, Zhu J, Fang S, Cheng T, Hong C, et al. Impact of Predicting Health Care Utilization Via Web Search Behavior: A Data-Driven Analysis. J Med Internet Res 2016 Sep 21;18(9):e251 [FREE Full text] [doi: 10.2196/jmir.6240] [Medline: 27655225]
25. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001 Oct;34(5):301-310 [FREE Full text] [doi: 10.1006/jbin.2001.1029] [Medline: 12123149]
26. Mutalik PG, Deshpande A, Nadkarni PM. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. J Am Med Inform Assoc 2001 Nov 01;8(6):598-609 [FREE Full text] [doi: 10.1136/jamia.2001.0080598] [Medline: 11687566]
27. Gindl S, Kaiser K, Miksch S. Syntactical negation detection in clinical practice guidelines. Stud Health Technol Inform 2008;136:187-192 [FREE Full text] [Medline: 18487729]
28. Aronow DB, Fangfang F, Croft WB. Ad hoc classification of radiology reports. J Am Med Inform Assoc 1999 Sep 01;6(5):393-411 [FREE Full text] [doi: 10.1136/jamia.1999.0060393] [Medline: 10495099]
29. Lapponi E, Read J, Øvrelid L. Representing and Resolving Negation for Sentiment Analysis. In: 2012 IEEE 12th International Conference on Data Mining Workshops. 2012 Presented at: 12th International Conference on Data Mining Workshops; December 10, 2012; Brussels, Belgium p. 687-692. [doi: 10.1109/ICDMW.2012.23]
30. Deléger L, Grouin C. Detecting negation of medical problems in French clinical notes. In: IHI '12: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. 2012 Presented at: 2nd ACM SIGHIT International Health Informatics Symposium; January 2012; Miami, Florida p. 697-702. [doi: 10.1145/2110363.2110443]
31. Costumero R, Lopez F, Gonzalo-Martín C, Millan M, Menasalvas E. An Approach to Detect Negation on Medical Documents in Spanish. In: Śl zak D, Tan AH, Peters JF, Schwabe L, editors. Brain Informatics and Health. BIH 2014. Lecture Notes in Computer Science, vol 8609. Cham: Springer; 2014:366-375.
32. Friedman C, Alderson PO, Austin JHM, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994 Mar 01;1(2):161-174 [FREE Full text] [doi: 10.1136/jamia.1994.95236146] [Medline: 7719797]
33. Chapman W, Dowling J, Chu D. ConText: An Algorithm for Identifying Contextual Features from Clinical Text. ACL Anthology. 2007. URL: https://www.aclweb.org/anthology/W07-1011/ [accessed 2020-11-22]
34. Aramaki E, Miura Y, Tonoike M, Ohkuma T, Mashuichi H, Ohe K. TEXT2TABLE: Medical Text Summarization System Based on Named Entity Recognition and Modality Identification. ACL Anthology. 2009. URL: https://www.aclweb.org/anthology/W09-1324/ [accessed 2020-11-22]

35. Conway M, Doan S, Collier N. Using Hedges to Enhance a Disease Outbreak Report Text Mining System. In: BioNLP '09: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing. 2009 Presented at: Workshop on Current Trends in Biomedical Natural Language Processing; June 2009; Boulder, Colorado p. 142-143. [doi: 10.3115/1572364.1572384]
36. Campillos Llanos L, Martinez P, Segura-Bedmar I. A preliminary analysis of negation in a Spanish clinical records dataset. In: Actas del Taller de NEGación en Español. NEGES-2017. 2017 Presented at: Taller de NEGación en Español; 2017; Spain p. 33-37.
37. Medlock B, Briscoe T. Weakly Supervised Learning for Hedge Classification in Scientific Literature. ACL Anthology. 2007. URL: https://www.aclweb.org/anthology/P07-1125/ [accessed 2020-11-22]
38. Morante R, Daelemans W. Learning the Scope of Hedge Cues in Biomedical Texts. In: BioNLP '09: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing. 2009 Presented at: Workshop on Current Trends in Biomedical Natural Language Processing; June 2009; Boulder, Colorado p. 28-36. [doi: 10.3115/1572364.1572369]
39. Cruz Díaz NP, Maña López MJ, Vázquez J, Álvarez V. A machine‐learning approach to negation and speculation detection in clinical texts. J Am Soc Inf Sci Tec 2012 May 31;63(7):1398-1410. [doi: 10.1002/asi.22679]
40. Agarwal S, Yu H. Biomedical negation scope detection with conditional random fields. J Am Med Inform Assoc 2010 Nov 01;17(6):696-701 [FREE Full text] [doi: 10.1136/jamia.2010.003228] [Medline: 20962133]
41. Konstantinova N, de Sousa SCM, Cruz NP, Maña MJ, Taboada M, Mitkov R. A review corpus annotated for negation, speculation and their scope. ACL Anthology. 2012. URL: https://www.aclweb.org/anthology/L12-1298/ [accessed 2020-11-22]
42. Zou B, Zhou G, Zhu Q. Tree Kernel-based Negation and Speculation Scope Detection with Structured Syntactic Parse Features. ACL Anthology. 2013. URL: https://www.aclweb.org/anthology/D13-1099/ [accessed 2020-11-22]
43. White JP. UWashington: Negation Resolution using Machine Learning Methods. ACL Anthology. 2012. URL: https://www.aclweb.org/anthology/S12-1044/ [accessed 2020-11-22]
44. Casillas A, Pérez A, Oronoz M, Gojenola K, Santiso S. Learning to extract adverse drug reaction events from electronic health records in Spanish. Expert Systems with Applications 2016 Nov;61:235-245. [doi: 10.1016/j.eswa.2016.05.034]
45. Donatelli L. Cues, Scope, and Focus: Annotating Negation in Spanish Corpora. In: Proceedings of NEGES 2018: Workshop on Negation in Spanish. 2018 Presented at: Workshop on Negation in Spanish; September 18, 2018; Seville, Spain p. 29-34 URL: http://ceur-ws.org/Vol-2174/paper3.pdf
46. Lazib L, Zhao Y, Qin B, Liu T. Negation Scope Detection with Recurrent Neural Networks Models in Review Texts. In: Social Computing. ICYCSEE 2016. Communications in Computer and Information Science, vol 623. Singapore: Springer; 2016:494-508.
47. Lazib L, Qin B, Zhao Y, Zhang W, Liu T. A syntactic path-based hybrid neural network for negation scope detection. Front. Comput. Sci 2018 Aug 2;14(1):84-94. [doi: 10.1007/s11704-018-7368-6]
48. Jiménez-Zafra SM, Taulé M, Martín-Valdivia MT, Ureña-López LA, Martí MA. SFU ReviewSP-NEG: a Spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns. Lang Resources & Evaluation 2017 May 22;52(2):533-569. [doi: 10.1007/s10579-017-9391-x]
49. Fancellu F, Lopez A, Webber B, He H. Detecting negation scope is easy, except when it isn't. ACL Anthology. 2017. URL: https://www.aclweb.org/anthology/E17-2010/ [accessed 2020-11-22]
50. Morante R, Blanco E. *SEM 2012 Shared Task: Resolving the Scope and Focus of Negation. ACL Anthology. 2012. URL: https://www.aclweb.org/anthology/S12-1035/ [accessed 2020-11-22]
51. Mehrabi S, Krishnan A, Sohn S, Roch AM, Schmidt H, Kesterson J, et al. DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. J Biomed Inform 2015 Apr;54:213-219 [FREE Full text] [doi: 10.1016/j.jbi.2015.02.010] [Medline: 25791500]
52. spaCy. URL: https://spacy.io/ [accessed 2020-11-22]
53. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. brat: a Web-based Tool for NLP-Assisted Text Annotation. ACL Anthology. 2012. URL: https://www.aclweb.org/anthology/E12-2021/ [accessed 2020-11-22]
54. Borthwick A, Sterling J, Agichtein E, Grishman R. Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition. ACL Anthology. 1998. URL: https://www.aclweb.org/anthology/W98-1118/ [accessed 2020-11-22]
55. Giorgi JM, Bader GD. Transfer learning for biomedical named entity recognition with neural networks. Bioinformatics 2018 Dec 01;34(23):4087-4094 [FREE Full text] [doi: 10.1093/bioinformatics/bty449] [Medline: 29868832]
56. Wang D, Zheng TF. Transfer learning for speech and language processing. In: 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). 2015 Presented at: 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA); December 16-19, 2015; Hong Kong, China. [doi: 10.1109/APSIPA.2015.7415532]
57. Mou L, Meng Z, Yan R, Li G, Xu Y, Zhang L, et al. How Transferable are Neural Networks in NLP Applications? ACL Anthology. 2016 Nov. URL: https://www.aclweb.org/anthology/D16-1046/ [accessed 2020-11-22]
58. Lee JY, Dernoncourt F, Szolovits P. Transfer Learning for Named-Entity Recognition with Neural Networks. ACL Anthology. 2018. URL: https://www.aclweb.org/anthology/L18-1708/ [accessed 2020-11-22]

59. Ling W, Dyer C, Black AW, Trancoso I. Two/Too Simple Adaptations of Word2Vec for Syntax Problems. ACL Anthology. 2015. URL: https://www.aclweb.org/anthology/N15-1142/ [accessed 2020-11-22]
60. Taulé M, Martí MA, Recasens M. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. ACL Anthology. 2008. URL: https://www.aclweb.org/anthology/L08-1222/ [accessed 2020-11-22]
61. word2vec. URL: http://word2vec.googlecode.com/svn/trunk/ [accessed 2020-08-25]
62. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems. 2013 Presented at: 26th International Conference on Neural Information Processing Systems; December 2013; Red Hook, New York p. 3111-3119. [doi: 10.5555/2999792.2999959]
63. Lafferty JD, McCallum AK, Pereira FCN. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning. 2001 Presented at: Eighteenth International Conference on Machine Learning; June 2001; San Francisco, California p. 282-289. [doi: 10.5555/645530.655813]
64. Pennington J, Socher R, Manning C. GloVe: Global Vectors for Word Representation. ACL Anthology. 2014. URL: https://www.aclweb.org/anthology/D14-1162/ [accessed 2020-11-22]
65. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. TACL 2017 Dec;5:135-146. [doi: 10.1162/tacl_a_00051]
66. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep Contextualized Word Representations. ACL Anthology. 2018. URL: https://www.aclweb.org/anthology/N18-1202/ [accessed 2020-11-22]
67. McCann B, Bradbury J, Xiong C, Socher R. Learned in Translation: Contextualized Word Vectors. arXiv. 2017. URL: https://arxiv.org/abs/1708.00107 [accessed 2020-11-22]
68. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ACL Anthology. 2019. URL: https://www.aclweb.org/anthology/N19-1423/ [accessed 2020-11-22]
69. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv. 2017. URL: https://arxiv.org/abs/1706.03762 [accessed 2020-11-22]
70. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. 2018. URL: http://arxiv.org/abs/1810.04805 [accessed 2020-11-22]
71. Jiménez-Zafra SM, Cruz Díaz NP, Morante R, Martín-Valdivia MT. NEGES 2018 Task 2: Negation Cues Detection. In: Proceedings of NEGES 2018: Workshop on Negation in Spanish. 2018 Presented at: Workshop on Negation in Spanish, NEGES 2018; September 18, 2018; Seville, Spain p. 35-41.
72. Montserrat M, Vivaldi J, Bel N. Annotation of negation in the IULA Spanish Clinical Record Corpus. ACL Anthology. 2017. URL: https://www.aclweb.org/anthology/W17-1807/ [accessed 2020-11-22]
73. Collier N, Park HS, Ogata N, Tateishi Y, Nobata C, Ohta T, et al. The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers. In: EACL '99: Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics. 1999 Presented at: Ninth conference on European chapter of the Association for Computational Linguistics; June 1999; Bergen, Norway p. 271-272. [doi: 10.3115/977035.977081]
74. Ciao. URL: https://www.ciao.es/ [accessed 2020-11-22]
75. CoNLL-2010 Shared Task. MTA-SZTE Research Group on Artificial Intelligence. URL: https://rgai.inf.u-szeged.hu/node/118 [accessed 2020-08-25]
76. Georgescul M. A Hedgehop over a Max-Margin Framework Using Hedge Cues. ACL Anthology. 2010. URL: https://www.aclweb.org/anthology/W10-3004/ [accessed 2020-11-22]
77. Ji F, Qiu X, Huang X. Detecting Hedge Cues and their Scopes with Average Perceptron. ACL Anthology. 2010. URL: https://www.aclweb.org/anthology/W10-3005/ [accessed 2020-11-22]
78. Chen L, Di Eugenio B. A Lucene and Maximum Entropy Model Based Hedge Detection System. ACL Anthology. 2010. URL: https://www.aclweb.org/anthology/W10-3016 [accessed 2020-11-22]
79. Tang B, Wang X, Wang X, Yuan B, Fan S. A Cascade Method for Detecting Hedges and their Scope in Natural Language Text. ACL Anthology. 2010. URL: https://www.aclweb.org/anthology/W10-3002 [accessed 2020-11-22]
80. Li X, Shen J, Gao X, Wang X. Exploiting Rich Features for Detecting Hedges and their Scope. ACL Anthology. 2010. URL: https://www.aclweb.org/anthology/W10-3011 [accessed 2020-11-22]
81. Özgür A, Radev DR. Detecting Speculations and their Scopes in Scientific Text. In: EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. 2009 Presented at: 2009 Conference on Empirical Methods in Natural Language Processing; August 2009; Singapore p. 1398-1407. [doi: 10.3115/1699648.1699686]
82. Zhou H, Li X, Huang D, Li Z, Yang Y. Exploiting Multi-Features to Detect Hedges and their Scope in Biomedical Texts. ACL Anthology. 2010. URL: https://www.aclweb.org/anthology/W10-3015 [accessed 2020-11-22]
83. Morante R, Van Asch V, Daelemans W. Memory-Based Resolution of In-Sentence Scopes of Hedge Cues. ACL Anthology. 2010. URL: https://www.aclweb.org/anthology/W10-3006 [accessed 2020-11-22]
84. Velldal E, Øvrelid L, Oepen S. Resolving Speculation: MaxEnt Cue Classification and Dependency-Based Scope Rules. ACL Anthology. 2010. URL: https://www.aclweb.org/anthology/W10-3007 [accessed 2020-11-22]

85. Santiso S, Casillas A, Pérez A, Oronoz M. Word embeddings for negation detection in health records written in Spanish. Soft Comput 2018 Nov 23;23(21):10969-10975. [doi: 10.1007/s00500-018-3650-7]
86. Fabregat H, Martinez-Romo J, Araujo L. Deep Learning Approach for Negation Cues Detection in Spanish. In: Proceedings of NEGES 2018: Workshop on Negation in Spanish. 2018 Presented at: Workshop on Negation in Spanish; September 18, 2019; Seville, Spain p. 43-48 URL: http://ceur-ws.org/Vol-2174/paper5.pdf
87. Loharja H, Padró L, Turmo J. Negation Cues Detection Using CRF on Spanish Product Review Texts. In: Proceedings of NEGES 2018: Workshop on Negation in Spanish. 2018 Presented at: Workshop on Negation in Spanish; September 18, 2018; Seville, Spain p. 49-54 URL: http://ceur-ws.org/Vol-2174/paper6.pdf
88. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 01;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
89. De Silva TS, MacDonald D, Paterson G, Sikdar KC, Cochrane B. Systematized nomenclature of medicine clinical terms (SNOMED CT) to represent computed tomography procedures. Comput Methods Programs Biomed 2011 Mar;101(3):324-329. [doi: 10.1016/j.cmpb.2011.01.002] [Medline: 21316117]

Abbreviations

BERT: bidirectional encoder representations from transformers
Bi-LSTM: bidirectional long short-term memory
CNN: convolutional neural network
CRF: conditional random field
NER: named entity recognition
NLP: natural language processing
PoS: part of speech
RNN: recurrent neural network

Edited by G Eysenbach; submitted 29.03.20; peer-reviewed by L Zhang, J Kim, G Lim; comments to author 29.06.20; revised version received 25.08.20; accepted 28.10.20; published 03.12.20.

Please cite as:
Rivera Zavala R, Martinez P
The Impact of Pretrained Language Models on Negation and Speculation Detection in Cross-Lingual Medical Text: Comparative Study
JMIR Med Inform 2020;8(12):e18953
URL: https://medinform.jmir.org/2020/12/e18953
doi:10.2196/18953
PMID:33270027

Original Paper

Model-Based Reasoning of Clinical Diagnosis in Integrative Medicine: Real-World Methodological Study of Electronic Medical Records and Natural Language Processing Methods

Wenye Geng[1*], MD; Xuanfeng Qin[2*], MD; Tao Yang[3], MD; Zhilei Cong[3], MD; Zhuo Wang[4], MD; Qing Kong[1], MD; Zihui Tang[1], MD; Lin Jiang[5], MD

[1]Department of Integrative Medicine, Fudan University Huashan Hospital, Shanghai, China
[2]Department of Neurosurgery, Fudan University Huashan Hospital, Shanghai, China
[3]Emergency Department, Huashan Hospital of Fudan University, Shanghai, China
[4]Shanghai Sunjian Informatics Technology Company Limited, Shanghai, China
[5]Healthcare Center, Fudan University Huashan Hospital, Shanghai, China
[*]these authors contributed equally

Corresponding Author:
Zihui Tang, MD
Department of Integrative Medicine
Fudan University Huashan Hospital
No 12 Urumuqi Mid Road
Shanghai
China
Phone: 86 021 5288 8236
Email: dr_zhtang@yeah.net

Abstract

Background: Integrative medicine is a form of medicine that combines practices and treatments from alternative medicine with conventional medicine. The diagnosis in integrative medicine involves the clinical diagnosis based on modern medicine and syndrome pattern diagnosis. Electronic medical records (EMRs) are the systematized collection of patients health information stored in a digital format that can be shared across different health care settings. Although syndrome and sign information or relative information can be extracted from the EMR and content texts can be mapped to computability vectors using natural language processing techniques, application of artificial intelligence techniques to support physicians in medical practices remains a major challenge.

Objective: The purpose of this study was to investigate model-based reasoning (MBR) algorithms for the clinical diagnosis in integrative medicine based on EMRs and natural language processing. We also estimated the associations among the factors of sample size, number of syndrome pattern type, and diagnosis in modern medicine using the MBR algorithms.

Methods: A total of 14,075 medical records of clinical cases were extracted from the EMRs as the development data set, and an external test data set consisting of 1000 medical records of clinical cases was extracted from independent EMRs. MBR methods based on word embedding, machine learning, and deep learning algorithms were developed for the automatic diagnosis of syndrome pattern in integrative medicine. MBR algorithms combining rule-based reasoning (RBR) were also developed. A standard evaluation metrics consisting of accuracy, precision, recall, and F1 score was used for the performance estimation of the methods. The association analyses were conducted on the sample size, number of syndrome pattern type, and diagnosis of lung diseases with the best algorithms.

Results: The Word2Vec convolutional neural network (CNN) MBR algorithms showed high performance (accuracy of 0.9586 in the test data set) in the syndrome pattern diagnosis of lung diseases. The Word2Vec CNN MBR combined with RBR also showed high performance (accuracy of 0.9229 in the test data set). The diagnosis of lung diseases could enhance the performance of the Word2Vec CNN MBR algorithms. Each group sample size and syndrome pattern type affected the performance of these algorithms.

Conclusions: The MBR methods based on Word2Vec and CNN showed high performance in the syndrome pattern diagnosis of lung diseases in integrative medicine. The parameters of each group's sample size, syndrome pattern type, and diagnosis of lung diseases were associated with the performance of the methods.

Trial Registration: ClinicalTrials.gov NCT03274908; https://clinicaltrials.gov/ct2/show/NCT03274908

(JMIR Med Inform 2020;8(12):e23082) doi:10.2196/23082

KEYWORDS

model-based reasoning; integrative medicine; electronic medical records; natural language processing

Introduction

Integrative medicine is a form of medicine that combines practices and treatments from alternative medicine with conventional medicine [1-3]. In China, integrative medicine combines traditional Chinese medicine (TCM) and modern medicine for clinical practice [1-3]. The diagnosis in integrative medicine comprises the clinical diagnosis based on modern medicine and syndrome pattern diagnosis [4]. Syndrome pattern based on TCM theory is an outcome of the analysis of TCM information by the TCM practitioner, and TCM treatments rely on this concept [4]. A syndrome pattern can be defined as a categorized pattern of symptoms and signs in a patient at a specific stage during the course of a disease. Syndrome elements are the smaller units of syndrome classification and the basic elements of a syndrome pattern [5]. The correct combination of syndrome elements can infer an appropriate syndrome pattern. Syndrome elements are also derived from the syndrome and signs from the patient [5,6]. Generally, practitioners of integrative medicine making diagnosis decisions need to combine syndrome pattern diagnosis and the diagnosis in modern medicine [5,6]. As TCM treatments rely on syndrome pattern diagnosis, the treatment combined with the therapies of TCM and modern medicine is expected to be more efficient for patients. Therefore, syndrome pattern for the diagnosis in integrative medicine is an essential part of diagnosis.

Electronic medical records (EMRs) are the systematized collection of patients' and the population's electronically stored health information in a digital format that can be shared across different health care settings [7,8]. In China, EMRs are a collection of diagnoses of syndrome patterns and model medicine as well as syndromes and signs with the TCM format [7,8]. Natural language processing (NLP) is a field of artificial intelligence and computational linguistics concerned with the interactions between computers and human natural languages [9,10]. Currently, NLP techniques combining EMRs have been comprehensively applied to medical data mining and medical decision support system [9,10]. Word embedding, as one of the techniques in NLP, attempted to map a word using a dictionary to a vector of real numbers in a low-dimensional space [11,12]. It is important in EMR data mining or artificial intelligence application in medicine for medical texts to be transferred to vectors because computers can handle or understand medical texts through computability vectors.

Applying artificial intelligence techniques to support physicians in medical practices is a major challenge. The processing of uncertainty information mainly contributes to the challenge. Syndrome and sign information is under the classic uncertainty information. The artificial neural network (ANN) can successfully and efficiently handle syndrome and sign information with uncertainty [13]. ANN is a computational model based on the structure and functions of biological neural networks [14]. The remarkable information processing characteristics of the ANN in terms of nonlinearity, fault and noise tolerance, high parallelism, and learning and generalization capabilities contribute to uncertain information processing and quantitative analysis. Furthermore, model-based reasoning (MBR) methods based on machine learning or ANN can successfully process syndrome and sign information with uncertainty to make a precise and accurate diagnosis in integrative medicine.

As mentioned previously, syndrome and sign information or relative information can be extracted from the EMRs, and content texts can be mapped to computability vectors using NLP techniques. Furthermore, MBR methods can be used to create a computer-aided system to support the diagnosis in integrative medicine. However, only a few studies have been conducted on MBR methods with EMRs and NLP to support the diagnosis in integrative medicine. Fortunately, our previous work was carried out to analyze syndrome patterns and syndrome elements in lung diseases based on real-world EMR data [5]. This study aimed to explore MBR algorithms in the diagnosis in integrative medicine based on EMRs and NLP techniques applied on lung disease data sets. We also estimated the associations among the factors of sample size, number of syndrome pattern type, and diagnosis in modern medicine using the MBR algorithms.

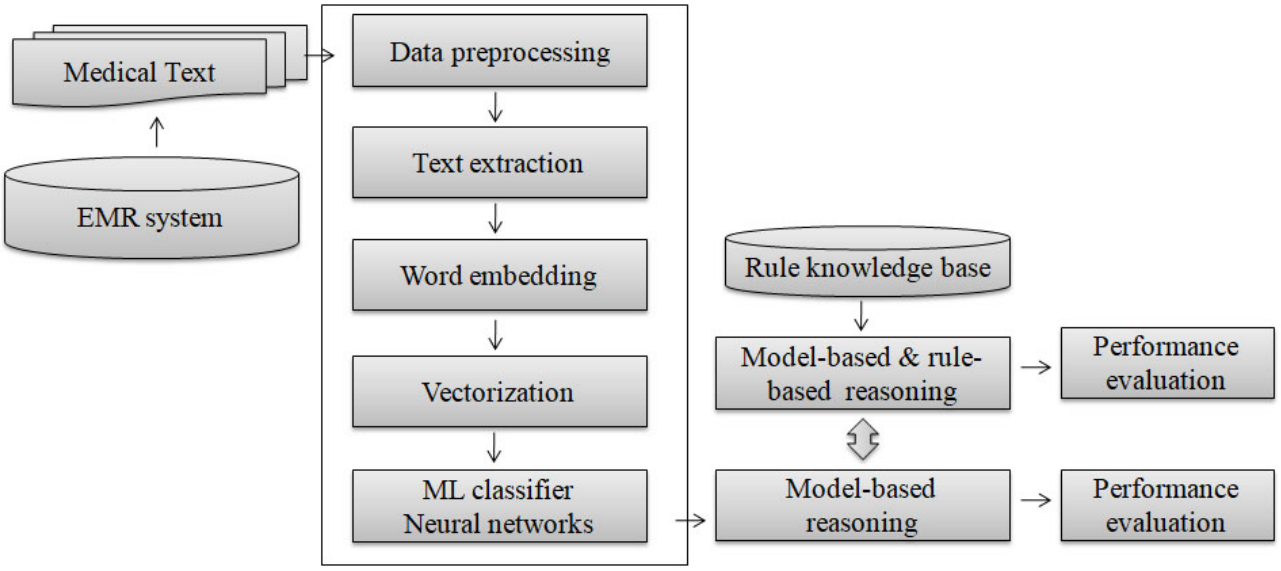
Methods

Analysis of Workflow

The workflow of the analysis of the MBR methods in the diagnosis in integrative medicine based on EMRs and NLP is illustrated in Figure 1. The EMRs on lung diseases were exported from the hospital information system, and the syndrome and sign information and relative information were extracted as a text format. The corresponding syndrome pattern diagnosis, clinical diagnosis in modern medicine, and syndrome elements were extracted and saved to the database with the structure data according to the unique code of patients. The content texts of the syndrome and sign information were mapped to the computability vectors through word embedding. The classification models that include the vectors of syndrome and sign information and syndrome patterns or syndrome elements were developed using machine learning or neural network methods. MBR algorithms were developed on the basis of

classification models concerning the syndrome pattern, and the model-based and rule reasoning algorithms were developed using the classification models and rule knowledge based on the combination of syndrome elements and syndrome patterns. The performances of the MBR methods in the diagnosis of lung diseases in integrative medicine have been evaluated and compared (for the main program codes for the module, please see [15]).

Figure 1. Workflow of the analysis of MBR methods in the diagnosis in integrative medicine based on EMRs and NLP. EMR: electronic medical record; MBR: model-based reasoning; ML: machine learning; NLP: natural language processing.



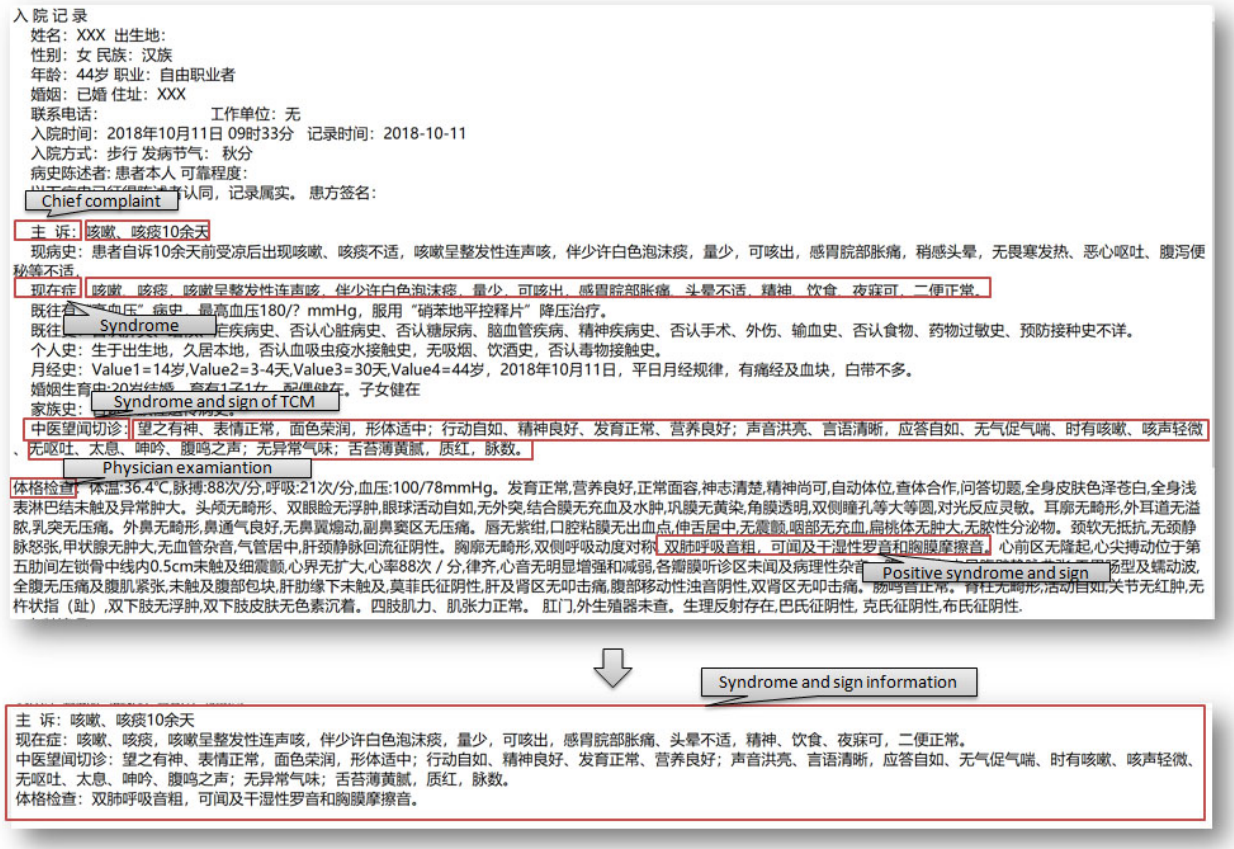
Data Collection and Processing

In our previous real-world study on the syndrome pattern and syndrome element of lung disease, EMRs were collected from lung disease wards in 5 hospitals [5]. A data set consisting of 14,075 medical records of clinical cases from 4 hospitals was assigned as the development data set, and it was divided into the train data set and the test data set at a ratio of 4:1. Another independent data set comprising 1000 medical records of clinical cases from a hospital was set as the external test data set. The information comprised patients' identity number, ward number, admission time, admission notes, first medical records, general medical records, discharge note, diagnosis of syndrome pattern, and diagnosis in modern medicine. In this work, we selected 10 common syndrome pattern types and 8 common lung diseases in the lung disease wards. Nine syndrome element types were generated and combined with the corresponding 10 syndrome pattern types.

Medical Information Extraction

The Chinese text information on the chief complaints, syndromes, and positive signs in the chest, tongue, and pulse was extracted from the admission notes, first medical records, and discharge records (Figure 2). The extracted Chinese text information was combined into contexts called "four diagnoses in TCM." The contexts of the syndromes and signs underwent word-cutting process to split them into tokens. In this work, the first corpus included the context of syndrome and sign information. In the analysis of the diagnosis in modern medicine and syndrome pattern diagnosis, another corpus included an additional token of diagnosis in modern medicine.

Figure 2. The Chinese text information on the chief complaints, syndromes, and positive signs in the chest, tongue, and pulse that was extracted from the admission notes, first medical records, and discharge records. TCM: traditional Chinese medicine.



Word2Vec

Word embedding is an NLP feature-learning technique in which words are mapped to vectors of real numbers [16]. Word embedding involves mathematical embedding from a space with 1 dimension per word to a continuous vector space with a much lower number of dimensions. The Word2Vec model is an NLP system that is used to produce word embedding, which takes a large corpus of text as its input and produces a vector space, and each unique word in the corpus is assigned a corresponding vector in the space [16]. The Word2Vec model generates vectors for each word present in a document. In this study, the corpus from a Chinese language Wikipedia dump, which is available at [17], was used to pretrain the word vector model. The parameters utilized with the Word2Vec model were developed for dimension reduction into 256 dimension vectors, 5 context windows, and a minimum sentence word count of 10. The Word2Vec model was implemented using the Gensim Python library [18].

Doc2Vec

The Doc2Vec model is an extension of Word2Vec that constructs embeddings from entire documents or sentences (instead of individual words) to learn a randomly initialized vector for the document (or sentence) along with the words [19]. The Doc2Vec model modifies the Word2Vec algorithm into an unsupervised learning algorithm that produces continuous representations for large blocks of texts, such as sentences, paragraphs, or entire documents. In this work, Doc2Vec was used to produce vectors for texts. The corpus from a Chinese language Wikipedia dump was again used to pretrain the Doc2Vec model. The parameters utilized with the Doc2Vec model were developed in the dimension reduction into 192 dimension vectors, 5 context windows, and a minimum sentence word count of 10. The Doc2Vec model was also implemented using the Gensim Python library.

Machine Learning

In this work, the 4 different machine learning classifiers algorithms, namely, random forest (RF), extreme gradient boosting (XGBoost), support vector machines (SVMs), and K-nearest neighbor (KNN), were used to develop MBR [20-22]. The 4 algorithms were the classic machine leaning algorithms, which were the best algorithms suitable for classification tasks.

RF, a classic machine learning classifier, is composed of tree predictors, with each tree depending on the values of a random vector sampled independently and having the same distribution for all trees in the forest [23]. RF aims to reduce the tree correlation issue by choosing only a subsample of the feature space at each split. In this work, RF was used on 1000 trees in the forest, and it was implemented using the scikit-learn Python library.

XGBoost is an optimized distributed gradient-boosting system designed to be highly efficient, flexible, and portable [24]. It implements machine learning algorithms under the gradient boosting framework, which attempts to accurately predict a

target variable by combining an ensemble of estimates from a set of simpler, weaker models. XGBoost can also be implemented using the scikit-learn Python library.

SVM is a well-known supervised learning model associated with learning algorithms that analyze data used for classification and regression analysis [25]. SVM was useful in text-based classification tasks and is not prone to errors in high-dimensional data sets. In this work, SVM was used with a linear kernel and implemented using the scikit-learn Python library.

The KNN classifier, one of the most popular machine learning algorithms, is based on the Euclidean distance between a test sample and the specified training samples [26]. It is used for data classification that attempts to determine in which group a data point is included by examining the data points around it. In this study, KNN was implemented using the scikit-learn Python library.

Artificial Neural Network

ANNs, one of the main tools used in machine learning, are a group of models inspired by biological neural networks used for estimating functions that depend on a large number of inputs [13]. ANN algorithms have 2 different classifiers: multilayer perceptron (MLP) and convolutional neural network (CNN). MLP is a feed-forward ANN model that maps sets of input data onto a set of appropriate outputs [27]. It consists of multiple layers of nodes with a nonlinear activation function in a directed graph, with each layer fully connected to the next one. Back-propagation is used as a supervised learning technique in MLP. In this work, MLP was performed with 6 hidden layers, with the nodes per layer varying from 64 to 1024. It was also implemented using the scikit-learn Python library.

CNN is one of the most popular algorithms for deep learning [28]. It is a category of ANN in which a model learns to perform classification tasks directly from images, text, or sound, and it has been proven effective in the areas of text classification and image recognition. CNN comprises one or more convolutional layers with a subsampling step, followed by one or more fully connected layers as in a standard multilayer neural network [29]. In this work, CNN consisted of an embedding layer, a convolutional layer, a max pooling layer, and 2 fully connected layers, and it was implemented using the Keras Python library.

MBR

In this study, the development of MBR was based on word embedding and machine learning classifiers for syndrome pattern [30,31]. A total of 11 MBR algorithms were used: Word2Vec RF, Word2Vec XGBoost, Word2Vec SVM, Word2Vec KNN, Word2Vec MLP, Word2Vec CNN, Doc2Vec RF, Doc2Vec XGBoost, Doc2Vec SVM, Doc2Vec KNN, and Doc2Vec MLP. These models with multiclass outputs were consistent with the syndrome pattern types. A comparison of the performance of the 11 MBR algorithms was conducted.

MBR Combined With Rule-Based Reasoning

MBR was based on word embedding and machine learning classifiers for syndrome elements. Nine MBR algorithms were used: Word2Vec RF, Word2Vec XGBoost, Word2Vec KNN, Word2Vec MLP, Word2Vec CNN, Doc2Vec RF, Doc2Vec XGBoost, Doc2Vec KNN, and Doc2Vec MLP. These models with multilabel outputs were consistent with the syndrome element types. The syndrome patterns were generated by combining the syndrome elements, which follow the rule knowledge base of the syndrome elements, with the syndrome pattern. A comparison of the performance of the 9 MBR combined with rule-based reasoning (RBR) algorithms was performed. The rules of combination of TCM elements for TCM syndrome are presented in Multimedia Appendix 1.

Evaluation

The performances of the MBR algorithms in syndrome pattern were evaluated in the test data set and the external data set using standard metrics, which included accuracy, precision, recall, and F1 score [32]. Moreover, the performances of the Word2Vec CNN MBR algorithms in each syndrome pattern and each syndrome element were evaluated in the test data set using standard metrics. A fivefold cross validation was conducted 20 times on the train data set for each algorithm to estimate the 95% CI for the performance parameters.

The accuracy comparison analysis of the Word2Vec CNN MBR algorithms in corpus 1 and corpus 2 was conducted in different proportions of the sample size of the development data set. In the accuracy analysis of the data set, each group sample size was set as a proportion of total sample size and the number of syndrome pattern type was selected randomly. The linear regression analyses were conducted to evaluate the associations between each group sample size and the number of syndrome pattern type at accuracies of 0.90% and 0.95% of the methods.

Ethics Approval and Consent to Participate

The study was approved by the Ethics Committee of the Huashan Hospital and performed in accordance with the Declaration of Helsinki.

Availability of Data and Material

The data sets generated or analyzed during this study are not publicly available due to private information but are available from the corresponding author on reasonable request. Data sets are from the study whose authors may be contacted at the Center of Bioinformatics and Biostatistics, Institutes of Integrative Medicine, Fudan University. The data concerning external test data set and an example of development data set are available online [15].

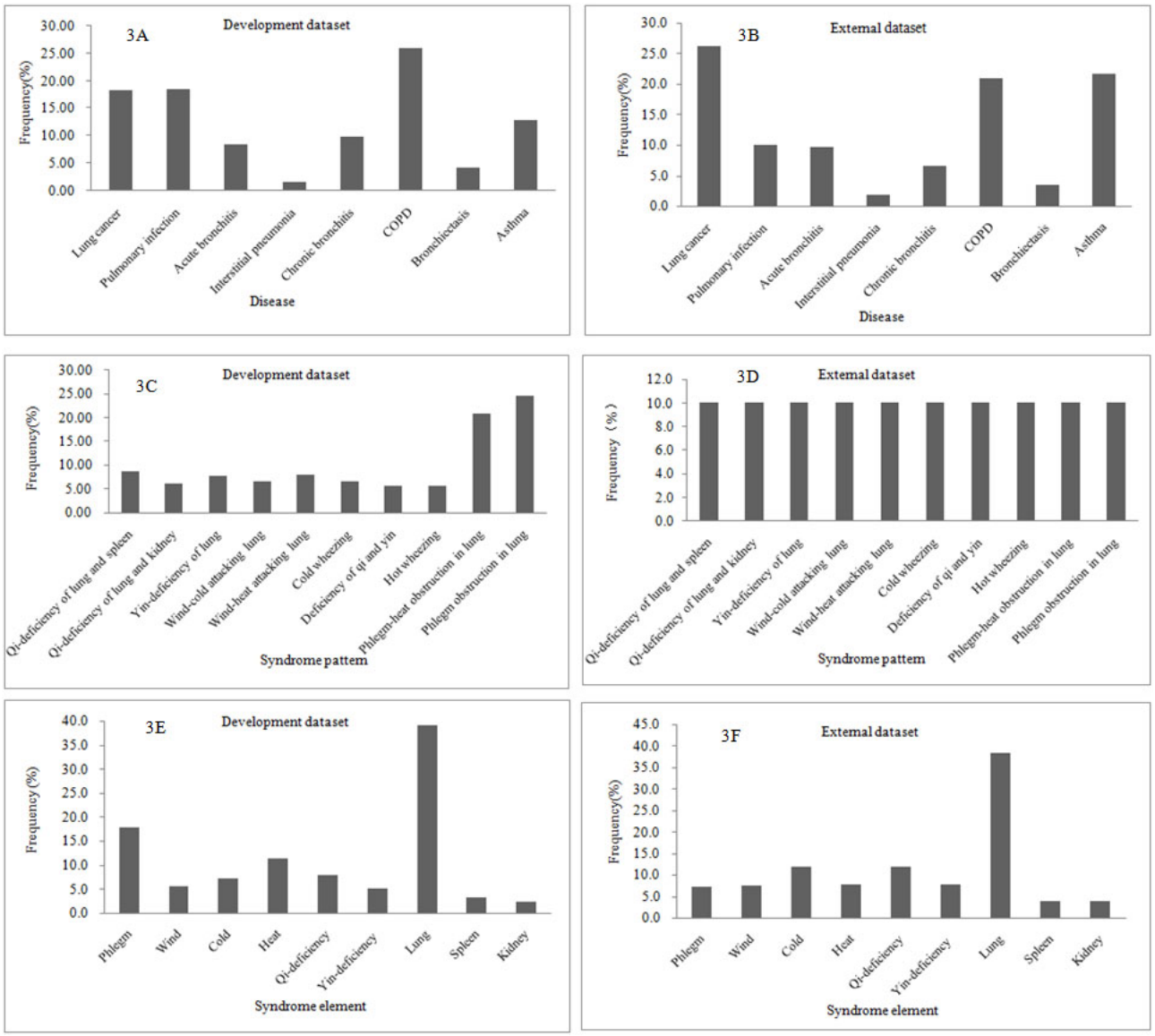
Results

Development and External Data Sets

The characteristics of the data set are shown in Figure 3. The development data set consisted of 14,075 medical records of clinical cases, and the external data set had 1000 medical records of clinical cases. Eight common lung diseases were found in the development data set: lung cancer (18.42%), pulmonary infection (18.59%), acute bronchitis (8.39%), interstitial pneumonia (1.66%), chronic bronchitis (9.78%), chronic obstructive pulmonary disease (25.98%), bronchiectasis (4.31%), and asthma (12.88%; Figure 3A). The same common lung diseases with the same proportions were also found in the external data set (Figure 3B). Ten common syndrome pattern

types were found in the development data set: qi-deficiency of lung and spleen, qi-deficiency of lung and kidney, yin-deficiency of lung, wind-cold attacking lung, wind-heat attacking lung, cold wheezing, deficiency of qi and yin, hot wheezing, phlegm-heat obstruction in lung, and phlegm obstruction in lung (Figure 3C). The same 10 syndrome pattern types with the same proportions were also found in the external data set (Figure 3D). The development data set had 35,992 syndrome elements for 14,075 syndrome patterns, and a syndrome pattern consisted of 2.56 syndrome elements on average. The development data set included 9 syndrome element types: phlegm, wind, cold, heat, qi-deficiency, yin-deficiency, lung, spleen, and kidney (Figure 3E). A total of 2602 syndrome elements with the same 9 types were found in 1000 syndrome patterns (Figure 3F).

Figure 3. The characteristics of the data set. COPD: chronic obstructive pulmonary disease.



MBR

In the test data set, the performance analysis of the MBR based on Word2Vec to identify syndrome patterns showed an average accuracy of 0.9397 (95% CI 0.9312-0.9468) in the Word2Vec RF model and 0.9323 (95% CI 0.9213-0.9443) in the Word2Vec ANN model (Table 1). The highest average accuracy was 0.9471 (95% CI 0.9382-0.9549) in the Word2Vec CNN model. The parameters of precision, recall, and F1 score were 0.9478 (95% CI 0.9393-0.9557), 0.9471 (95% CI 0.9382-0.9549), and 0.9470 (95% CI 0.9383-0.9550) in the Word2Vec CNN model, respectively. Similar performance values were found in the corresponding external data set.

Table 1. Performance analysis of model-based reasoning methods applied for syndrome pattern diagnosis of lung disease based on Word2Vec in the test and external data sets.

Model and data set	Accuracy, mean (95% CI)	Precision, mean (95% CI)	Recall, mean (95% CI)	F1 score, mean (95% CI)
Word2Vec + RF[a]				
Test	0.9397 (0.9312-0.9468)	0.9411 (0.9331-0.9481)	0.9397 (0.9312-0.9468)	0.9396 (0.9311-0.9468)
External	0.9121 (0.9001-0.9251)	0.9125 (0.8985-0.9189)	0.9120 (0.9030-0.9220)	0.9118 (0.8988-0.9208)
Word2Vec + XGBoost[b]				
Test	0.8832 (0.8732-0.8942)	0.8844 (0.8714-0.8954)	0.8832 (0.8722-0.8932)	0.8832 (0.8742-0.8972)
External	0.8720 (0.8641-0.8842)	0.8753 (0.8643-0.8893)	0.8720 (0.8630-0.8860)	0.8728 (0.8598-0.8838)
Word2Vec + KNN[c]				
Test	0.8485 (0.8355-0.8605)	0.8489 (0.8349-0.8569)	0.8485 (0.8355-0.8575)	0.8478 (0.8398-0.8598)
External	0.8481 (0.8371-0.8611)	0.8514 (0.8404-0.8624)	0.8481 (0.8351-0.8561)	0.8481 (0.8351-0.8591)
Word2Vec + SVM[d]				
Test	0.8172 (0.8062-0.8252)	0.8245 (0.8135-0.8325)	0.8172 (0.8052-0.8312)	0.8161 (0.8071-0.8251)
External	0.7791 (0.7711-0.7931)	0.8047 (0.7957-0.8177)	0.7791 (0.7681-0.7881)	0.7826 (0.7706-0.7956)
Word2Vec + MLP[e]				
Test	0.9323 (0.9213-0.9443)	0.9326 (0.9226-0.9436)	0.9323 (0.9243-0.9403)	0.9319 (0.9229-0.9409)
External	0.9203 (0.9101-0.9302)	0.9211 (0.9101-0.9341)	0.9201 (0.9090-0.9340)	0.9193 (0.9063-0.9293)
Word2Vec + CNN[f]				
Test	0.9471 (0.9382-0.9549)	0.9478 (0.9393-0.9557)	0.9471 (0.9382-0.9549)	0.9470 (0.9383-0.9550)
External	0.9250 (0.9110-0.9360)	0.9277 (0.9153-0.9382)	0.9250 (0.9110-0.9360)	0.9250 (0.9114-0.9362)

[a]RF: random forest.

[b]XGBoost: extreme gradient boosting.

[c]KNN: K nearest neighbor.

[d]SVM: support vector machine.

[e]MLP: multilayer perceptron.

[f]CNN: convolutional neural network.

The performance analysis of the MBR based on Doc2Vec to identify syndrome patterns in the test data set showed the highest average accuracy of 0.8840 (95% CI 0.8730-0.8970) in the Doc2Vec CNN model (Table 2). The parameters of precision, recall, and F1 score were 0.8876 (95% CI 0.8776-0.8976), 0.8840 (95% CI 0.8710-0.8932), and 0.8843 (95% CI 0.8753-0.8973) in the Doc2Vec CNN model, respectively. Similar performance values were found in the corresponding external data set.

Table 2. Performance analysis of model-based reasoning methods applied for syndrome pattern diagnosis of lung disease based on Doc2Vec in the test and external data sets.

Model and data set	Accuracy, mean (95% CI)	Precision, mean (95% CI)	Recall, mean (95% CI)	F1 score, mean (95% CI)
Doc2Vec + RF[a]				
Test	0.8320 (0.8198-0.8442)	0.8457 (0.8345-0.8567)	0.8320 (0.8198-0.8442)	0.8337 (0.8217-0.8458)
External	0.8190 (0.8090-0.8310)	0.8506 (0.8366-0.8610)	0.8190 (0.8110-0.8323)	0.8267 (0.8147-0.8397)
Doc2Vec + XGBoost[b]				
Test	0.7584 (0.7444-0.7724)	0.7682 (0.7602-0.7812)	0.7584 (0.7504-0.7704)	0.7589 (0.7499-0.7719)
External	0.7270 (0.719-0.7400)	0.7735 (0.7645-0.7835)	0.7270 (0.7130-0.7390)	0.7391 (0.7261-0.7501)
Doc2Vec + KNN[c]				
Test	0.8527 (0.8407-0.8637)	0.8588 (0.8488-0.8668)	0.8527 (0.8407-0.8627)	0.8535 (0.8425-0.8665)
External	0.8202 (0.8092-0.8282)	0.8246 (0.8116-0.8326)	0.8220 (0.8090-0.8331)	0.8215 (0.8105-0.8295)
Doc2Vec +SVM[d]				
Test	0.6748 (0.6628-0.6848)	0.7424 (0.7334-0.7504)	0.6748 (0.6668-0.6858)	0.7577 (0.7467-0.7667)
External	0.5820 (0.5700-0.5950)	0.5743 (0.5663-0.5883)	0.5920 (0.5830-0.6033)	0.5288 (0.5168-0.5388)
Doc2Vec + MLP[e]				
Test	0.8840 (0.8730-0.8970)	0.8876 (0.8776-0.8976)	0.8840 (0.8710-0.8932)	0.8843 (0.8753-0.8973)
External	0.8760 (0.8620-0.8890)	0.8897 (0.8757-0.9027)	0.8760 (0.8630-0.8851)	0.8791 (0.8701-0.8921)

[a]RF: random forest.

[b]XGBoost: extreme gradient boosting.

[c]KNN: K nearest neighbor.

[d]SVM: support vector machine.

[e]MLP: multilayer perceptron.

MBR Combined With RBR

The performance analysis of the MBR combined with RBR based on Word2Vec in the test data set indicated that the highest average accuracy was 0.9229 (95% CI 0.9099-0.9319) in the Word2Vec CNN model (Table 3). The parameters of precision, recall, and F1 score were 0.9884 (95% CI 0.9744-0.9964), 0.9679 (95% CI 0.9589-0.9809), and 0.9778 (95% CI 0.9698-0.9888) in the Word2Vec CNN model, respectively. Similar performance values were found in the corresponding external data set.

Table 3. Performance analysis of model-based reasoning methods in combination with rule-based reasoning methods applied for syndrome pattern diagnosis of lung disease based on Word2Vec in the test and external data sets.

Model and data set	Accuracy, mean (95% CI)	Precision, mean (95% CI)	Recall, mean (95% CI)	F1 score, mean (95% CI)
Word2Vec + RF[a]				
Test	0.9131 (0.8990-0.9261)	0.9934 (0.9814-0.9983)	0.9628 (0.9538-0.9748)	0.9774 (0.9644-0.9864)
External	0.9040 (0.8903-0.9180)	0.9657 (0.9547-0.9747)	0.9580 (0.9501-0.9721)	0.9617 (0.9477-0.9697)
Word2Vec + XGBoost[b]				
Test	0.7703 (0.7583-0.7803)	0.9666 (0.9556-0.9786)	0.9044 (0.8924-0.9144)	0.9333 (0.9233-0.9433)
External	0.7980 (0.7871-0.8112)	0.9702 (0.9582-0.9812)	0.9227 (0.9137-0.9337)	0.9444 (0.9364-0.9544)
Word2Vec + KNN[c]				
Test	0.8414 (0.8324-0.8534)	0.9380 (0.9270-0.9502)	0.9254 (0.9164-0.9334)	0.9312 (0.9202-0.9432)
External	0.8521 (0.8403-0.8612)	0.9441 (0.9321-0.9571)	0.9373 (0.9263-0.9473)	0.9446 (0.9306-0.9556)
Word2Vec + MLP[d]				
Test	0.9052 (0.8930-0.9181)	0.9751 (0.9621-0.9830)	0.9758 (0.9678-0.9858)	0.9752 (0.9652-0.9862)
External	0.9021 (0.8940-0.9151)	0.9791 (0.9671-0.9911)	0.9780 (0.9660-0.9904)	0.9784 (0.9704-0.9904)
Word2Vec + CNN[e]				
Test	0.9229 (0.9099-0.9319)	0.9884 (0.9744-0.9964)	0.9679 (0.9589-0.9809)	0.9778 (0.9698-0.9888)
External	0.9160 (0.9030-0.9261)	0.9765 (0.9655-0.9885)	0.9662 (0.9582-0.9782)	0.9698 (0.9608-0.9778)

[a]RF: random forest.

[b]XGBoost: extreme gradient boosting.

[c]KNN: K nearest neighbor.

[d]MLP: multilayer perceptron.

[e]CNN: convolutional neural network.

The performance analysis of the MBR combined with RBR based on Doc2Vec showed that the highest average accuracy was 0.8190 (95% CI 0.8082-0.8281) in the Doc2Vec CNN model (Table 4). The parameters of precision, recall, and F1 score were 0.9550 (95% CI 0.9441-0.9673), 0.9507 (95% CI 0.9387-0.9597), and 0.9524 (95% CI 0.9444-0.9654) in the Doc2Vec CNN model, respectively. Similar performance values were found in the corresponding external data set.

Table 4. Performance analysis of model-based reasoning methods in combination with rule-based reasoning methods applied for syndrome pattern diagnosis of lung disease based on Doc2Vec in the test and external data sets.

Model and data set	Accuracy, mean (95% CI)	Precision, mean (95% CI)	Recall, mean (95% CI)	F1 score, mean (95% CI)
Doc2Vec + RF[a]				
Test	0.6410 (0.6281-0.6520)	0.8586 (0.8496-0.8698)	0.9745 (0.9635-0.9865)	0.9049 (0.8939-0.9139)
External	0.5940 (0.5810-0.6061)	0.9728 (0.9648-0.9828)	0.8002 (0.7892-0.8112)	0.8642 (0.8542-0.8762)
Doc2Vec + XGBoost[b]				
Test	0.6177 (0.6087-0.6307)	0.8525 (0.8415-0.8625)	0.9413 (0.9273-0.9513)	0.8891 (0.8771-0.8981)
External	0.536 (0.5272-0.5440)	0.9346 (0.9266-0.9486)	0.7863 (0.7763-0.7953)	0.8401 (0.8301-0.8531)
Doc2Vec + KNN[c]				
Test	0.8488 (0.8358-0.8618)	0.9393 (0.9283-0.9523)	0.9503 (0.9383-0.9613)	0.9440 (0.9331-0.9582)
External	0.8260 (0.8174-0.8383)	0.9203 (0.9073-0.9323)	0.9415 (0.9275-0.9535)	0.9301 (0.9211-0.9401)
Doc2Vec + MLP[d]				
Test	0.8190 (0.8082-0.828)1	0.9550 (0.9441-0.9673)	0.9507 (0.9387-0.9597)	0.9524 (0.9444-0.9654)
External	0.8031 (0.7911-0.8111)	0.9478 (0.9398-0.9618)	0.9446 (0.9316-0.9546)	0.9444 (0.9314-0.9544)

[a]RF: random forest.

[b]XGBoost: extreme gradient boosting.

[c]KNN: K nearest neighbor.

[d]MLP: multilayer perceptron.

Word2Vec CNN MBR in Corpus 1 and Corpus 2

Corpus 1 included the syndrome and sign information without a clinical diagnosis of lung disease, whereas corpus 2 included the syndrome and sign information with a clinical diagnosis of lung disease. A higher average accuracy (0.9584; 95% CI 0.9510-0.9655) was found in the Word2Vec CNN model for syndrome pattern diagnosis in corpus 2 than in corpus 1 (0.9471; 95% CI 0.9382-0.9549) in the test data set (Table 5). Moreover, higher performance parameter values of precision, recall, and F1 score were found in the Word2Vec CNN model for each syndrome pattern diagnosis in corpus 2 than in corpus 1 (Table 5). Similar results were found in the Word2Vec CNN method combined with the RBR model for syndrome pattern diagnosis in corpus 2 in comparison with the model in corpus 1 in the test data set with a full sample size (Table 6). A higher average accuracy of the Word2Vec CNN model was found for syndrome pattern diagnosis in the test data set with different sample sizes in corpus 2 than in corpus 1 (Figure 4).

Table 5. Performance analysis of model-based reasoning methods for each syndrome pattern in the test data set with corpus 1 and corpus 2.[a]

Syndrome pattern	Corpus 1				Corpus 2			
	Precision	Recall	F1 score	Support	Precision	Recall	F1 score	Support
Qi-deficiency of lung and spleen	0.9363	0.9514	0.9438	247	0.9957	0.9665	0.9809	239
Qi-deficiency of lung and kidney	0.9362	0.9999	0.9670	176	0.9781	0.9944	0.9861	179
Yin-deficiency of lung	0.9777	0.9733	0.9755	225	0.9902	0.9999	0.9951	203
Wind-cold attacking lung	0.9943	0.9943	0.9956	176	0.9878	0.9999	0.9939	162
Wind-heat attacking lung	0.9899	0.9120	0.9494	216	0.9150	0.9826	0.9476	230
Cold wheezing	0.9724	0.9832	0.9778	179	0.9750	0.9653	0.9701	202
Deficiency of qi and yin	0.9934	0.9804	0.9868	153	0.9932	0.9932	0.9945	147
Hot wheezing	0.9051	0.9931	0.947	144	0.9563	0.9808	0.9684	156
Phlegm-heat obstruction in lung	0.9389	0.9021	0.9201	613	0.9357	0.9125	0.9240	606
Phlegm obstruction in lung	0.9183	0.9344	0.9263	686	0.9461	0.9407	0.9434	691
Average (weighted)	0.9477	0.9471	0.9470	2815	0.9586	0.9584	0.9584	2815

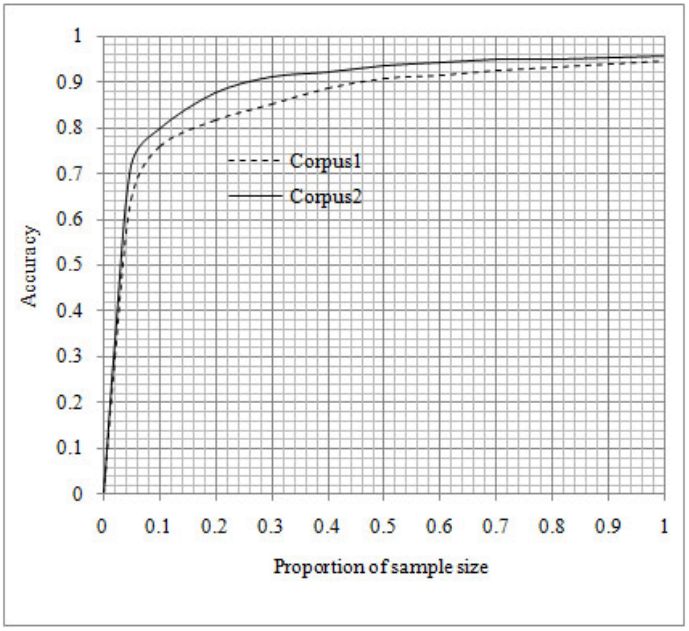
[a]Corpus 1 consists of syndrome and sign information, and corpus 2 consists of syndrome and sign information plus clinical diagnosis information. The average accuracy was 0.9471 (95% CI 0.9382-0.9549) for syndrome pattern in the test data set with corpus 1, and 0.9584 (95% CI 0.9510-0.9655) for syndrome pattern in the test data set with corpus 2.

Table 6. Performance analysis of model-based reasoning methods in combination with rule-based reasoning methods for each syndrome element in the test data set with corpus 1 and corpus 2.[a]

Syndrome element	Corpus 1				Corpus 2			
	Precision	Recall	F1 score	Support	Precision	Recall	F1 score	Support
Phlegm	0.9907	0.9538	0.9719	1233	0.9935	0.9951	0.9943	1233
Wind	0.9926	0.9218	0.9559	435	0.9953	0.9770	0.9861	435
Cold	0.9800	0.9722	0.976	503	0.996	1.000	0.998	503
Heat	0.9704	0.8903	0.9286	811	0.9675	0.9174	0.9418	811
Qi-deficiency	0.9616	0.9756	0.9686	616	0.9871	0.9935	0.9903	616
Yin-deficiency	1.000	0.9851	0.9925	403	0.9975	0.9801	0.9887	403
Lung	1.000	1.000	1.000	2815	1.000	1.000	1.000	2815
Spleen	0.9644	0.9457	0.955	258	0.9771	0.9922	0.9846	258
Kidney	0.9882	0.9825	0.9853	171	0.9826	0.9883	0.9854	171
Average (weighted)	0.9885	0.968	0.9779	7245	0.9922	0.9863	0.9892	7245

[a]Corpus 1 consists of syndrome and sign information, and corpus 2 consists of syndrome and sign information plus clinical diagnosis information. The average accuracy was 0.9229 (95% CI 0.9099-0.9319) for syndrome pattern in the test data set with corpus 1, and 0.9559 (95% CI 0.9429-0.9699) for syndrome pattern in the test data set with corpus 2.

Figure 4. Accuracy and sample size proportions in corpus 1 and corpus 2.



Association of Accuracy and Sample Size With Syndrome Pattern Type

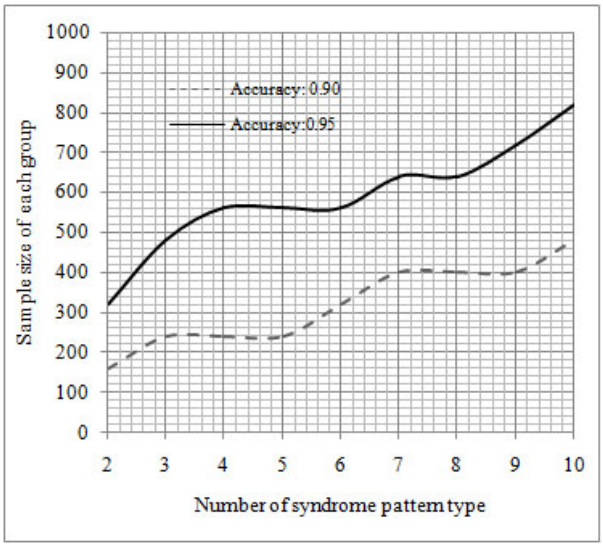
We performed an average accuracy analysis in the development data set classified by the number of syndrome pattern type and each group's sample size. The results showed that the average accuracy increased with the increase in sample size of each group and decreased with the increase in number of syndrome pattern (Table 7). The linear regression analysis showed that each group's sample size was significantly associated with the number of syndrome pattern with an accuracy of 0.90 ($Y = 34.39 \times X + 109.43$, $P<.001$, where Y is each group's sample size and X is the number of syndrome pattern type) and 0.95 ($Y = 48.55 \times X + 296.78$, $P<.001$, where Y is each group's sample size and X is the number of syndrome pattern type), respectively (Figure 5).

Table 7. Average accuracy analysis grouped by sample size of each group and number of syndrome pattern type.[a]

Each group sample size	N=2	N=3	N=4	N=5	N=6	N=7	N=8	N=9	N=10
16	0.5714	0.4001	0.3876	0.3122	0.2521	0.3113	0.3076	0.2068	0.1875
40	0.6575	0.5001	0.4375	0.3511	0.2916	0.3751	0.3751	0.2916	0.2251
64	0.7238	0.6412	0.5384	0.5125	0.4636	0.4444	0.4174	0.4127	0.3921
80	0.8751	0.7291	0.6406	0.6311	0.5521	0.4732	0.5468	0.4513	0.4001
160	*0.9375*	0.8542	0.8437	0.8432	0.8345	0.7901	0.7621	0.7577	0.7325
240	0.9375	*0.9097*	*0.9014*	*0.9011*	0.8993	0.8482	0.8515	0.8487	0.8083
320	*0.9658*	0.9114	0.9074	0.9151	*0.9227*	0.8973	0.8984	0.8836	0.8515
400	0.9688	0.9433	0.9384	0.9281	0.9301	*0.9266*	*0.9023*	*0.9025*	0.8929
480	0.9752	*0.9553*	0.9414	0.9412	0.9418	0.9464	0.9444	0.9234	*0.9135*
560	0.9762	0.9583	*0.9534*	*0.9521*	*0.9532*	0.9482	0.9487	0.9394	0.9304
640	0.9776	0.9653	0.9633	0.9661	0.9626	*0.9526*	*0.9619*	0.9456	0.9354
720	0.9786	0.9708	0.9688	0.9712	0.9709	0.9672	0.9678	*0.9591*	0.9356
800	0.9813	0.9776	0.9756	0.9735	0.9739	0.9785	0.9734	0.9597	0.9429

[a]The first average accuracy was arrived at 0.90 and 0.95 and corresponding values are presented in italics.

Figure 5. Sample size of each group.



Discussion

Principal Findings

We developed MBR methods for diagnosis of lung diseases in integrative medicine based on a real-world EMR data set with NLP. In our previous studies, we accumulated large-scale real-world data for artificial intelligence on integrative medicine. In this work, real-world medical records of clinical cases were used to develop models, and medical texts were mapped to vectors of real numbers that a computer could process. CNN approaches can automatically extract features from word vectors, thus contributing to the high performance of MBR methods in syndrome pattern diagnosis for diagnosis of lung diseases in integrative medicine. To the best of our knowledge, this study is the first to investigate MBR methods for diagnosis in integrative medicine on a large real-world data set using NLP and deep learning methods in China. These MBR methods can be recommended for a clinical decision-making system and can also provide a novel approach for diagnosis in integrative medicine. This work would be of significance for applications of artificial intelligence on integrative medicine.

An interesting finding is the high performance of the MBR methods for syndrome pattern diagnosis in integrative medicine. The best Word2Vec CNN MBR method for syndrome pattern diagnosis in integrative medicine had an accuracy of 0.9471 and 0.9250 in the development and external data sets, respectively. Word embedding and CNN contributed to the high performance. Word embedding techniques can map texts to computability vectors, which can perform text analysis with quantitative analysis. CNN can automatically extract features from medical texts, significantly contributing to the performance of the MBR. Additionally, the diagnosis information of modern medicine being added to the corpus enhances the accuracy of the syndrome pattern diagnosis in integrative medicine with

reasoning, thus indicating that physicians can more efficiently make a syndrome pattern diagnosis after determining the diagnosis in modern medicine.

We performed an association analysis to evaluate the relationship between the number of syndrome pattern type and each group's sample size for the accuracy of MBR algorithms. Moreover, we conducted a linear regression analysis to estimate the linear function of each group's sample size and syndrome pattern type at an accuracy of 0.95. Only a few studies reported on the quantitative associations. In the Word2Vec CNN MBR algorithms at an accuracy of 0.95, the smallest group sample size was 300 for 2 syndrome pattern types, and for each group the sample size was at least 800 for 10 syndrome pattern types. According to the linear model, the Word2Vec CNN MBR method based on each group's sample size of at least 1200 showed high performance in syndrome pattern with 20 types. A total of 400 common syndrome pattern types were grouped into 20 systems in integrative internal medicine. A total of 25,000 medical records of clinical cases could satisfy the Word2Vec CNN MBR methods in syndrome pattern diagnosis in an integrative system at an accuracy of 0.95. A total of 500,000 medical records of clinical cases could satisfy the Word2Vec CNN MBR methods in the diagnosis of 400 syndrome patterns in the entire integrative internal medicine at an accuracy of 0.95. We could thus combine data-driven artificial intelligence and knowledge-driven artificial intelligence for developing an intelligent clinical decision system on integrative medicine.

Interestingly, the combination of MBR and RBR methods applied for syndrome pattern diagnosis in integrative medicine showed high performance. Specifically, Word2Vec CNN MBR combined with RBR methods had an accuracy of 0.9559 in syndrome pattern diagnosis in corpus 2 with additional information on modern medicine diagnosis. This reasoning method showed a more understandable and clearer knowledge of lung diseases for physicians in comparison with the Word2Vec CNN MBR methods. Moreover, it was more suitable for users of or physicians practicing integrative medicine. Generally, a hybrid reasoning is more suitable for application in clinical practice. The data- and knowledge-driven artificial intelligence contributed to the hybrid reasoning, which has the advantages of high performance reasoning and being explainable for clinicians. In clinical practice, the TCM elements reasoning could be used for TCM diagnosis or differentiation.

Although this study used novel methods to develop MBR in syndrome pattern diagnosis in integrative medicine, it has several limitations. First, we selected only 10 of the 20 common syndrome pattern types in lung diseases, partly because the other 10 syndrome pattern types did not have enough medical records of clinical cases. Therefore, future studies should use comprehensive syndrome patterns in lung diseases or other systems. Second, the size of the corpus for pretrained word vectors was not large to cover all Chinese words or special items on lung diseases.

Conclusion

MBR methods based on Word2Vec CNN showed high performance in syndrome pattern diagnosis of lung diseases in integrative medicine. The parameters of each group's sample size, syndrome pattern type, and clinical diagnosis of lung diseases were associated with the performance of the methods. The hybrid reasoning with data- and knowledge-driven artificial intelligence could well contribute to the development of medical artificial intelligence on integrative medicine. We aim to develop a clinical diagnosis or decision-making model with knowledge graph and hybrid reasoning to better combine data- and knowledge-driven artificial intelligence on integrative medicine in the near future.

Acknowledgments

This work was supported by grants from the Institutes of Integrative Medicine of Fudan University (ClinicalTrials.gov Identifier: NCT03274908) and China Postdoctoral Science Foundation-funded project (2017M611461).

Authors' Contributions

WG and XQ drafted the manuscript. TY, ZC, ZW, and QK participated in the design of the study and performed the statistical analysis. ZT and LJ conceived the study, and participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Rule knowledge base.
[XLSX File (Microsoft Excel File), 10 KB - medinform_v8i12e23082_app1.xlsx]

References

1. Wang J, Xiong X. Current situation and perspectives of clinical study in integrative medicine in china. Evid Based Complement Alternat Med 2012;2012:268542-268511 [FREE Full text] [doi: 10.1155/2012/268542] [Medline: 22550539]

2. Leung TH, Wong W. Development of integrative medicine in Hong Kong, China. Chin J Integr Med 2017 Jul 17;23(7):486-489. [doi: 10.1007/s11655-017-2815-z] [Medline: 28623621]
3. Xu H, Chen K. Integrative medicine: the experience from China. J Altern Complement Med 2008 Jan;14(1):3-7. [doi: 10.1089/acm.2006.6329] [Medline: 18199020]
4. Lee T, Lo L, Wu F. Traditional Chinese Medicine for Metabolic Syndrome via TCM Pattern Differentiation: Tongue Diagnosis for Predictor. Evid Based Complement Alternat Med 2016;2016:1971295-1971298 [FREE Full text] [doi: 10.1155/2016/1971295] [Medline: 27313640]
5. Xu F, Cui W, Kong Q, Tang Z, Dong J. A Real-World Evidence Study for Distribution of Traditional Chinese Medicine Syndrome and Its Elements on Respiratory Disease. Evid Based Complement Alternat Med 2018;2018:8305892 [FREE Full text] [doi: 10.1155/2018/8305892] [Medline: 30643538]
6. Wei J, Wu R, Zhao D. Analysis of TCM syndrome elements and relevant factors for senile diabetes. Journal of Traditional Chinese Medicine 2013 Aug;33(4):473-478. [doi: 10.1016/s0254-6272(13)60151-x] [Medline: 24187868]
7. Xu Y, Li N, Lu M, Myers RP, Dixon E, Walker R, et al. Development and validation of method for defining conditions using Chinese electronic medical record. BMC Med Inform Decis Mak 2016 Aug 20;16(1):110 [FREE Full text] [doi: 10.1186/s12911-016-0348-6] [Medline: 27542973]
8. Xue Y, Liang H, Wu X, Gong H, Li B, Zhang Y. Effects of electronic medical record in a Chinese hospital: a time series study. Int J Med Inform 2012 Oct;81(10):683-689. [doi: 10.1016/j.ijmedinf.2012.05.017] [Medline: 22727614]
9. Wang H, Zhang W, Zeng Q, Li Z, Feng K, Liu L. Extracting important information from Chinese Operation Notes with natural language processing methods. J Biomed Inform 2014 Apr;48:130-136 [FREE Full text] [doi: 10.1016/j.jbi.2013.12.017] [Medline: 24486562]
10. Chen L, Song L, Shao Y, Li D, Ding K. Using natural language processing to extract clinically useful information from Chinese electronic medical records. Int J Med Inform 2019 Apr;124:6-12. [doi: 10.1016/j.ijmedinf.2019.01.004] [Medline: 30784428]
11. Wu Y, Xu J, Jiang M, Zhang Y, Xu H. A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text. AMIA Annu Symp Proc 2015;2015:1326-1333 [FREE Full text] [Medline: 26958273]
12. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, et al. A comparison of word embeddings for the biomedical natural language processing. J Biomed Inform 2018 Nov;87:12-20 [FREE Full text] [doi: 10.1016/j.jbi.2018.09.008] [Medline: 30217670]
13. Hramov AE, Frolov NS, Maksimenko VA, Makarov VV, Koronovskii AA, Garcia-Prieto J, et al. Artificial neural network detects human uncertainty. Chaos 2018 Mar;28(3):033607. [doi: 10.1063/1.5002892] [Medline: 29604631]
14. Tang ACY, Chung JWY, Wong TKS. Validation of a novel traditional chinese medicine pulse diagnostic model using an artificial neural network. Evid Based Complement Alternat Med 2012;2012:685094 [FREE Full text] [doi: 10.1155/2012/685094] [Medline: 21918652]
15. Tang Z. Clinical Decision Support System. URL: https://github.com/zihuitang/clincial_decision_support_system_im [accessed 2020-12-09]
16. Zhu Y, Yan E, Wang F. Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. BMC Med Inform Decis Mak 2017 Jul 03;17(1):95 [FREE Full text] [doi: 10.1186/s12911-017-0498-1] [Medline: 28673289]
17. Wikipedia Dump. URL: https://dumps.wikimedia.org/zhwiki/latest/zhwiki-latest-pages-articles.xml.bz2 [accessed 2020-12-09]
18. Ince RAA, Petersen RS, Swan DC, Panzeri S. Python for information theoretic analysis of neural data. Front Neuroinform 2009;3:4 [FREE Full text] [doi: 10.3389/neuro.11.004.2009] [Medline: 19242557]
19. Xing W, Yuan X, Li L, Hu L, Peng J. Phenotype Extraction Based on Word Embedding to Sentence Embedding Cascaded Approach. IEEE Trans Nanobioscience 2018 Jul;17(3):172-180. [doi: 10.1109/tnb.2018.2838137]
20. Baştanlar Y, Ozuysal M. Introduction to machine learning. Methods Mol Biol 2014;1107:105-128. [doi: 10.1007/978-1-62703-748-8_7] [Medline: 24272434]
21. Kotoku J. An Introduction to Machine Learning. Igaku Butsuri 2016;36(1):18-22 [FREE Full text] [doi: 10.11323/jjmp.36.1_18] [Medline: 28428491]
22. Rowe M. An Introduction to Machine Learning for Clinicians. Acad Med 2019 Oct;94(10):1433-1436. [doi: 10.1097/ACM.0000000000002792] [Medline: 31094727]
23. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological Methods 2009 Dec;14(4):323-348. [doi: 10.1037/a0016973]
24. Sheridan RP, Wang WM, Liaw A, Ma J, Gifford EM. Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships. J Chem Inf Model 2016 Dec 27;56(12):2353-2360. [doi: 10.1021/acs.jcim.6b00591] [Medline: 27958738]
25. Baumes LA, Serra JM, Serna P, Corma A. Support vector machines for predictive modeling in heterogeneous catalysis: a comprehensive introduction and overfitting investigation based on two real applications. J Comb Chem 2006;8(4):583-596. [doi: 10.1021/cc050093m] [Medline: 16827571]

26. Abu Alfeilat HA, Hassanat AB, Lasassmeh O, Tarawneh AS, Alhasanat MB, Eyal Salman HS, et al. Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review. Big Data 2019 Dec 01;7(4):221-248. [doi: 10.1089/big.2018.0175] [Medline: 31411491]
27. Araujo P, Astray G, Ferrerio-Lage JA, Mejuto JC, Rodriguez-Suarez JA, Soto B. Multilayer perceptron neural network for flow prediction. J. Environ. Monit 2011;13(1):35-41. [doi: 10.1039/c0em00478b] [Medline: 21088795]
28. Zheng T, Gao Y, Wang F, Fan C, Fu X, Li M, et al. Detection of medical text semantic similarity based on convolutional neural network. BMC Med Inform Decis Mak 2019 Aug 07;19(1):156. [doi: 10.1186/s12911-019-0880-2] [Medline: 31391038]
29. Hamm CA, Wang CJ, Savic LJ, Ferrante M, Schobert I, Schlachter T, et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. Eur Radiol 2019 Jul;29(7):3338-3347 [FREE Full text] [doi: 10.1007/s00330-019-06205-9] [Medline: 31016442]
30. Jang B, Kim I, Kim JW. Word2vec convolutional neural networks for classification of news articles and tweets. PLoS One 2019;14(8):e0220976 [FREE Full text] [doi: 10.1371/journal.pone.0220976] [Medline: 31437181]
31. Turner CA, Jacobs AD, Marques CK, Oates JC, Kamen DL, Anderson PE, et al. Word2Vec inversion and traditional text classifiers for phenotyping lupus. BMC Med Inform Decis Mak 2017 Aug 22;17(1):126 [FREE Full text] [doi: 10.1186/s12911-017-0518-1] [Medline: 28830409]
32. Handelman GS, Kok HK, Chandra RV, Razavi AH, Huang S, Brooks M, et al. Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods. AJR Am J Roentgenol 2019 Jan;212(1):38-43. [doi: 10.2214/AJR.18.20224] [Medline: 30332290]

Abbreviations

ANN: artificial neural network
CNN: convolutional neural network
EMRs: electronic medical records
KNN: K-nearest neighbor
MBR: model-based reasoning
MLP: multilayer perceptron
NLP: natural language processing
RBR: rule-based reasoning
RF: random forest
SVM: support vector machine
TCM: traditional Chinese medicine
XGBoost: extreme gradient boosting

Edited by C Lovis; submitted 31.07.20; peer-reviewed by Q Zeng, V Foufi; comments to author 20.09.20; revised version received 18.10.20; accepted 07.11.20; published 21.12.20.

Please cite as:
Geng W, Qin X, Yang T, Cong Z, Wang Z, Kong Q, Tang Z, Jiang L
Model-Based Reasoning of Clinical Diagnosis in Integrative Medicine: Real-World Methodological Study of Electronic Medical Records and Natural Language Processing Methods
JMIR Med Inform 2020;8(12):e23082
URL: http://medinform.jmir.org/2020/12/e23082/
doi:10.2196/23082
PMID:33346740

Original Paper

The Generalizability of a Medication Administration Discrepancy Detection System: Quantitative Comparative Analysis

Eric Kirkendall[1,2,3], MD, MBI; Hannah Huth[1,4], BA; Benjamin Rauenbuehler[1,5], BS; Adam Moses[1,6], MHA, PMP; Kristin Melton[3,7], MD; Yizhao Ni[3,8], PhD

[1]Center for Healthcare Innovation, Wake Forest School of Medicine, Winston Salem, NC, United States
[2]Department of Pediatrics, Wake Forest School of Medicine, Winston Salem, NC, United States
[3]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, United States
[4]College of Medicine, University of Tennessee Health Science Center, Memphis, TN, United States
[5]University of Iowa, Iowa City, IA, United States
[6]Department of Internal Medicine, Wake Forest School of Medicine, Winston Salem, NC, United States
[7]Division of Neonatology and Pulmonary Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States
[8]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States

Corresponding Author:
Eric Kirkendall, MD, MBI
Center for Healthcare Innovation
Wake Forest School of Medicine
486 North Patterson Avenue
Office 512
Winston Salem, NC, 27101
United States
Phone: 1 (336) 716 0462
Fax: 1 (336) 713 4947
Email: ekirkend@wakehealth.edu

Abstract

Background: As a result of the overwhelming proportion of medication errors occurring each year, there has been an increased focus on developing medication error prevention strategies. Recent advances in electronic health record (EHR) technologies allow institutions the opportunity to identify medication administration error events in real time through computerized algorithms. MED.Safe, a software package comprising medication discrepancy detection algorithms, was developed to meet this need by performing an automated comparison of medication orders to medication administration records (MARs). In order to demonstrate generalizability in other care settings, software such as this must be tested and validated in settings distinct from the development site.

Objective: The purpose of this study is to determine the portability and generalizability of the MED.Safe software at a second site by assessing the performance and fit of the algorithms through comparison of discrepancy rates and other metrics across institutions.

Methods: The MED.Safe software package was executed on medication use data from the implementation site to generate prescribing ratios and discrepancy rates. A retrospective analysis of medication prescribing and documentation patterns was then performed on the results and compared to those from the development site to determine the algorithmic performance and fit. Variance in performance from the development site was further explored and characterized.

Results: Compared to the development site, the implementation site had lower audit/order ratios and higher MAR/(order + audit) ratios. The discrepancy rates on the implementation site were consistently higher than those from the development site. Three drivers for the higher discrepancy rates were alternative clinical workflow using orders with dosing ranges; a data extract, transfer, and load issue causing modified order data to overwrite original order values in the EHRs; and delayed EHR documentation of verbal orders. Opportunities for improvement were identified and applied using a software update, which decreased false-positive discrepancies and improved overall fit.

Conclusions: The execution of MED.Safe at a second site was feasible and effective in the detection of medication administration discrepancies. A comparison of medication ordering, administration, and discrepancy rates identified areas where MED.Safe

could be improved through customization. One modification of MED.Safe through deployment of a software update improved the overall algorithmic fit at the implementation site. More flexible customizations to accommodate different clinical practice patterns could improve MED.Safe's fit at new sites.

(JMIR Med Inform 2020;8(12):e22031) doi:10.2196/22031

KEYWORDS

medication administration; error; automated algorithm; generalizability; quantitative comparative analysis; discrepancy; detection; quantitative analysis; portability; performance algorithm; electronic health record

Introduction

Patient safety is maximized when medical errors are efficiently detected and mitigated or prevented in the first place. The most common type of medical errors are medication errors, which are defined as any preventable event that may cause or lead to inappropriate medication use or patient harm while the medication is in the control of the health care professional, patient, or consumer [1]. Medication errors can occur at all stages in the patient care process including ordering, transcribing, dispensing, administration, and monitoring [2-4]. In recent years, medication administration has been identified as an error-prone stage in the patient care process and comprises a large percentage of all medical errors [3]. Despite extensive efforts, medication administration errors (MAEs) continue to inundate patient care [5,6].

The persistence of medication errors has led to a need for clinical informatics methods and technological interventions to improve medication error detection and prevention [7,8]. Common informatics approaches to prevent errors include the use of dedicated systems such as clinical decision support during medication ordering in the electronic health record (EHR) or drug error reduction systems contained in smart infusion pumps; both provide overdose and other types of alerts [9,10]. The former system works to detect errors and reduce the total number of medication errors early in the medication use process (at the ordering stage) [11], but does not detect error types that are introduced downstream in the later phases such as medication administration. Improved efforts to detect different error types during the administration and monitoring phases can serve to capture issues that have propagated from early stages—in the event they are not already addressed by upstream systems—as well as detecting errors introduced later in the system [12]. By effectively detecting and identifying errors at any point of the medication use life cycle, it is possible to inform intervention and prevention strategies to prevent future errors of the same type and possibly mitigate harm [13-17].

The availability of digitized EHRs and medication administration records (MARs) make it possible to perform algorithmic analysis of the data to detect MAEs quickly and efficiently [12,14,18,19]. Furthermore, the EHR and the creation of care-related data afford the ability to detect MAEs or discrepancies across entire populations and large data sets. This is in contrast to current methods of detection, which usually rely on sampling strategies followed by selective manual review of records or by reviewing the output from voluntary reporting [13,15-17]. In our prior work [12,20-22], discrepancies were identified when an algorithm detected a difference between the dosage intended to be delivered (prescriber's orders) and how it was documented as being delivered (MAR data). A dosing-related MAE was defined as any discrepancy between the medication dose or infusion rate administered to a patient and the dose/rate prescribed by physicians during patient care. However, a discrepancy only becomes an error when it is clinically valid and has the potential to cause harm to the patient. As a result, error rates (ie, clinically valid errors) and discrepancy rates (ie, algorithm-based detections) are not completely synonymous; high discrepancy rates do not directly correspond to high error rates or indicate suboptimal practice until the discrepancy is investigated and deemed an actual error. However, discrepancies give reviewers a starting point to efficiently find actual errors.

In this study, we sought to implement MED.Safe, a software package of medication discrepancy detection algorithms, and benchmark the results to our earlier work at the development site to determine its portability and generalizability. We analyzed the system outputs at an external site, highlighting where and in what context the system performed well, and suggested customizations to further improve its performance. This analysis will provide the basis for further implementation and scaling of the current software package into other health care institutions.

Methods

Study Setting

The study took place at Wake Forest Baptist Medical Center (WFBMC), a tertiary level 1 trauma center and level 1 pediatric trauma center with 885 beds in Winston-Salem, North Carolina. WFBMC implemented an EHR system (Epic Systems) in 2012. This study focuses on the pediatric intensive care unit (PICU) with 12 beds, the neonatal ICU (NICU) with 40 beds, and the adult medical ICU with 172 beds.

Data Sources

Order and MAR data were extracted from the EHR for 11 medications prescribed at WFBMC: dobutamine, dopamine, epinephrine, fentanyl, insulin, intravenous (IV) fluids, lipids, milrinone, morphine, total parenteral nutrition (TPN), and vasopressin. The medications were originally selected by the investigative team (EK, KM, YN) because they were the continuously infused medications associated with the highest harm in the NICU setting. Structural differences in the format of 2 of the medication orders between the sites were taken into account during data extraction. At Cincinnati Children's Hospital Medical Center (CCHMC), all TPN and IV fluids are contained in orders under 1 parent order for each

medication/fluid category. At WFBMC, there is no single parent order, and additional mapping of the individual fluid and TPN orders was necessary. After accounting for this difference, the data from WFBMC were retrospectively extracted for the calendar year 2018 (January 1, 2018, to December 31, 2018). To compare system outputs, NICU data from CCHMC were also retrospectively extracted over the same period.

MED.Safe System

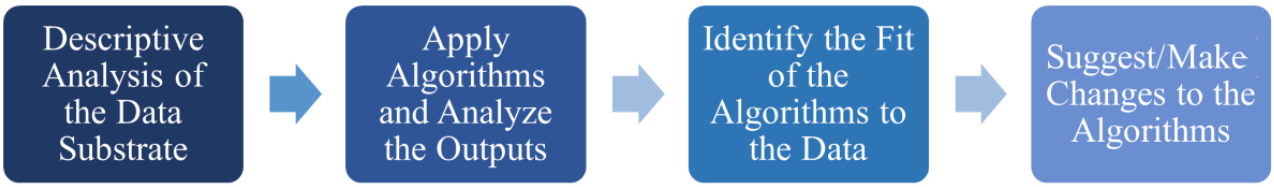
MED.Safe is an automated software package that analyzes medication use information in EHRs to identify medication administration discrepancies [12,20,21]. The MED.Safe package was originally developed by CCHMC with the purpose of monitoring high-risk IV medications in the NICU setting.

The analyzed information includes (1) medication orders that document medication doses (or infusion rates) prescribed to the patients, (2) structured order modifications (audits) that adjust the original doses/rates via computerized physician order entry, (3) MARs that document actual doses/rates administered to patients, and (4) free-text physician to nurse communication orders that deliver complex dose/rate adjustment during patient care. The free-text communications were parsed with a set of regular expression–based natural language processing algorithms to identify discrete dose/rate changes. The output consists of matching ordered medication doses with those recorded on the MAR in chronologic order. Using the extracted information, the detector module identifies discrepant doses/rates between MARs and other data sources using a set of logic-based rules. The detector was built upon our earlier research on MAE detection, where the logic-based rules were abstracted from standard care practices, refined by neonatologists, and implemented by programmers. By analyzing the dynamic EHR information, the detector determines the latest dose/rate prescribed to a patient and matches it with an MAR dose/rate to determine whether a match or discrepancy is present. MED.Safe allows users to map data elements required by the computerized algorithms to the site's EHR instance data model. Once the mapping is complete, MED.Safe automatically extracts data from the EHR instance, executes the discrepancy detection algorithms, and visualizes chronological ordering of the medication use data and the identified discrepancies (if any). It also generates descriptive statistics of the medication use data including numbers of orders, audits, MARs, and discrepancies for the studied medications.

Study Design

The investigative team (EK, BR, AM) executed the MED.Safe software package developed at CCHMC on the local WFBMC EHR data followed by a rigorous analysis of algorithm outputs. This step was completed entirely at WFBMC with guidance from the CCHMC study team (KM and YN). Analysis of the outputs was performed with the intent of learning the context within which the discrepancy detection algorithms were a good "fit" and performed accurately, and where they seemed to be inaccurate and needed customization for the new clinical environment. Figure 1 presents an overview of the study, and the individual methodological steps are further described in the following sections.

Figure 1. The overall processes of the study, for executing MED.Safe at a second site.



Phase 1: Analysis of WFBMC's Medication Ordering Environment

To determine the fit and feasibility of MED.Safe at WFBMC, the investigative team (all study authors) analyzed the quantity and distribution of medication use data available. Descriptive statistics on medication orders, order modifications (ie, audits), and MARs generated by MED.Safe were aggregated by department (NICU, PICU, and adult medical ICU) and medication to study prescriber preferences and workflows. The analyzed MARs were restricted to actions including new bag, start, restart, rate verify, and rate change, to include administrations where potential administration errors could occur. Ratios comparing the numbers of audits, orders, and MARs were calculated for all ICUs at WFBMC and the NICU at CCHMC. The audit/order ratio represented the average number of times an order was modified during its life cycle, which implied prescribing patterns in a clinical environment (if prescribers frequently changed an order or kept a more stable prescribing habit). The MAR/(order + audit) ratio represented the average number of MARs documented by clinicians for each order or order modification, which suggested documentation patterns in a clinical unit.

Phase 2: Analysis of the MED.Safe Outputs to the Data From Another EHR Instance at WFBMC

After data element configuration, MED.Safe was executed against WFBMC's clinical data repository to extract medication use data retrospectively. MED.Safe's discrepancy detection algorithms were then performed for each WFBMC ICU department. We analyzed the results aggregated across the ICU departments and for WFBMC NICU solely and compared them with those from the development site (CCHMC) to determine specific settings (medications and clinical departments) that demonstrated the best fit and areas of improvement needed for the system. Results were visualized numerically and graphically to compare trends in discrepancy rates between WFBMC and CCHMC.

Phase 3: Analysis of System Generalizability and Areas of Improvement

We assumed that good system generalizability to the WFBMC data would be expected to yield discrepancy rates similar to the baseline rates at CCHMC. Discrepancy rates substantially higher than the baselines were assumed to indicate a poor fit, which prompted further investigation to confirm this assumption and suggest areas of improvement.

If the discrepancy rate for a medication was higher than expected compared to the baseline, the system outputs were inspected manually to identify potential causes. The numbers of processed medication orders, audits, and MARs were interrogated to understand and examine the possible effect of local medication use patterns. For example, a specific type of order or MAR entry triggering discrepancies on more than 1 occasion might indicate a pattern of interest. These patterns were investigated, and the inspection was completed for each medication.

Phase 4: Suggested Customization of the System to Enable Better Detection of Medication Administration Errors

Manual analysis of the patterns identified in phase 3 was completed by the investigative team (all study authors) to pinpoint whether the source of discrepancy deviation was technical (caused by algorithm logic) or a result of clinical factors (a change of prescribing practices between sites that the system was not capable of capturing), in order to improve accuracy in MAE detection.

The technical barriers to good fit that were identified were addressed through the addition of a software update where feasible. The updated system was then re-executed on the same 2018 WFBMC data set. The updated system outputs were compared to the original system outputs in terms of order counts, order audit counts, MAR counts, and discrepancy rates to understand the impact of the customizations.

Results

Phase 1: Analysis of WFBMC's Medication Ordering Environment

Table 1 presents the distribution of medical use data for each ICU department at WFBMC. A total of 10,304 orders, 2647 audits, and 268,446 MARs were created during the study period. The NICU placed the most orders, made the most order modifications (audits), and created the most MAR entries. By contrast, the adult medical ICU had the least in all 3 categories, reflecting the fact that the MED.Safe system was originally designed for a pediatric population (the CCHMC NICU). Multimedia Appendices 1 and 2 present more specific breakdowns by medication and department, which suggested that IV fluids, TPN, lipids, and fentanyl were the most ordered medications and had the highest MARs in each of the investigated departments. The WFBMC NICU was the only investigated department without use of vasopressin and morphine; the other departments had orders and subsequent audits and MARs for all 11 medications studied. Additionally, the WFBMC NICU had almost 3 times the number of MARs when compared to the CCHMC NICU despite having only about half as many orders and audits. This was found to be the result of a practice of documenting rate verifications on the MAR much more frequently than the practice in the CCHMC NICU.

The audit/order and MAR/(order + audit) ratios are presented in Multimedia Appendices 3-5 to compare the differences in prescribing habits and order fluidity between WFBMC and CCHMC. Figure 2 compares the audit/order ratios between all WFBMC ICUs, WFBMC NICU (NICU subset of all WFBMC ICUs), and CCHMC NICU. The ratios differed substantially between the 3 data sets across the studied medications. The CCHMC NICU had higher audit/order ratios for 7 of the 11 medications. For example, dopamine at CCHMC had an audit/order ratio of 3.0, whereas that medication at WFBMC had an audit/order ratio of 0.9.

Table 1. Distribution of medication orders, audits, and medication administration records in the WFBMC ICUs compared to the CCHMC NICU.

Distribution of data elements	WFBMC[a] adult medical ICU[b]	WFBMC PICU[c]	WFBMC NICU[d]	CCHMC[e] NICU
Number of orders	1950	1964	6390	12,603
Number of audits	576	934	1137	4386
Number of MARs[f]	38,787	62,780	166,879	56,715

[a]WFBMC: Wake Forest Baptist Medical Center.

[b]ICU: intensive care unit.

[c]PICU: pediatric intensive care unit.

[d]NICU: neonatal intensive care unit.

[e]CCHMC: Cincinnati Children's Hospital Medical Center.

[f]MAR: medication administration record.

Figure 2. Comparison of audit/order ratios between (A) CCHMC NICU, (B) WFBMC NICU, and (C) WFBMC All ICUs. CCHMC: Cincinnati Children's Hospital Medical Center; ICU: intensive care unit; NICU: neonatal intensive care unit; WFBMC: Wake Forest Baptist Medical Center.

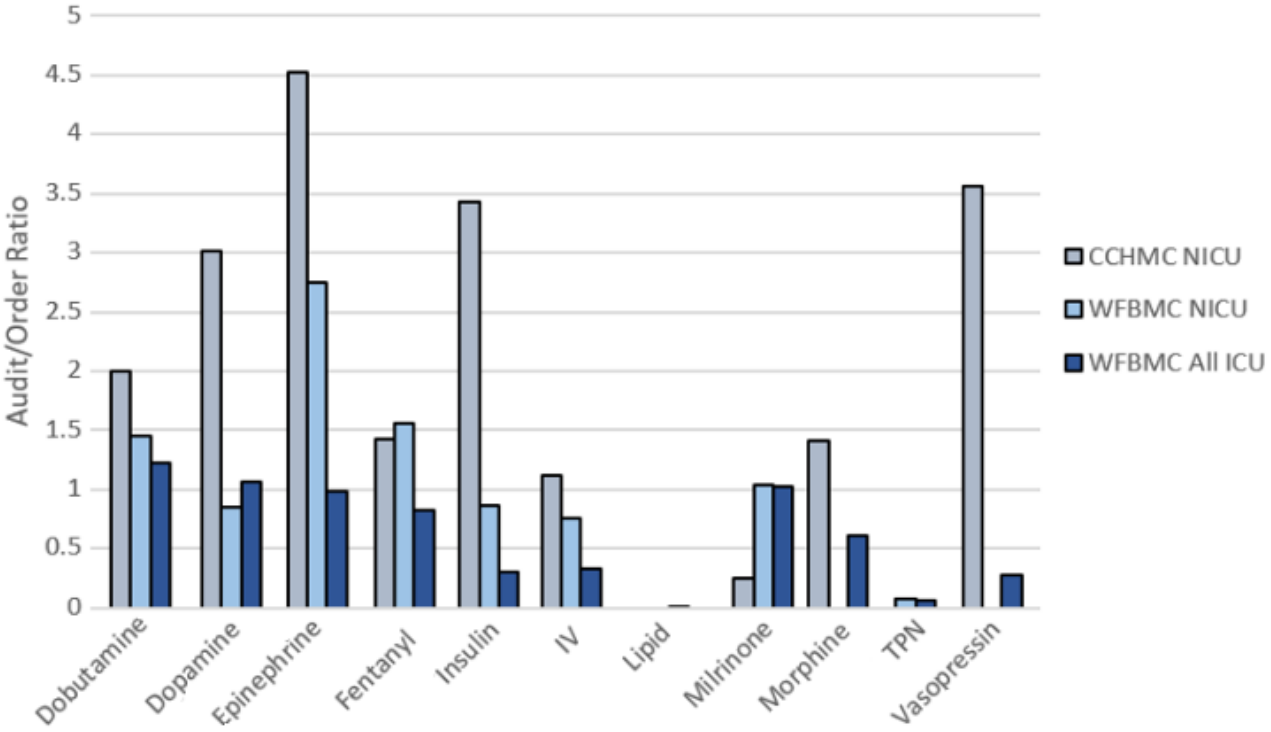
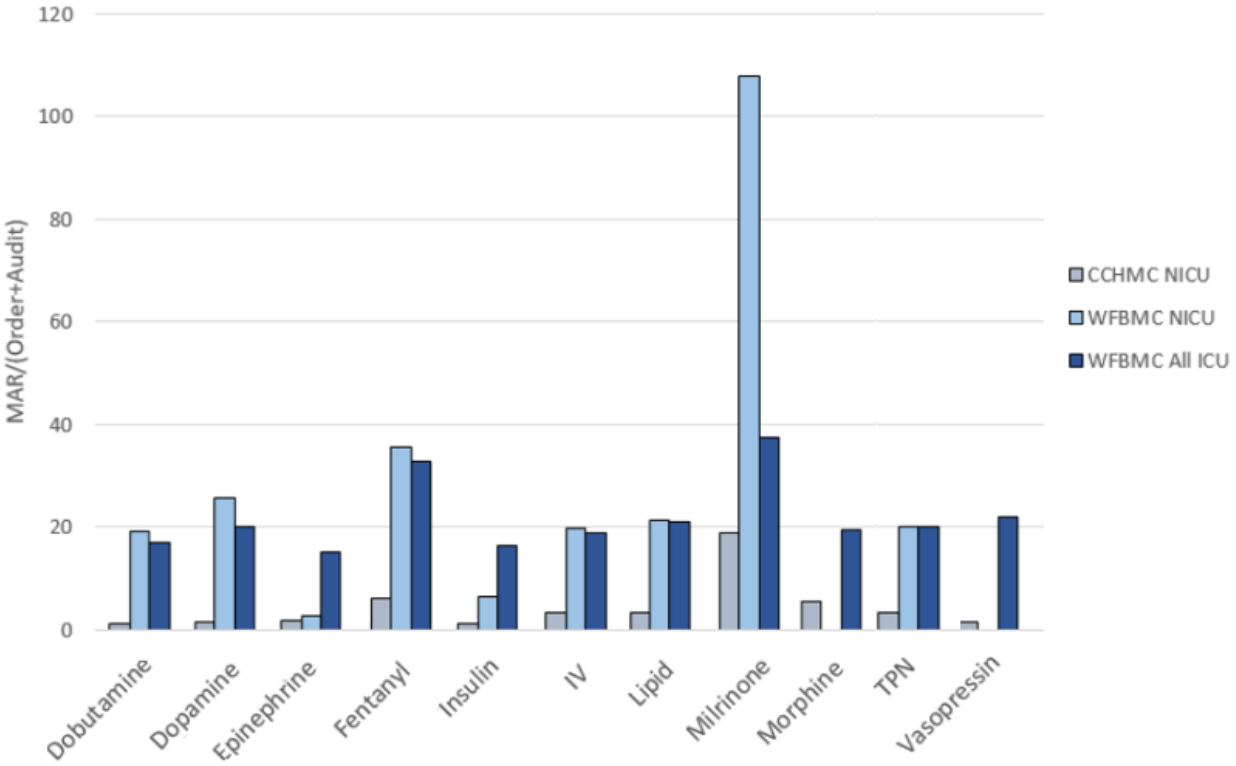


Figure 3 and Multimedia Appendices 3-5 present MAR/(order + audit) ratios between WFBMC departments and CCHMC NICU. The WFBMC NICU and all ICUs at WFBMC had comparable ratios. When compared to the CCHMC NICU, the ratios for WFBMC were higher for each studied medication. The average ratio for WFBMC NICU was 23.6 and the average for CCHMC was 4.4. The MAR/(order + audit) ratio for milrinone in the WFBMC NICU was higher than the other medications and departments. This is a result of WFBMC NICU's practice to verify the rate of milrinone approximately every hour for the entire duration of the medication.

Figure 3. Comparison of MAR/(order + audit) ratios between the CCHMC NICU, the WFBMC NICU, and WFBMC All ICUs. CCHMC: Cincinnati Children's Hospital Medical Center; ICU: intensive care unit; MAR: medication administration record; NICU: neonatal intensive care unit; WFBMC: Wake Forest Baptist Medical Center.



Phase 2: Comparison of the MED.Safe Outputs to the Data From Another EHR Instance at the Second Site

Table 2 presents the discrepancy rate output by MED.Safe for each studied medication. Compared to the baseline discrepancy rates from CCHMC NICU, 5 out of 9 medications used at WFBMC NICU (excluding vasopressin and morphine that did not have orders) showed close discrepancy rates, with less than 1% difference. Epinephrine had similar discrepancy rates, with less than 3% difference. However, the discrepancy rates for insulin, dobutamine, and dopamine were exceptionally large, with over 5% difference. Compared to WFBMC NICU, the discrepancy rates at all WFBMC ICUs tended to deviate more from CCHMC NICU.

Table 2. A comparison of medication administration discrepancy rates generated by MED.Safe at Wake Forest Baptist Medical Center and Cincinnati Children's Hospital Medical Center during the study period.

Medication	Discrepancy rate at all ICUs[a] in WFBMC[b], %	Discrepancy rate at NICU[c] in WFBMC, %	Discrepancy rate at NICU in CCHMC[d], %
Dobutamine	7.9	19.8	0.0
Dopamine	6.7	6.0	0.9
Epinephrine	20.9	4.7	2.1
Fentanyl	5.9	0.5	0.3
Insulin	41.7	59.3	4.3
Intravenous fluids	1.1	1.7	2.5
Lipids	0.1	0.0	0.1
Milrinone	1.1	0.3	0.0
Morphine	6.7	N/A[e]	0.1
Total parenteral nutrition	1.4	1.4	1.3
Vasopressin	2.1	N/A	2.3

[a]ICU: intensive care unit.

[b]WFBMC: Wake Forest Baptist Medical Center.

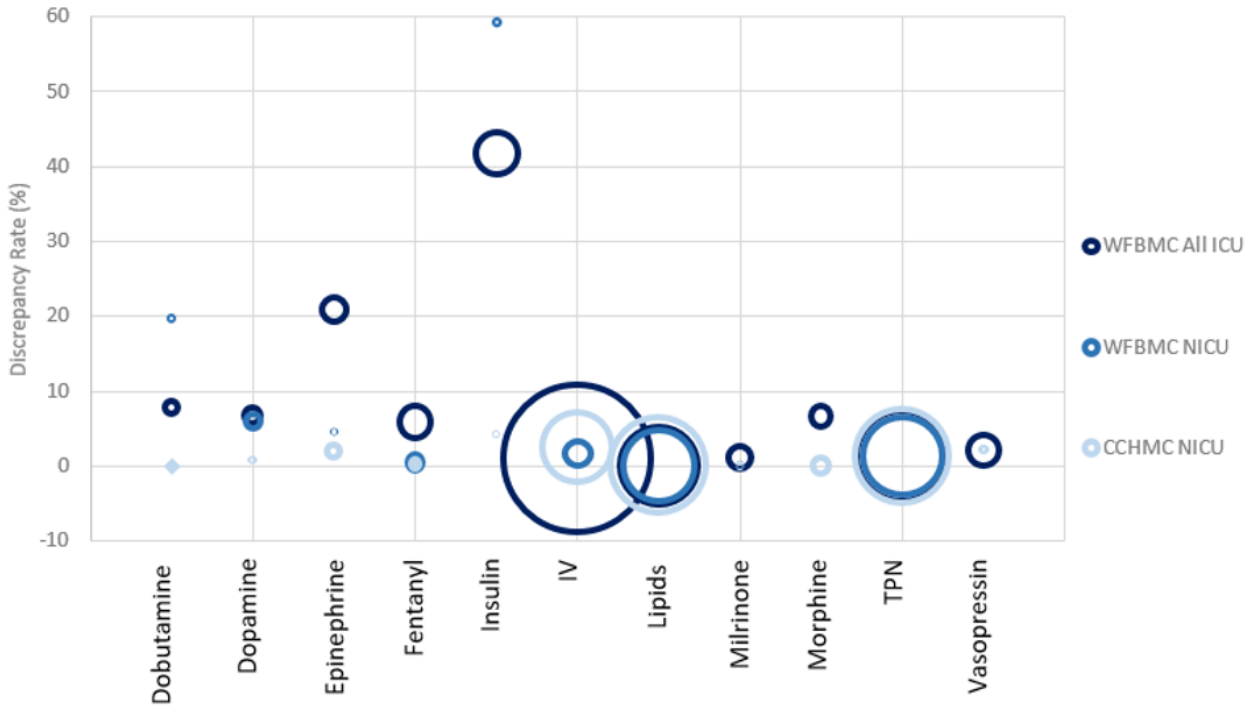
[c]NICU: neonatal intensive care unit.

[d]CCHMC: Cincinnati Children's Hospital Medical Center.

[e]N/A: not applicable. In 2018, no orders for continuous morphine or vasopressin were placed in the WFBMC NICU.

Figure 4 further depicts the relationship between site, discrepancy rate, and medication. A circle size represents the number of orders for a medication during the study period while plotting the discrepancy rate by medication and institutional site. For nearly all medications, the CCHMC NICU had lower discrepancy rates when compared to WFBMC sites and a larger number of orders when compared to the WFBMC NICU specifically. We observed that the outliers in discrepancy rates (epinephrine, dopamine, dobutamine, and insulin) were often due to a small number of orders as represented by the small circle radius.

Figure 4. A comparison of discrepancy rates by medication and number of orders between (A) WFBMC All ICUs, (B) WFBMC NICU, and (C) CCHMC NICU. Circle radius correlates with the number of medication orders for the sites. CCHMC: Cincinnati Children's Hospital Medical Center; ICU: intensive care unit; NICU: neonatal intensive care unit; WFBMC: Wake Forest Baptist Medical Center.



Phase 3: Analysis of System Generalizability and Areas of Improvement

We further investigated the medications with discrepancy rates that substantially deviated from the CCHMC baseline. Three primary causes for the deviation of discrepancy rates were identified: (1) range-based dosing (a common prescribing practice); (2) a data extraction, transforming, and loading issue causing initial order values to be overwritten in the data (a technical data processing issue); and (3) verbal ordering practices (site-specific prescribing practice).

At WFBMC, some medication orders are written as a dosing range (eg, insulin 1-10 Units/hr, with an associated titration protocol) rather than as a discrete dose (eg, insulin 1 Units/hr, titrate by 0.5 Units/hr). Because MED.Safe expects a determinate dose for high-risk IV medications per guidelines at CCHMC, the dosing range practice resulted in very high levels of discrepancies for some medications (eg, insulin) at WFBMC, as seen in Table 2. Figure 5 demonstrates an example system output for an order with a dosing range, including the order, audit, and MARs for a single patient spanning 2 calendar days. After reviewing the patient chart, it was discovered that the original order in the EHR was set to a range of 1-10 Units/hr and was changed to 1-20 Units/hr approximately 6 hours later. However, the MED.Safe system expected a discrete dose for insulin and converted the dosing range to a single value, accepting only the lower-bound range value as an order dose/rate input despite the original physician order for 1-10 Units/hr. Consequently, it marked all of the MAR dose/rate values as causing discrepancies in this single patient. This is a technical limitation of the system design. If the system had been able to accommodate dosing ranges in orders, it should have analyzed the MARs appropriately and avoided false-positive alerts.

Figure 5. Example of a dosing range order interpretation issue by the algorithm. In this example, orders placed with dosing ranges are not interpreted correctly by the system in place to detect medical administration discrepancies. The algorithms, in their current state, do not expect a dosing range and mark the MAR as a discrepancy if the value doesn't match the first value in the order dose range. Subsequent titrations that would fall within the acceptable range of the order are erroneously identified as discrepancies by the algorithm. *The Order Dose/Rate in this figure represents the value that the algorithm parses from the original order. In the instance of orders being placed with a dose range (ie, 1-10 Units/hr), the algorithms only parse and use the first value of the dose range. MAR: medication administration record.

Algorithm Output	Entry Type	Time Stamp	Medication Name	Order Dose/Rate	MAR Dose/Rate	Audit Values
	ORDER	11/1/2018 14:42	insulin infusion 1 unit/mL (100 units/100mL)	1.0 Units/hr *		
Discrepancy	MAR	11/1/2018 15:20	insulin infusion 1 unit/mL (100 units/100mL)		4.0 Units/hr	
Discrepancy	MAR	11/1/2018 17:00	insulin infusion 1 unit/mL (100 units/100mL)		8.0 Units/hr	
Discrepancy	MAR	11/1/2018 19:11	insulin infusion 1 unit/mL (100 units/100mL)		9.0 Units/hr	
	AUDIT	11/1/2018 20:27				1-10 --> 1-20
Discrepancy	MAR	11/1/2018 21:00	insulin infusion 1 unit/mL (100 units/100mL)		12.0 Units/hr	
Discrepancy	MAR	11/1/2018 21:15	insulin infusion 1 unit/mL (100 units/100mL)		14.0 Units/hr	
Discrepancy	MAR	11/1/2018 22:14	insulin infusion 1 unit/mL (100 units/100mL)		15.0 Units/hr	
Discrepancy	MAR	11/1/2018 23:02	insulin infusion 1 unit/mL (100 units/100mL)		16.0 Units/hr	
Discrepancy	MAR	11/2/2018 0:00	insulin infusion 1 unit/mL (100 units/100mL)		16.0 Units/hr	
Discrepancy	MAR	11/2/2018 0:24	insulin infusion 1 unit/mL (100 units/100mL)		16.5 Units/hr	

The second cause of deviation is related to an issue where original order doses/rates were overwritten or replaced by each new audit value, a consequence of the data extraction, transforming, and loading operations of the EHR software. We previously reported on this phenomenon in detail; it is the result of how the proprietary EHR system updates and stores audited order values in the retrospective database [22]. Figure 6 presents an example of this phenomenon. The original order value should be "5.0 Units/hr" (as evidenced by the first audit that changed dose from 5 to 4) but was listed as "3.0 Units/hr" that reflected the last dose modification (the second audit). Consequently, the first MAR was marked as discrepant. This issue resulted in inflated discrepancy rates because the first MAR could always be marked as discrepant if the original order value was no longer presented in our data. This data extraction, transforming, and loading pattern was confirmed by the team's suspicions upon inspecting order values in the real-time production EHR system and comparing them to the retrospective data extracts. Astute readers may also notice that only the first MAR was considered discrepant by the system in Figure 6. This is because the system implements a "check the value with previous MAR data" logic that overrides subsequent discrepancy calls when the MAR values do not change in order to avoid overcalling discrepancies. As such, the first is considered a discrepancy, while subsequent consecutive MARs do not trigger a discrepancy to be called, by design.

Figure 6. Example of an "order/audit value overwriting" issue leading to false positive calls from the system. Due to an ETL process, the original order value is repeatedly overwritten by the newer order audit values and ends up with the value of the last order audit record. When compared to the MAR documentations (which are correct), the false value in the order causes the algorithms to 'detect' a discrepancy, which is a false positive. ETL: extract, transform, load; MAR: medication administration record.

Algorithm Output	Entry Type	Time Stamp	Medication Name	Order Dose/Rate	MAR Dose/Rate	Audit Values
	ORDER	11/1/2018 14:42	DOBUTamine (DOBUTREX) infusion 1mg/mL (250 mg/250mL)	3.0 Units/hr		
Discrepancy	MAR	11/1/2018 15:20	DOBUTamine (DOBUTREX) infusion 1mg/mL (250 mg/250mL)		5.0 Units/hr	
Dose Checked	MAR	11/1/2018 17:00	DOBUTamine (DOBUTREX) infusion 1mg/mL (250 mg/250mL)		5.0 Units/hr	
Dose Checked	MAR	11/1/2018 19:11	DOBUTamine (DOBUTREX) infusion 1mg/mL (250 mg/250mL)		5.0 Units/hr	
	AUDIT	11/1/2018 20:27				5--> 4
Dose Checked	MAR	11/1/2018 21:00	DOBUTamine (DOBUTREX) infusion 1mg/mL (250 mg/250mL)		4.0 Units/hr	
Dose Checked	MAR	11/1/2018 21:15	DOBUTamine (DOBUTREX) infusion 1mg/mL (250 mg/250mL)		4.0 Units/hr	
Dose Checked	MAR	11/1/2018 22:14	DOBUTamine (DOBUTREX) infusion 1mg/mL (250 mg/250mL)		4.0 Units/hr	
	AUDIT	11/1/2018 20:27				4 --> 3
Dose Checked	MAR	11/1/2018 23:02	DOBUTamine (DOBUTREX) infusion 1mg/mL (250 mg/250mL)		3.0 Units/hr	
Dose Checked	MAR	11/2/2018 0:00	DOBUTamine (DOBUTREX) infusion 1mg/mL (250 mg/250mL)		3.0 Units/hr	
Dose Checked	MAR	11/2/2018 0:24	DOBUTamine (DOBUTREX) infusion 1mg/mL (250 mg/250mL)		3.0 Units/hr	

Lastly, there were discrepancies associated with changes to dosage (manifested as MAR documentations) that occurred greater than 30 minutes before the order was entered into the EHR. Such might occur as a result of an emergency during which a verbal order at the bedside is performed but not timely documented in the EHR. As such, the system implemented a 30-minute time window to account for these known lags in documentation due to verbal ordering while meeting the institutional expectations. This phenomenon is depicted in Figure 7, where the rate was changed to "4.0 Units/hr" 76 minutes before the order was modified. By reviewing the patient chart, we confirmed that the dose was changed via a verbal order and the administration was correct. However, the system marked the corresponding MAR as a discrepancy given that there was no audit or new order entered into the EHR for over 30 minutes after the administration. As a quick sensitivity analysis, we modified the algorithms to accept orders within a 60-minute time window; a comparison of discrepancy rates demonstrated a minimal impact, with rates changing less than 0.142% across all medications.

Figure 7. Example of the delayed entry of a verbal order causing a discrepancy to be detected. A verbal order was given at the bedside and the medication was appropriately adjusted, but the order was not documented until after the MAR documentation was placed. The algorithms allow a 30-minute window for verbal orders to be entered before calling a discrepancy, but in this example the order audit for the verbal order rate was not entered until 76 minutes later. MAR: medication administration record.

Algorithm Output	Entry Type	Time Stamp	Medication Name	Order Dose/Rate	MAR Dose/Rate	Audit Values
	ORDER	11/1/2018 14:42	fentaNYL (SUBLIMAZE) 10mcg/mL in dextrose 5% 20mL infusion	3.0 Units/hr		
Dose Checked	MAR	11/1/2018 15:20	fentaNYL (SUBLIMAZE) 10mcg/mL in dextrose 5% 20mL infusion		3.0 Units/hr	
Dose Checked	MAR	11/1/2018 17:00	fentaNYL (SUBLIMAZE) 10mcg/mL in dextrose 5% 20mL infusion		3.0 Units/hr	
Discrepancy	MAR	11/1/2018 19:11	fentaNYL (SUBLIMAZE) 10mcg/mL in dextrose 5% 20mL infusion		4.0 Units/hr	
	AUDIT	11/1/2018 20:27				3--> 4
Dose Checked	MAR	11/1/2018 21:00	fentaNYL (SUBLIMAZE) 10mcg/mL in dextrose 5% 20mL infusion		4.0 Units/hr	
Dose Checked	MAR	11/1/2018 21:15	fentaNYL (SUBLIMAZE) 10mcg/mL in dextrose 5% 20mL infusion		4.0 Units/hr	
Dose Checked	MAR	11/1/2018 22:14	fentaNYL (SUBLIMAZE) 10mcg/mL in dextrose 5% 20mL infusion		4.0 Units/hr	
	AUDIT	11/1/2018 20:27				4 --> 3
Dose Checked	MAR	11/1/2018 23:02	fentaNYL (SUBLIMAZE) 10mcg/mL in dextrose 5% 20mL infusion		3.0 Units/hr	
Dose Checked	MAR	11/2/2018 0:00	fentaNYL (SUBLIMAZE) 10mcg/mL in dextrose 5% 20mL infusion		3.0 Units/hr	
Dose Checked	MAR	11/2/2018 0:24	fentaNYL (SUBLIMAZE) 10mcg/mL in dextrose 5% 20mL infusion		3.0 Units/hr	

Phase 4: Suggested Customization of the System or Clinical Workflows to Enable Better Detection of Medication Administration Errors

The system found discrepancies in medication administration that were attributed to both technical and clinical factors, which contributed to the initial poor fit of discrepancy detection on some medications at the implementation site (WFBMC). To overcome these barriers to successful implementation, the algorithms should be customized to adapt to the local institution. As an initiative, we customized the algorithms with a software update to solve 1 of the 3 major sources of false-positive discrepancies: order/audit value overwrites (the second issue identified in phase 3).

The investigative team (all study authors) implemented a patch to MED.Safe to recover the original order values from the sequences of medication use data. We then re-executed the updated system on the data used in the initial analysis to study its effects. Figure 8 and Table 3 demonstrate its effects in decreasing the output discrepancy rates for fentanyl, dobutamine, epinephrine, milrinone, and IV fluids. The other medications retained their discrepancy rates prior to the update, implying that they were not affected by order/audit value overwriting errors. As a result of this update, discrepancy rates from the WFBMC NICU became comparable to those from the CCHMC NICU for 5 of 9 medications with orders. The remaining medications maintained rates approximately twofold higher than the baseline CCHMC rates. Although this customization corrected for order/audit value overwriting errors, false-positive discrepancies persist as a result of delayed documentation of verbal orders and dosing range issues.

Figure 8. A comparison of discrepancy rates between (A) CCHMC NICU, (B) WFBMC NICU using the updated MED.Safe, and (C) WFBMC using the original MED.Safe. CCHMC: Cincinnati Children's Hospital Medical Center; IV: intravenous; NICU: neonatal intensive care unit; TPN: total parenteral nutrition; WFBMC: Wake Forest Baptist Medical Center.

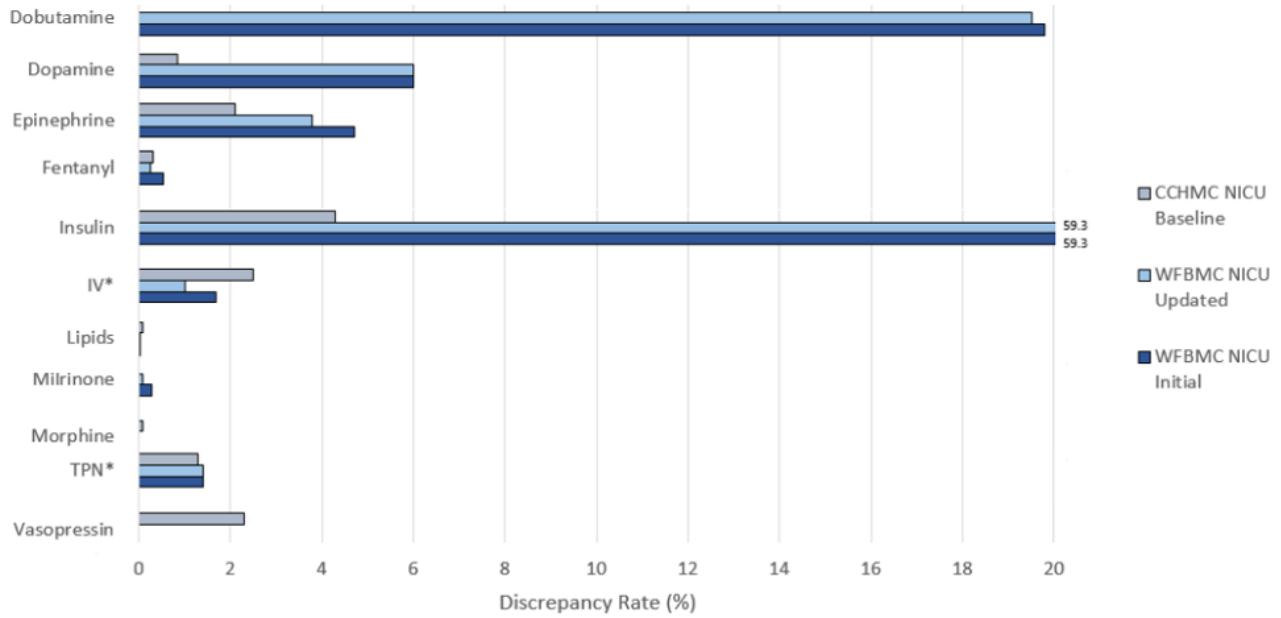


Table 3. Discrepancy rates of medication administration in the NICU before and after implementation of a software update at WFBMC in comparison to the site of development CCHMC.

Medication	*Initial* discrepancy rates in WFBMC[a] NICU[b], %	*Updated* discrepancy rates in WFBMC NICU, %	Absolute change in discrepancy rate, %	Initial discrepancy rates in CCHMC[c] NICU, %
Dobutamine	19.8	19.5	–0.3	0.0
Dopamine	6.0	6.0	0.0	0.9
Epinephrine	4.7	3.8	–0.9	2.1
Fentanyl	0.5	0.25	–0.25	0.3
Insulin	59.3	59.3	0.0	4.3
Intravenous fluids	1.7	1.0	–0.7	2.5
Lipids	0.0	0.0	0.0	0.1
Milrinone	0.3	0.19	–0.11	0.0
Morphine	N/A[d]	N/A	N/A	0.1
Total parenteral nutrition	1.4	1.4	0.0	1.3
Vasopressin	N/A	N/A	N/A	2.3

[a]WFBMC: Wake Forest Baptist Medical Center.

[b]NICU: neonatal intensive care unit.

[c]CCHMC: Cincinnati Children's Hospital Medical Center.

[d]N/A: not applicable. In 2018, no orders for continuous morphine or vasopressin were placed in the WFBMC NICU.

Discussion

Principal Findings

The ability to effectively implement the MED.Safe package at a second site is the first critical step toward creating a scalable and impactful solution for detecting and mitigating medication errors. This study investigated the feasibility and success of implementation for MED.Safe at a second site distinct from the origin of the software. The system outputs, such as descriptive statistics from local EHR data and discrepancy rates, served as a means to understand the institutional clinical workflows and prescribing patterns, assess the system generalizability, and help develop site-specific customizations. It is our hope that this study will serve as a guide for future institutions to efficiently assess the applicability of MED.Safe and lead to its implementation in an effort that maximizes medication safety in clinical settings.

Consideration of the clinical policies and workflows surrounding medication ordering, auditing, and MARs was vital in determining the feasibility of MED.Safe implementation at

WFBMC. We observed that the NICU, PICU, and adult medical ICU were fundamentally different in their prescribing and auditing patterns (Table 1 and Multimedia Appendices 1-5). The WFBMC NICU had the most orders, audits, and MARs for the studied medications, reflecting the fact that MED.Safe was originally designed for an NICU setting that did not include common adult vasopressors such as norepinephrine. The adult medical ICUs had far less medication orders despite greater bed count. This was partially due to the fact that norepinephrine would have contributed 1466 orders to the total order count in this environment if an algorithm was available in MED.Safe to detect discrepancies; if included, the descriptive statistics would have more closely correlated with the bed count across the units. Regardless, the descriptive statistics output by the system allowed us to quickly understand, at the aggregate level, how prevalent the medications and MAR documentations were in different clinical environments and where the system may be the most useful. For instance, we found from the descriptive statistics that the NICU did not have vasopressin and morphine orders. As such, the algorithms for those medications not prescribed would not have any utility in the NICU and implementing MED.Safe there would yield no benefit. Beyond the basic descriptive characteristics, the comparison between audit/order ratios at WFBMC and CCHMC (Figure 2 and Multimedia Appendices 1-5) allowed us to understand the differences in prescribing workflows between the institutions. The lower audit/order ratios at WFBMC in comparison to CCHMC lead us to believe that WFBMC tends to create new orders for medication dose/rate changes, whereas CCHMC modifies existing orders for such changes more frequently. The more frequent use of order dose range intervals in combination with practices of documenting MAR rate to verify values very frequently may have contributed to the higher MAR/(order + audit) ratios at the WFBMC NICU despite fewer orders and audits overall (compared to CCHMC NICU). Our findings highlight potential practice differences across institutions, which may change the distribution of discrepancy rates, introduce additional opportunities to identify errors, or suggest the need for customizations to the MED.Safe system.

In phase 2, we executed the discrepancy detection algorithms of the software and analyzed the output discrepancy rates at WFBMC (Table 2). The rates at WFBMC aligned well with the ones at CCHMC for the majority of the studied medications. However, the rates at WFBMC varied widely, ranging from 0% to 59%, compared to CCHMC rates that ranged from 0% to 4.3%. The results suggested that the algorithms generalized well to the data and clinical practices for some medications but fit poorly for the others. Further inspection for the poorly performing medications in phase 3 identified 3 phenomena that contributed to the inflated discrepancy rates: range-based dosing, order/audit value overwriting in the data, and verbal ordering practices.

WFBMC uses dosing ranges to allow for bedside adjustment of a medication so long as the dosing is in range of the order and follows ancillary instructions, protocols, or policies. Such practice is common in adult medication prescribing, particularly in the administration of insulin, where dosing might shift within a given range depending on the trend of blood glucose values or intake of food. However, the algorithms were not equipped to deal with ordering ranges because at CCHMC site-specific practices required that an order dose/rate should be determinate and an audit (modification) be documented each time a dose/rate was changed. Consequently, WFBMC had comparatively fewer audits and more discrepancies for values within the acceptable dosing range. This difference in site-specific practices resulted in high discrepancy rates for insulin (59.3% at WFBMC NICU versus 4.3% at CCHMC NICU). A quick glance at the descriptive data and discrepancy rates generated by the algorithms will cue future customizations as to the cause of the high rates and shortcut much of the time spent in exploration and validation.

Second, the investigative team (all study authors) determined that the institutional EHR was overwriting the original order values with each new audit. The overwriting resulted in a notable amount of false-positive discrepancies on the first MARs. We were able to overcome this EHR-derived technical limitation with a software update that recovered the original order dose/rate by reasoning through from the sequences of order-audit data.

Lastly, a portion of discrepancies originated from dose/rate changes with delayed order documentation. This often occurs in emergency settings where verbal orders are first placed, while electronic orderings are documented after the care is delivered. The “grace period” for entering the electronic orders varies between institutions based on the site-specific clinical practices. Operating under verbal orders without proper documentation and procedure is high risk, and it creates a blind spot for errors that may have occurred but lacked the appropriate data for the system to detect them. The inability to identify medication errors during this elapsed time might lead to perpetuation of similar errors for an extended period, ultimately lessening the value of the system in identifying errors efficiently. A change in policy to eliminate the practice of verbal ordering is one potential solution, but this does not fit with the reality of clinical practice. Another solution is to adapt the system to the “grace period” that complies with local policies surrounding verbal ordering. For instance, the MED.Safe algorithms adopted a period of 30 minutes given the institutional expectations at CCHMC, which could be extended to 45-60 minutes to comply with WFBMC’s verbal ordering policies. In our quick sensitivity analysis we found that an extension to a 60-minute window, however, did not greatly reduce the discrepancy rate. This effect appears to be site specific as we have seen this change decrease rates to a greater degree at other sites. In the future, we will add this customizable feature to the software so that the grace period can be adjusted depending on the care setting and local policy. This will also allow an automated version of the sensitivity analysis. Ultimately, the system could be more flexible and customizable to fit each institution and even department that varies in health care policy and procedures surrounding the medication use life cycle.

In phase 4, we addressed the order/audit value overwriting issue through a software update. It reduced false-positive discrepancies output by the system for most of the studied medications. The remaining 2 medications (dobutamine and insulin) with discrepancy rates notably higher than baseline CCHMC rates are largely due to the range-based dosing issue.

Further reduction in false-positive discrepancies can therefore be obtained by addressing the other 2 issues, range-based dosing and verbal ordering practices. Efforts to do so are planned for future work.

Our study suggested that it was feasible to implement MED.Safe in a setting external to the development environment. However, the software package did not account for all the differences in medication administration practices at the implementation site, with a resultant impact on its performance. The identified barriers to proper fitting of the system can be overcome through both clinical practice change/policy reform and the addition of algorithm customizations where appropriate. We were able to identify targets for algorithm customization to account for these practices and to address one of those issues efficiently. These efforts have greatly advanced our knowledge of the portability of the MED.Safe and have shown us what work is left to do in order to further improve its generalizability.

Conclusions

The implementation of the MED.Safe system at a second site was a feasible and efficient way to track medical administration discrepancies. Analysis of medication use data and discrepancy rates output by the system revealed local medication prescribing patterns, and comparison against implementation at the original site suggested areas of both good and poor fit. Overall fit was enhanced through the implementation of a software update. To maximize efficiency in accurately detecting and correcting medication errors, modifications must be made to both the MED.Safe software package and suboptimal clinical practices. Such modifications should increase the system's customizability to the local clinical workflows and policies, ultimately improving its accuracy and generalization for external use.

Acknowledgments

The investigative team thanks Meredith Hollinger, Wake Forest Pharmacy Shared Services, for her assistance in providing data for the study. In addition, Lara Kanbar from the CCHMC Division of Biomedical Informatics provided a critical review of the near-final manuscript. This research was supported by the National Institute of Health under award number R01LM012230 (sponsored by the National Library of Medicine).

Conflicts of Interest

None declared.

Multimedia Appendix 1
A comparison of medication ordering, auditing, and MAR data by drug generated by MED.Safe at Wake Forest Baptist Medical Center. MAR: medication administration record.
[DOCX File , 24 KB - medinform_v8i12e22031_app1.docx]

Multimedia Appendix 2
A comparison of medication ordering, auditing, and MAR data by generated by MED.Safe at Wake Forest Baptist Medical Center. MAR: medication administration record.
[DOCX File , 23 KB - medinform_v8i12e22031_app2.docx]

Multimedia Appendix 3
Descriptive statistics of orders, audits, and medication administration record data at all Wake Forest Baptist Health ICUs during the study period. ICU: intensive care unit.
[DOCX File , 14 KB - medinform_v8i12e22031_app3.docx]

Multimedia Appendix 4
Descriptive statistics of orders, audits, and medication administration record data in the Wake Forest Baptist Health NICU during the study period. NICU: neonatal intensive care unit.
[DOCX File , 14 KB - medinform_v8i12e22031_app4.docx]

Multimedia Appendix 5
Descriptive statistics of orders, audits, and medication administration record data in the Cincinnati Children's Hospital NICU during the study period. NICU: neonatal intensive care unit.
[DOCX File , 14 KB - medinform_v8i12e22031_app5.docx]

References

1. Hartwig SC, Denger SD, Schneider PJ. Severity-indexed, incident report-based medication error-reporting program. Am J Hosp Pharm 1991 Dec;48(12):2611-2616. [Medline: 1814201]

2. Fontan J, Maneglier V, Nguyen VX, Loirat C, Brion F. Medication errors in hospitals: computerized unit dose drug dispensing system versus ward stock distribution system. Pharm World Sci 2003 Jun;25(3):112-117. [doi: 10.1023/a:1024053514359] [Medline: 12840964]
3. Miller MR, Clark JS, Lehmann CU. Computer based medication error reporting: insights and implications. Qual Saf Health Care 2006 Jun;15(3):208-213 [FREE Full text] [doi: 10.1136/qshc.2005.016733] [Medline: 16751472]
4. Aronson JK. Medication errors: what they are, how they happen, and how to avoid them. QJM 2009 Aug;102(8):513-521. [doi: 10.1093/qjmed/hcp052] [Medline: 19458202]
5. Keers RN, Williams SD, Cooke J, Ashcroft DM. Prevalence and nature of medication administration errors in health care settings: a systematic review of direct observational evidence. Ann Pharmacother 2013 Feb;47(2):237-256. [doi: 10.1345/aph.1R147] [Medline: 23386063]
6. Kale A, Keohane CA, Maviglia S, Gandhi TK, Poon EG. Adverse drug events caused by serious medication administration errors. BMJ Qual Saf 2012 Nov;21(11):933-938 [FREE Full text] [doi: 10.1136/bmjqs-2012-000946] [Medline: 22791691]
7. Shah PK, Irizarry J, O'Neill S. Strategies for Managing Smart Pump Alarm and Alert Fatigue: A Narrative Review. Pharmacotherapy 2018 Aug;38(8):842-850. [doi: 10.1002/phar.2153] [Medline: 29883535]
8. Morriss FH, Abramowitz PW, Nelson SP, Milavetz G, Michael SL, Gordon SN, et al. Effectiveness of a barcode medication administration system in reducing preventable adverse drug events in a neonatal intensive care unit: a prospective cohort study. J Pediatr 2009 Mar;154(3):363-8, 368.e1. [doi: 10.1016/j.jpeds.2008.08.025] [Medline: 18823912]
9. Nash IS, Rojas M, Hebert P, Marrone SR, Colgan C, Fisher LA, et al. Reducing excessive medication administration in hospitalized adults with renal dysfunction. Am J Med Qual 2005;20(2):64-69. [doi: 10.1177/1062860604273752] [Medline: 15851383]
10. Ghaleb MA, Barber N, Franklin BD, Wong ICK. The incidence and nature of prescribing and medication administration errors in paediatric inpatients. Arch Dis Child 2010 Feb;95(2):113-118. [doi: 10.1136/adc.2009.158485] [Medline: 20133327]
11. Garrett PR, Sammer C, Nelson A, Paisley KA, Jones C, Shapiro E, et al. Developing and implementing a standardized process for global trigger tool application across a large health system. Jt Comm J Qual Patient Saf 2013 Jul;39(7):292-297. [doi: 10.1016/s1553-7250(13)39041-2] [Medline: 23888638]
12. Ni Y, Lingren T, Hall ES, Leonard M, Melton K, Kirkendall ES. Designing and evaluating an automated system for real-time medication administration error detection in a neonatal intensive care unit. J Am Med Inform Assoc 2018 May 01;25(5):555-563 [FREE Full text] [doi: 10.1093/jamia/ocx156] [Medline: 29329456]
13. Murff HJ, Patel VL, Hripcsak G, Bates DW. Detecting adverse events for patient safety research: a review of current methodologies. J Biomed Inform 2003;36(1-2):131-143 [FREE Full text] [Medline: 14552854]
14. Jacobs B. Electronic medical record, error detection, and error reduction: a pediatric critical care perspective. Pediatr Crit Care Med 2007 Mar;8(2 Suppl):S17-S20. [doi: 10.1097/01.PCC.0000257484.86356.39] [Medline: 17496828]
15. Naessens JM, Campbell CR, Huddleston JM, Berg BP, Lefante JJ, Williams AR, et al. A comparison of hospital adverse events identified by three widely used detection methods. Int J Qual Health Care 2009 Aug;21(4):301-307. [doi: 10.1093/intqhc/mzp027] [Medline: 19617381]
16. Härkänen M, Turunen H, Vehviläinen-Julkunen K. Differences Between Methods of Detecting Medication Errors: A Secondary Analysis of Medication Administration Errors Using Incident Reports, the Global Trigger Tool Method, and Observations. J Patient Saf 2020 Jun;16(2):168-176. [doi: 10.1097/PTS.0000000000000261] [Medline: 27010325]
17. Kirkendall ES, Kloppenborg E, Papp J, White D, Frese C, Hacker D, et al. Measuring adverse events and levels of harm in pediatric inpatients with the Global Trigger Tool. Pediatrics 2012 Nov;130(5):e1206-e1214. [doi: 10.1542/peds.2012-0179] [Medline: 23045558]
18. Classen D, Li M, Miller S, Ladner D. An Electronic Health Record-Based Real-Time Analytics Program For Patient Safety Surveillance And Improvement. Health Aff (Millwood) 2018 Nov;37(11):1805-1812. [doi: 10.1377/hlthaff.2018.0728] [Medline: 30395491]
19. Classen DC, Pestotnik SL, Evans RS, Burke JP. Computerized surveillance of adverse drug events in hospital patients. JAMA 1991 Nov 27;266(20):2847-2851. [Medline: 1942452]
20. Li Q, Kirkendall ES, Hall ES, Ni Y, Lingren T, Kaiser M, et al. Automated detection of medication administration errors in neonatal intensive care. J Biomed Inform 2015 Oct;57:124-133 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.012] [Medline: 26190267]
21. Li Q, Melton K, Lingren T, Kirkendall ES, Hall E, Zhai H, et al. Phenotyping for patient safety: algorithm development for electronic health record based automated adverse event and medical error detection in neonatal intensive care. J Am Med Inform Assoc 2014;21(5):776-784 [FREE Full text] [doi: 10.1136/amiajnl-2013-001914] [Medline: 24401171]
22. Kirkendall ES, Ni Y, Lingren T, Leonard M, Hall ES, Melton K. Data Challenges With Real-Time Safety Event Detection And Clinical Decision Support. J Med Internet Res 2019 May 22;21(5):e13047 [FREE Full text] [doi: 10.2196/13047] [Medline: 31120022]

Abbreviations

CCHMC: Cincinnati Children's Hospital Medical Center

EHR: electronic health record
ICU: intensive care unit
IV: intravenous
MAE: medication administration error
MAR: medication administration record
NICU: neonatal intensive care unit
PICU: pediatric intensive care unit
TPN: total parenteral nutrition
WFBMC: Wake Forest Baptist Medical Center

Edited by G Eysenbach; submitted 09.07.20; peer-reviewed by J Chaparro, KM Kuo; comments to author 16.08.20; revised version received 11.10.20; accepted 28.10.20; published 02.12.20.

Please cite as:
Kirkendall E, Huth H, Rauenbuehler B, Moses A, Melton K, Ni Y
The Generalizability of a Medication Administration Discrepancy Detection System: Quantitative Comparative Analysis
JMIR Med Inform 2020;8(12):e22031
URL: https://medinform.jmir.org/2020/12/e22031
doi:10.2196/22031
PMID:33263548

Original Paper

Unpacking Prevalence and Dichotomy in Quick Sequential Organ Failure Assessment and Systemic Inflammatory Response Syndrome Parameters: Observational Data–Driven Approach Backed by Sepsis Pathophysiology

Nazmus Sakib[1], MSc; Sheikh Iqbal Ahamed[1], PhD; Rumi Ahmed Khan[2], MD; Paul M Griffin[3], PhD; Md Munirul Haque[4], PhD

[1]Ubicomp Lab, Department of Computer Science, Marquette University, Milwaukee, WI, United States

[2]College of Medicine, University of Central Florida, Orlando, FL, United States

[3]Regenstrief Center for Healthcare Engineering, Purdue University, West Lafayette, IN, United States

[4]RB Annis School of Engineering, University of Indianapolis, Indianapolis, IN, United States

Corresponding Author:
Nazmus Sakib, MSc
Ubicomp Lab
Department of Computer Science
Marquette University
1313 W Wisconsin Ave
Milwaukee, WI, 53233
United States
Phone: 1 4147975981
Email: nazmus.sakib@marquette.edu

Abstract

Background: Considering morbidity, mortality, and annual treatment costs, the dramatic rise in the incidence of sepsis and septic shock among intensive care unit (ICU) admissions in US hospitals is an increasing concern. Recent changes in the sepsis definition (sepsis-3), based on the quick Sequential Organ Failure Assessment (qSOFA), have motivated the international medical informatics research community to investigate score recalculation and information retrieval, and to study the intersection between sepsis-3 and the previous definition (sepsis-2) based on systemic inflammatory response syndrome (SIRS) parameters.

Objective: The objective of this study was three-fold. First, we aimed to unpack the most prevalent criterion for sepsis (for both sepsis-3 and sepsis-2 predictors). Second, we intended to determine the most prevalent sepsis scenario in the ICU among 4 possible scenarios for qSOFA and 11 possible scenarios for SIRS. Third, we investigated the multicollinearity or dichotomy among qSOFA and SIRS predictors.

Methods: This observational study was conducted according to the most recent update of Medical Information Mart for Intensive Care (MIMIC-III, Version 1.4), the critical care database developed by MIT. The qSOFA (sepsis-3) and SIRS (sepsis-2) parameters were analyzed for patients admitted to critical care units from 2001 to 2012 in Beth Israel Deaconess Medical Center (Boston, MA, USA) to determine the prevalence and underlying relation between these parameters among patients undergoing sepsis screening. We adopted a multiblind Delphi method to seek a rationale for decisions in several stages of the research design regarding handling missing data and outlier values, statistical imputations and biases, and generalizability of the study.

Results: Altered mental status in the Glasgow Coma Scale (59.28%, 38,854/65,545 observations) was the most prevalent sepsis-3 (qSOFA) criterion and the white blood cell count (53.12%, 17,163/32,311 observations) was the most prevalent sepsis-2 (SIRS) criterion confronted in the ICU. In addition, the two-factored sepsis criterion of high respiratory rate (≥22 breaths/minute) and altered mental status (28.19%, among four possible qSOFA scenarios besides no sepsis) was the most prevalent sepsis-3 (qSOFA) scenario, and the three-factored sepsis criterion of tachypnea, high heart rate, and high white blood cell count (12.32%, among 11 possible scenarios besides no sepsis) was the most prevalent sepsis-2 (SIRS) scenario in the ICU. Moreover, the absolute Pearson correlation coefficients were not significant, thereby nullifying the likelihood of any linear correlation among the critical parameters and assuring the lack of multicollinearity between the parameters. Although this further bolsters evidence for their dichotomy, the absence of multicollinearity cannot guarantee that two random variables are statistically independent.

Conclusions: Quantifying the prevalence of the qSOFA criteria of sepsis-3 in comparison with the SIRS criteria of sepsis-2, and understanding the underlying dichotomy among these parameters provides significant inferences for sepsis treatment initiatives in the ICU and informing hospital resource allocation. These data-driven results further offer design implications for multiparameter intelligent sepsis prediction in the ICU.

(JMIR Med Inform 2020;8(12):e18352) doi:10.2196/18352

KEYWORDS

sepsis; MIMIC-III; SIRS; qSOFA; pathophysiology; medical internet research; medical informatics; critical care; intensive care unit; multicollinearity

Introduction

Sepsis remains one of the most elusive syndromes in medical science, which is a syndrome induced by infection and associated with biochemical, physiological, and pathological abnormalities as a result of an unregulated response from the human body [1-3]. In the United States, over 1.7 million adults are affected by sepsis, and more than 970,000 patients are admitted to hospitals because of sepsis each year. Sepsis both directly and indirectly contributes to more than 250,000 deaths annually, representing more than 50% of all hospital deaths [2,4-8]. Unfortunately, these excruciating statistics have been exacerbated over recent years, as identified in a two-decade study on US hospitalizations, costs, and disease epidemiology. These statistics reflect an 8.7% annual increase in the incidence of sepsis among hospitalized patients in the United States [5,9,10].

Besides the alarmingly increasing incidence of sepsis and associated mortality rate, the average length of stay in hospitals is considerably higher (approximately 75% higher than that reported for most other conditions) for sepsis patients in the United States, thereby increasing the burden associated with hospital utilization [10-13]. Furthermore, the Agency for Healthcare Research and Quality [14] reported that the average length of stay for patients with sepsis dilated compellingly in 2013, and there was a distinct proportion of patients with severe sepsis cases, including 4.5 days, 6.5 days, and 16.5 days of hospitalization for sepsis, severe sepsis, and septic shock, respectively, according to the systematic inflammatory response syndrome (SIRS) criteria. Moreover, although accounting for 3.6% of hospital stays, sepsis-related care represents 13% of total US hospital costs, resulting in hospital expenses exceeding US $24 billion in 2013. Not surprisingly, in 2013, the cost associated with sepsis management ranked the highest among the admissions for all diseases and medical conditions, followed by osteoarthritis at US $17 billion and childbirth (medical condition) at US $13 billion [15-17]. At present, the hospital costs associated with sepsis still rank first, and sepsis care currently requires more than twice the resources required for other medical conditions [18]. These costs are also expected to be exacerbated in the near future, and will likely approach a 3-fold increase compared to those of other admissions [3,19,20].

This notable increase in mortality rate and annual health care expenditure (affected by the increased length of stay) has made sepsis treatment and research a critical domain in medical internet research and medical informatics, resulting in a recent surge in the related literature [21-24]. Studies have shown that improved and effective methods of early sepsis identification can substantially reduce the severity and epidemiological burden of sepsis in the United States [24-29]. In addition, several authors have recommended that identifying the prevalent risk factor(s), followed by an instant diagnosis, can reduce the cost in treatment workflow, and further scale down the mortality rate for patients with sepsis to some extent [26,30-33]. However, most of these studies have only concentrated on one risk factor at a time for the clinical assessment of sepsis, thereby limiting the probability for sepsis detection as it requires complex reasoning and implications. In many cases, it is apparent that the results are sensitive to subtle variations in definition(s) of sepsis, as well as subjective suspicions of physicians [21,22,34-36].

The recent major release of Medical Information Mart for Intensive Care (MIMIC-III, Version 1.4) is an extensive, single-center, and comprehensive database comprising information pertaining to patients admitted to the critical care units at Beth Israel Deaconess Medical Center in Boston, Massachusetts, including vital signs, laboratory measurements, observations and notes charted by care providers, imaging reports, fluid balance, medications, procedure codes, diagnostic codes, and hospital length of stay [17,21,37,38]. MIMIC-III is a multidisciplinary collaborative effort of the Laboratory for Computational Physiology at MIT, Computer Science and Artificial Intelligence Laboratory at MIT, and Information Systems Department at Beth Israel Deaconess Medical Center. The underlying motivation behind this collaboration is to assure reproducibility and improve the quality of data-driven medical informatics research. The salient features of MIMIC-III (Version 1.4) include that it is the only freely accessible critical care database of its kind in the United States that promotes analysis without additional restriction after accepting the data use agreement.

Furthermore, a critical care dataset with detailed individual patient care information spanning more than a decade empowers medical informatics research and pedagogy around the world. MIMIC-III (Version 1.4) contains data from 58,976 hospital admissions for patients admitted to the critical care units from 2001 to 2012. Personal information is removed, and the original records are shifted and reformatted to ensure that the data are not identifiable to human patients. The database comprises 26 tables linked by identifiers for corresponding patients. Each of the tables is a spreadsheet including information on patient hospital stays and the physiological data collected in the intensive care unit (ICU), along with data dictionaries to explain the observational context. MIMIC-III (Version 1.4) allows for

a variety of data forms, ranging from text interpretations for radiology images to time-stamped physiological measures [21,37]. This open and unrestricted nature of extensive health care data allows for clinical studies to be improved and reproduced in ways that would not otherwise be possible [39]. Hence, MIMIC-III (Version 1.4) can facilitate exploratory and data-driven studies on sepsis, its diagnosis, and treatment in the ICU [17,21].

Sepsis was first formally defined by a 1991 consensus conference as a SIRS to infection in the host [1,40]. According to the then-prevailing definition, sepsis associated with organ dysfunction was referred to as severe sepsis, and severe sepsis followed by sepsis-induced persisting hypotension despite adequate fluid resuscitation was termed as septic shock. Subsequently, considering the limitations of 1991 consensus conference definitions, the 2001 task force extended the list of diagnostic criteria for sepsis [41]. Despite discrepancy in the 1991 interpretation, the 2001 task force could not offer an alternative definition due to lack of supporting evidence; therefore, the sepsis definition remained mostly unchanged from 1991 to 2016 [41,42]. In 2016, a task force comprising experts of sepsis pathobiology, pathophysiology, epidemiology, and clinical trials convened by the Society of Critical Care Medicine along with the European Society of Intensive Care Medicine revised the definition of sepsis and septic shock.

The substantial advances observed in pathobiology, epidemiology, immunology, and intervention management motivated efforts to reexamine the interpretation of sepsis. The definition devised by the 2016 task force has since been supported by 31 international sites [1]. Singer et al [1] concluded that it is necessary to change the perception about sepsis to establish a more reliable predictive indicator of mortality and impact in the survivability of patients. Consequently, the SIRS-based definition was replaced by the quick Sequential Organ Failure Assessment (qSOFA) criteria. The qSOFA suggests three criteria to evaluate patients who are more likely to have a poor outcome due to sepsis: hypotension, altered mental status, and high respiratory rate [21]. In addition to qSOFA, the sepsis-3 definition (given that this was the third updated definition of sepsis) includes the Sepsis-related Organ Failure Assessment (SOFA) for making a sepsis diagnosis. Albeit not substantially, SOFA provides better predictive accuracy with greater consistency compared to qSOFA. However, the intricacy and time-consuming lab tests involved in SOFA have remained poorly understood outside the critical care community since the definition was updated in 2016.

As sepsis is still perceived as a spectrum disease that subsequently ends in organ dysfunction, septic shock is a crucial juncture for multiparameter intelligent sepsis prediction in the ICU. However, we here focus on sepsis defined according to SIRS and qSOFA. We adopted a data-driven approach using MIMIC-III (Version 1.4) to offer unique contributions to the field. First, we aimed to unpack the most prevalent SIRS and qSOFA criteria. Second, we evaluated the most prevalent sepsis scenarios based on SIRS and qSOFA criteria. Third, we investigated the dichotomy among SIRS and qSOFA criteria to establish underlying statistical relations among these predictors, with design implications for predictive modeling. Quantifying the prevalence of the qSOFA criteria (in comparison with SIRS) and understanding the underlying dichotomy of these parameters have important implications for sepsis treatment initiatives in the ICU and for informing hospital resource allocation. Hence, this study has potential to improve preventable deaths from sepsis.

Methods

Theoretical Background

Sepsis Pathophysiology

Sepsis—commonly interpreted as a spectrum disease—ranges from milder symptoms and ends in septic shock, followed by multiple organ dysfunction syndromes. This entire spectrum begins with the introduction of pathogens in the blood vessels, such as gram-positive or gram-negative bacteria, fungi, viruses, and parasites. The appearance of pathogens in the blood vessels makes them no longer sterile; when the white blood cells confront these infective materials (pathogens), they become activated. Consequently, more white blood cells are called in to the site of infection to eradicate the pathogens. Generally, these infective materials exist outside in the interstitial tissue rather than in the bloodstream. Therefore, to access the infective materials and eradicate them, the white blood cells release substances such as nitric oxide. Three events occur once these substances interact with the blood vessels. First, the diameter of the blood vessel expands, resulting in vasodilation. The vasodilation reduces the localized systemic vascular resistance and affects the speed of the blood flow, including the blood flow in the infected area. Second, the permeability of the blood vessels increases so that the immune system can confront the peripheral infective material easily. In the context of this paper, blood pressure—in the mathematical sense—is considered to be the product of cardiac output and systemic vascular resistance, thus affecting tissue perfusion. Hence, the lower the systemic vascular resistance, the lower the blood pressure, and consequently tissue perfusion is reduced [43,44].

The decrease in tissue perfusion is further exacerbated by the increased permeability of the blood vessels since the fluid can reach out and build around the tissue, which eventually makes it challenging for oxygen to diffuse through the fluids and access the cells. This exacerbated tissue perfusion is the cardinal reason behind the shock. Third, when the white blood cells interact with the pathogens, they release lytic enzymes as well as reactive oxygen species to eliminate the infective materials. These enzymes damage not only the pathogens but also the blood vessels to some extent, resulting in serious complications. When the blood vessels are ruptured, proteins are released to cause clotting as a patch due to coagulation factors in the blood. This may initially preclude the blood from spilling into the extravascular space; however, over time, some of these clots can break off into the bloodstream to allow the blood to spill out of the blood vessels, resulting in disseminated intravascular coagulation. Since this complication is disseminated throughout the body, the damaging enzymes and cytokines associated with different immune molecules may also cause damage to the blood vessels in the lungs. Damage and rupture in all of the blood vessels in the lungs seriously affects oxygen absorption into the

bloodstream, resulting in acute respiratory distress syndrome. This can lead to severe respiratory distress since the respiratory system can no longer pull in oxygen into the bloodstream from the environment. In response, the human body initially pushes to increase the cardiac output to compensate for the decreased systemic vascular resistance so as to maintain blood pressure. However, if remained untreated, the septic shock will persist and the cardiac output will eventually start to be depressed, resulting in a serious decrease in cardiac output [43-46]. These pathophysiological incidents caused by sepsis are reflected in several physiological parameters as clinical clues, hence commonly named as symptom distributives. Although highly elusive in nature, the entire purpose of the sepsis-3 and sepsis-2 definitions is to capture the underlying symptom distributives that are the most relevant.

Bedside Monitoring: qSOFA vs SIRS

Sepsis, unlike most other human diseases, is not a specific disease entity but rather a syndrome consorted with an ambiguous pathobiology and the absence of gold-standard diagnostic tests for assessments [1,21]. Therefore, numerous endeavors have been made to capture the pathobiology, pathophysiology, and epidemiology of sepsis to explain the syndrome. An initial definition of sepsis (sepsis-1) was introduced at the 1991 Consensus Conference that described sepsis as SIRS [21,40]. Addressing the limitations of sepsis-1, the 2001 task force extended the list of diagnostic criteria for sepsis (sepsis-2), based on SIRS, with the following four criteria: fever or hypothermia (body temperature>100.4°F or <96.8°F), tachypnea (respiratory rate >20 breaths/minute), tachycardia (heart rate >90 beats/minute), and white blood cell count >12,000/mm^3 or <4000/mm^3 (or >10% immature bands) [47]. In particular, sepsis-2 interprets sepsis as a cascaded disease that is primarily diagnosed as SIRS, followed by sepsis, severe sepsis, and septic shock. At the very end of the spectrum, patients may experience multiple organ dysfunction syndrome, an incurable stage of sepsis. Table 1 lists the parameters and cascaded development of sepsis as per the SIRS criteria. However, this definition failed to distinguish sepsis from the other uncomplicated infections and diseases that exhibit identical criteria, and indispensably failed to define what sepsis really is [1]. The task force also coined definitions for *severe sepsis* and *septic shock*, interpreting *severe sepsis* as sepsis complicated by organ dysfunction and *septic shock* as sepsis-induced hypotension persisting despite sufficient fluid resuscitation [47].

Table 1. Systemic inflammatory response syndrome (SIRS) criteria for sepsis definition.

Parameters/Criteria	Phases of syndrome development
Criterion 1: Body Temperature	
>100.4°F or <96.8°F	**Phase 1**: SIRS ≥ 2 criteria
Criterion 2: Respiratory Rate	
>20 breaths/minute (or $PaCO_2$ <32 mmHg)	**Phase 2**: Sepsis (SIRS + suspected or confirmed infection)
Criterion 3: Heart Rate	
> 90 beats/minute	**Phase 3**: Severe sepsis (sepsis + organ dysfunction)
Criterion 4: White blood cell count	
>12,000/mm^3 or <4000/mm (or >10% bands)	**Phase 4**: Septic shock (severe sepsis + persistent hypotension)
Final Phase: Multiple Organ Dysfunction	
Reported ≥ 2 organs failing	

With significant advancements in the understanding of sepsis pathophysiology and pathobiology, after nearly two decades, a new definition of sepsis was proposed at the Third International Consensus in 2016 [1]. Currently, sepsis (sepsis-3) is defined as a syndrome pertaining to a life-threatening organ dysfunction introduced by a dysregulated host response to a microorganism. According to the definitions of sepsis-3, the SOFA score (criteria) is used in the ICU to determine the extent of a patient's organ functions (dysfunction) [1]. In addition, sepsis can be promptly identified for an individual with a suspected infection at bedside using the qSOFA (sepsis-3) score. qSOFA requires satisfying at least two of the following criteria to determine that a patient is likely to have poor outcome due to sepsis [21]: respiratory rate ≥22 breaths/minutes, altered mental status (≤13 on the Glasgow Coma scale), and low blood pressure (≤100 mm Hg).

With the goal of leveraging the greater consistency of sepsis-3 in clinical trials and epidemiologic studies, several predictive machine-learning models were developed using the qSOFA parameters. Khwannimit et al [48] found that the qSOFA score showed higher prognostic accuracy for mortality and organ failure compared with SIRS criteria. Moreover, in predicting mortality and ICU-free days, qSOFA rendered considerably better discrimination in comparison with SIRS [49]. Donnele et al [50] and Hwang et al [51] provided substantial evidence to support employing SOFA and qSOFA in the ICU sepsis diagnosis and treatment workflow over SIRS criteria. However, numerous studies implied conflicting results, and asserted that qSOFA manifests inconsistent performance in mortality prediction [21]. Several studies reported that qSOFA showed poor sensitivity and inconsistent precision in the predictive models [49,51,52]. Although counterintuitive to some extent, Haydar et al [49] and Fernando et al [52] indicated that qSOFA took much longer in the patients' trajectory in comparison with SIRS to identify patients with sepsis, which further delayed the initiation of medical interventions in the ICU, and thereby

subjected the patients to a higher risk of developing septic shock and multiple organ dysfunction.

Considering these stark contrasts in the results (reflected by evaluation metrics such as accuracy, sensitivity, precision, and G-mean) of predictive modeling using SIRS and qSOFA parameters, in this study, we decided to take a step back and have a more in-depth look at the qSOFA and SIRS parameters, and their underlying attributes and interrelations. Multicollinearity among parameters often intensifies the tension between optimization and generalizability, and eventually leads to model overfitting, which in turn hampers the generalizability of discriminant functions [53]. Moreover, model overfitting indicates that a small deviation in the input data can result in considerable, and sometimes aberrant, changes in the model, even leading to changes in the sign of parameter estimates [21,53]. Table 2 compares the SIRS and qSOFA criteria, highlighting the changes brought in with sepsis-3 from sepsis-2 throughout all of the cascaded steps.

Table 2. Comparison of sepsis-2 and sepsis-3 criteria.

Stage	Sepsis-2 criteria (SIRS[a])	Sepsis-3 criteria (qSOFA[b])
Sepsis	Suspected or confirmed infection + SIRS	Suspected or confirmed infection + qSOFA score ≥2
Severe sepsis	Sepsis + organ dysfunction (lab markers, including hypoxia, hypotension, elevated lactate)	Category removed
Septic shock	Severe sepsis + persistent hypotension (after adequate fluid resuscitation)	Sepsis + vasopressors to maintain mean arterial pressure ≥65 mmHg + serum lactate level >2 mmol/L

[a]SIRS: systemic inflammatory response syndrome.

[b]qSOFA: quick Sequential Organ Failure Assessment.

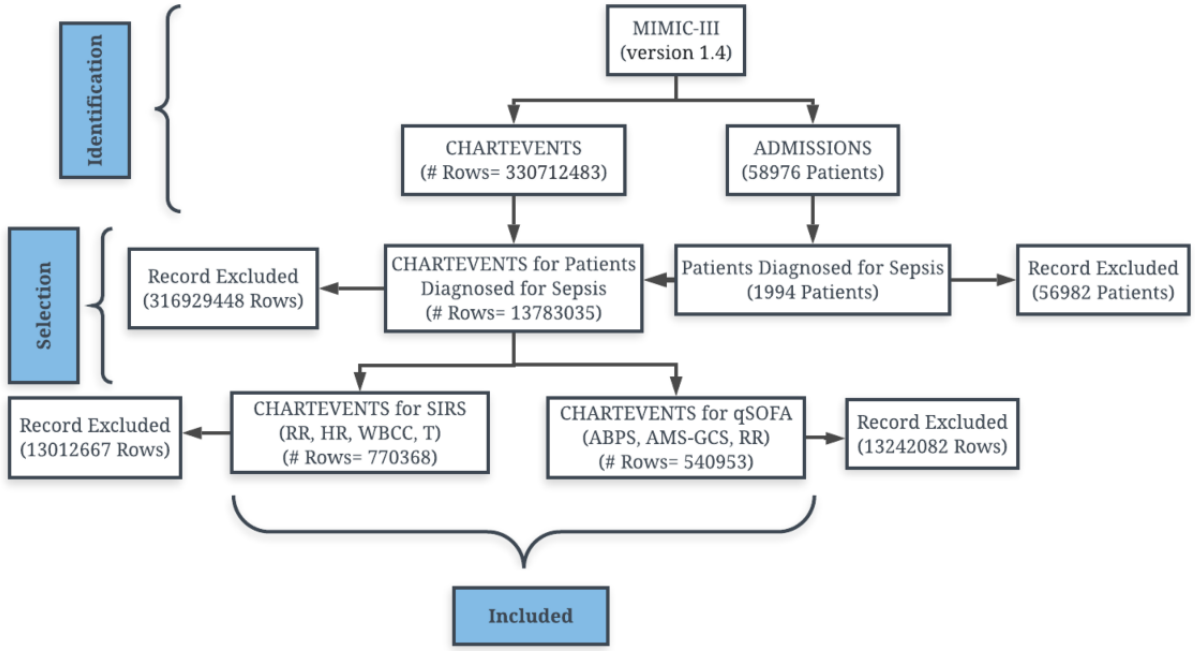
Data and Research Design

We used MIMIC-III (Version 1.4), a publicly available ICU patient database [1], for this study. The data, ranging from 2001 to 2012, involves 58,976 distinct hospital admissions. For the purpose of our study, we used the parameters of the qSOFA as well as SIRS to identify all ICU patients who had been diagnosed with sepsis or were most susceptible to the disease. We then analyzed the qSOFA and SIRS parameters of these identified sepsis patients, or the patients who had undergone sepsis screening, to study their intrarelationship. In our population, 1994 hospital admissions resulted in a diagnosis of sepsis among 58,976 overall admissions from 2001 to 2012. Among these 1994 patients, the mortality rate was 21.11% (n=421 deaths).

The selection criteria included identifying the unique key for the critical parameter records and omittable parameters that we deemed to be bias-free for the purpose of this study, such as patient gender, data storage time, and deidentified date of birth in the case of sepsis. During research design and data wrangling, we confronted missing data and outlier values that were not biologically reasonable, albeit not for a considerable amount of records. This modicum amount of unexpected data points opened up the possibility of two distinct research designs. First, we could ignore the observations that have such data point(s) because they are of negligible number compared to the total observations available. Second, we could follow the conventional central-value imputation or multiple imputations by chained equations to handle the missing data. A multiblind Delphi process, convened by Ubicomp Lab of the Department of Computer Science at Marquette University and Regenstrief Center for Healthcare Engineering at Purdue University, came to the decision that ignoring the observations that have such unexpected data point(s) will be more suitable for the purpose of this study, which requires avoiding imputation bias. Moreover, outlier values that are not biologically reasonable were excluded, considering them as mistaken data entries in the ICU [21].

To determine the prevalence and dichotomy of the qSOFA and SIRS parameters, we identified 13,783,035 patient records (Chartevent) from 330,712,483 records (Chartevent) available in MIMIC-III (Version 1.4), which are unique for each Hospital Admission ID and chart time and pertaining to patients who had received a sepsis diagnosis. Then, to identify the most prevalent qSOFA and SIRS criteria, we selected 540,953 and 770,368 patient records for SIRS and qSOFA, respectively (in which respiratory rate was common in both cases). Figure 1 summarizes the research design in a simple flow chart.

Figure 1. Outline of research design.



To assure the consistency and interpretability of the results while determining the most prevalent sepsis scenario, our selection criteria only filtered within chart times for which we had observations for all three qSOFA parameters since the observation frequency varies with the parameters based on the intricacy involved in measurement. For instance, observations for altered mental status (based on the Glasgow Coma Scale) are less frequently recorded than those of the respiratory rate. More importantly, since sepsis is a spectrum disease, studying and comparing the observations for different parameters at different record times for a particular patient can confound the result and its interpretability. For the same reason, studying the parameters that are observed at the same time can capture the patient's disease trajectory more consistently. For determining the most prevalent sepsis scenario for SIRS, our selection criteria only filtered within chart times for which we had observations for all four parameters (temperature, heart rate, respiratory rate, and white blood cell count). The white blood cell count observations are considerably less frequent compared to the other three parameters of SIRS, and therefore observations considered for the SIRS criteria are substantially reduced compared with those considered for the qSOFA criteria.

We further addressed two possible sources of selection bias. First, it is intuitive that the longer the patient stays in the ICU, there will be more observations available for that particular patient. We considered that this may influence the results of our study to some extent if there are considerably more patients with a longer length of stay. Second, when evaluating the respiratory rate for ICU patients, there may be a possible blend in the data between patients with intubated breathing and natural breathing. However, the possibility of these two selection biases also provided an opportunity to test the intrageneralizability of the results of this study (both for qSOFA and SIRS). Therefore, in the second phase of this study, we dissected our data for only the first observations of each hospital admission.

This research design is grounded in statistical theory such that the results can help in developing multiparameter intelligent sepsis prediction or treatment models that require predictors exhibiting the least or no collinearity.

Results

Statistical Distributions: qSOFA and SIRS

The means (SD) and median (IQR) values for qSOFA and SIRS parameters in each phase of the study are presented in Table 3. In the first phase of the study, with respect to the qSOFA criteria, we analyzed the distributions of systolic arterial blood pressure, Glasgow Coma Scale score, and respiratory rate. For the SIRS criteria, in the first phase we analyzed the distribution of heart rate, respiratory rate, temperature, and white blood cell count. In the second phase, we only considered the first observation of each hospital admission for each parameter.

Table 3. Statistical distributions of parameters for quick Sequential Organ Failure Assessment (qSOFA) and systemic inflammatory response syndrome (SIRS).

Parameter	Phase 1: Entire patient trajectory		Phase 2: First observation only	
	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)
qSOFA				
SABP[a] (mmHg)	116.4 (24.78)	114.0 (100-131)	106.7 (37.62)	110.0 (96.0-126.0)
GCS[b]	11.17 (3.66)	11.00 (9-15)	11.53 (4.32)	14.00 (8-15)
RR[c] (breaths/min)	21.07 (6.52)	21.00 (17-25)	20.48 (6.16)	20.00 (16.00-24.00)
SIRS				
HR[d] (beats/minute)	89.1 (18.61)	87 (76-100)	95.58 (20.76)	94.00 (80.00-109.00)
RR (breaths/minute)	21.07 (6.52)	21.00 (17-25)	20.48 (6.16)	20.00 (16.00-24.00)
BT[e] (°F)	98.37 (1.57)	98.30 (97.30-99.30)	98.25 (2.01)	98.20 (97.00-99.50)
WBC[f] count (/mm^3)	13.14 (7.30)	11.70 (8.10-16.70)	14.34 (8.28)	12.80 (8.50-18.90)

[a]SABP: systolic arterial blood pressure.

[b]GCS: Glasgow Coma Scale.

[c]RR: respiratory rate.

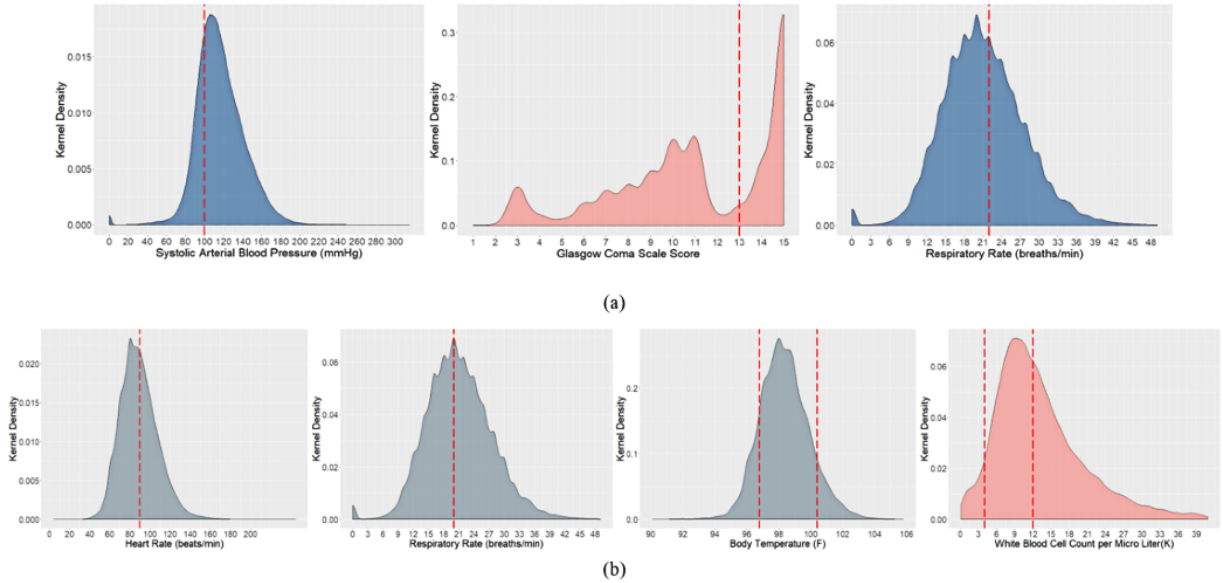
[d]HR: heart rate.

[e]BT: body temperature.

[f]WBC: white blood cell.

Kernel density estimation distributions for the qSOFA criteria (systolic arterial blood pressure, altered mental status in Glasgow Coma Scale, and respiratory rate) and SIRS criteria (heart rate, respiratory rate, temperature, and white blood cell count) are depicted in Figure 2 to investigate the most prevalent sepsis parameter. Visual statistics demonstrated that most of the patients' observations did not meet the qSOFA criterion for systolic arterial blood pressure (Figure 2a). The distribution for systolic arterial blood pressure implies that most of the observations were in the range of 100-125 mmHg, which is in the healthy range from the clinical point of view. Similarly, the Glasgow Coma Scale distribution (Figure 2a) indicated that a significant portion of these observations were in the safe zone (15 and 14). However, as the Glasgow Coma Scale ranges from 1 to 15, and the domain of consideration for the not-safe zone (qSOFA, 1-13) and the domain of consideration for the safe zone (14-15) are significantly disproportionate, the visual analytics may be confusing for an accurate interpretation. In the case of respiratory rate (Figure 2a), it is critical to interpret whether or not the majority of the observations met the qSOFA criterion, although it is evident that most of the data ranged between 15 and 24 breaths/minute. From the clinical point of view, at a resting state, a respiratory rate observation of 12-20 breaths/minute is considered to be healthy.

Figure 2. Kernel density estimation distribution of (a) quick Sequential Organ Failure Assessment (qSOFA) and (b) systemic inflammatory response syndrome (SIRS) parameters to understand the prevalence of each parameter.



For the SIRS criteria (Figure 2b), the distribution for heart rate observations was less confounding using visual analytics in inferring prevalence, as more of the kernel density was below the criterion margin (90 beats/minute), which indicates the presence of more healthy observations. In the case of respiratory rate measurement, it is worth mentioning that the cutoff for the SIRS criteria is different than that of the qSOFA criteria. For SIRS criteria, the criterion cutoff is 20 breaths/minute, and anything above that level is considered as tachypnea. It is visually discernible that as the cutoff shifted left (from 22 to 20) for SIRS, more patient observations met the sepsis criteria. The distribution for body temperature can be interpreted as a band: the observations inside two temperature cutoffs indicate the density of the healthy observations, and they represented a significant portion of the distribution. In the case of white blood cell count, as the domain of consideration for the not-safe zone and the domain of consideration for the safe zone were significantly disproportionate, the visual analytics may be confusing to imply prevalence. However, we can infer that the majority of observations met the SIRS criteria.

In the following subsections, we provide an explicit numerical interpretation to better understand the prevalence and underlying statistical relation between the predictors.

Patients' Entire Trajectory for qSOFA

The kernel density estimation distribution of qSOFA parameters for both safe and qSOFA criterion–met observations are presented in Figure 3 to better understand the prevalent qSOFA parameters. Overall, 25.12% of the systolic arterial blood pressure observations, 59.28% of the Glasgow Coma Scale measurements, and 45.11% of the respiratory rate observations met the respective qSOFA criterion. It is intuitive from the qSOFA criteria that determination of the most prevalent criterion from observational studies would help practitioners and researchers in further factorial experiments. This observational study entirely relied on passive retrospective observations without assigning any further treatment. The results suggest that altered mental status is the most prevalent qSOFA criterion experienced in the ICU. We further addressed a nearly double-barreled question: what is the most prevalent sepsis scenario in the ICU? We found that 28.19% of the observations (when three measurements were available at the same time) showed a two-factored qSOFA of high respiratory rate and altered mental status (among $3C_3+3C_2=4$ possibilities), resulting in this pair identified as the most prevalent qSOFA (sepsis-3) scenario in the ICU. Notably, no sepsis is another possible scenario besides these four possible qSOFA scenarios in the ICU (which is also true for our observations).

Figure 3. Kernel density estimation distribution of quick Sequential Organ Failure Assessment (qSOFA) parameters for both safe and qSOFA criterion-met observations to identify the prevalent qSOFA parameters.

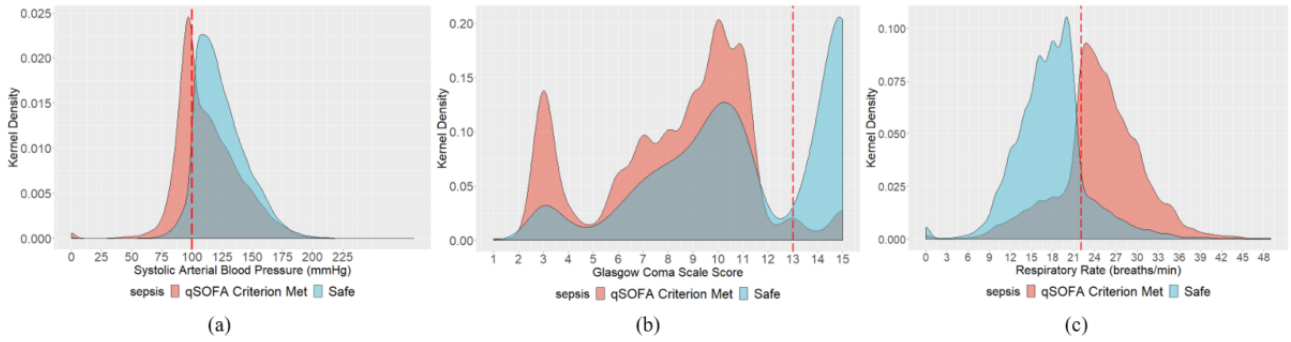
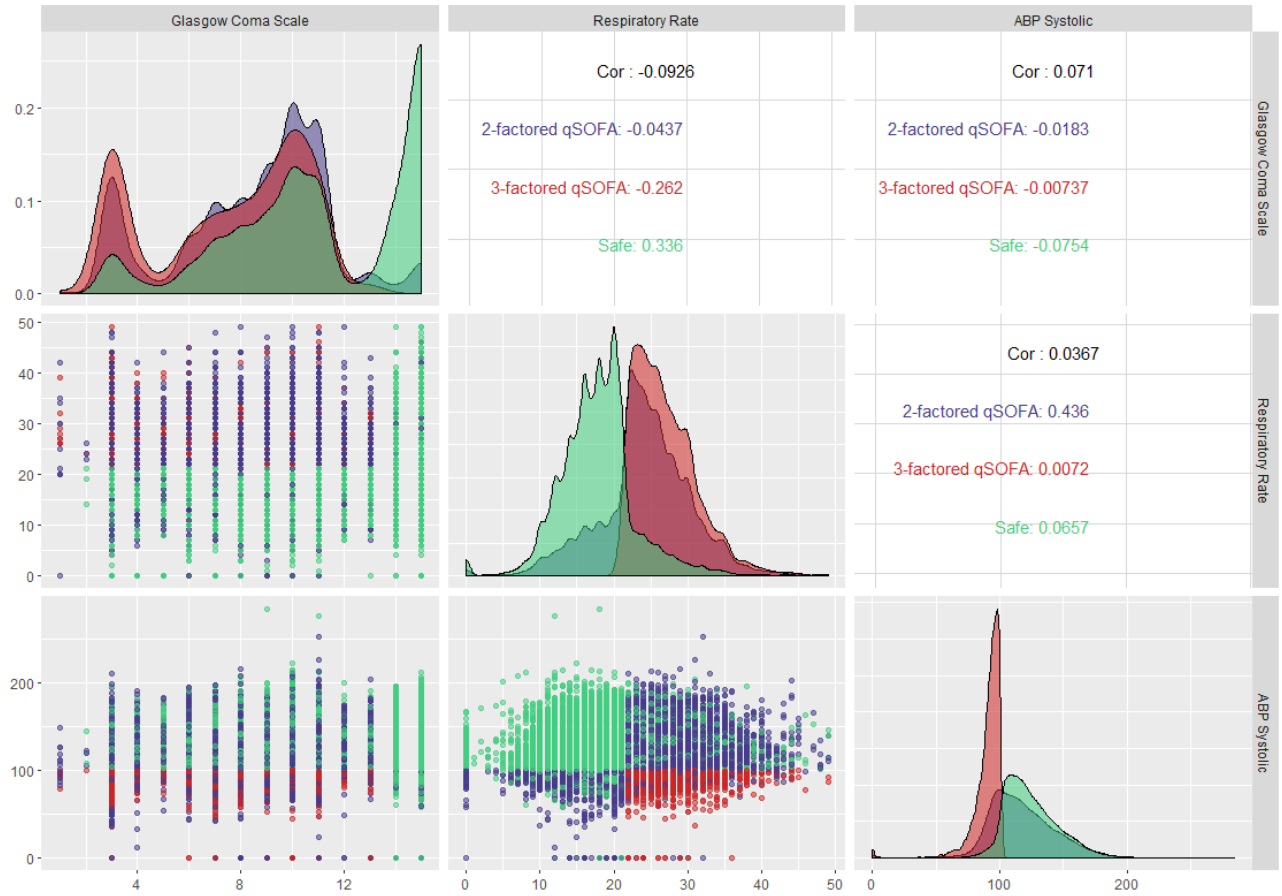


Figure 4 shows a facet grid plot of the qSOFA parameters to capture the most prevalent sepsis scenario and the underlying dichotomy among the parameters. This plot has multiple implications; however, the most obvious is the comparison of the Pearson correlation coefficients (absolute) of each of the qSOFA parameters' pairs. The absolute Pearson correlation coefficients for respiratory rate-Glasgow Coma Scale measurement, Glasgow Coma Scale measurement-systolic arterial blood pressure, and respiratory rate-systolic arterial blood pressure pairs were 0.09, 0.07, and 0.04, respectively. These insignificant correlation coefficients nullify the possibility of any linear correlation among the qSOFA parameters, thereby ensuring that multicollinearity does not exist between the parameters and further advocates for the dichotomy among them. Understanding this relationship can help in developing predictive models, as it implies that the overdetermined system involved in the modeling is a full-ranked matrix (ie, not rank-deficient). However, the lack of multicollinearity cannot guarantee that two random variables are statistically independent. Moreover, based on its pathophysiology, sepsis is a spectrum disease, and therefore one predictor may influence another during the development of sepsis and septic shock.

Figure 4. Facet grid illustration of sepsis-3 (qSOFA) parameters to capture the underlying relationship between parameters and the most prevalent sepsis scenario in the intensive care unit. qSOFA: quick Sequential Organ Failure Assessment.



Patients' Entire Trajectory for SIRS

Figure 5 shows the kernel density estimation distribution of SIRS parameters for both safe and SIRS criterion–met observations to understand the prevalent SIRS parameters. We found that 43.30% of the heart rate observations, 50.89% of the respiratory rate observations, 23.08% of the body temperature observations, and 53.12% of the white blood cell count observations met the respective SIRS criterion. Although both the white blood cell count and respiratory rate had a significant prevalence in the observations of patients who went through the sepsis screening, white blood cell count was the most prevalent SIRS criterion experienced in the ICU. In addition, 12.32% of the observations (when four measurements were available at the same time) showed a three-factored SIRS of tachypnea-high heart rate-high white blood cell count. It is critical to consider that there are 6 possible pairs of combinations, 4 possible trios of combinations, and 1 combination considering all the parameters as the possible sepsis scenario in the ICU. As mentioned above for qSOFA, no sepsis is another possible scenario besides these 11 possible SIRS scenarios in the ICU (which is also the case for our observations). Identifying the most prevalent criterion and sepsis scenario in the ICU for SIRS can help practitioners and researchers in the diagnosis, treatment, and design of further factorial experiments.

Figure 5. Kernel density estimation distribution of systemic inflammatory response syndrome (SIRS) parameters for both safe and sepsis criterion–met observations to identify the prevalent SIRS parameters.

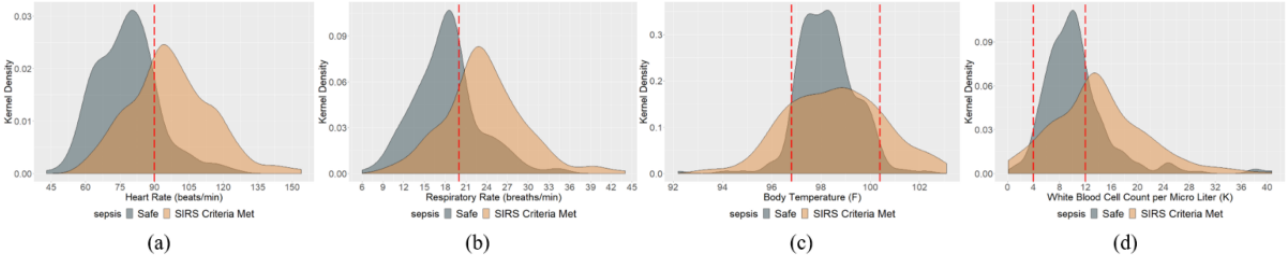
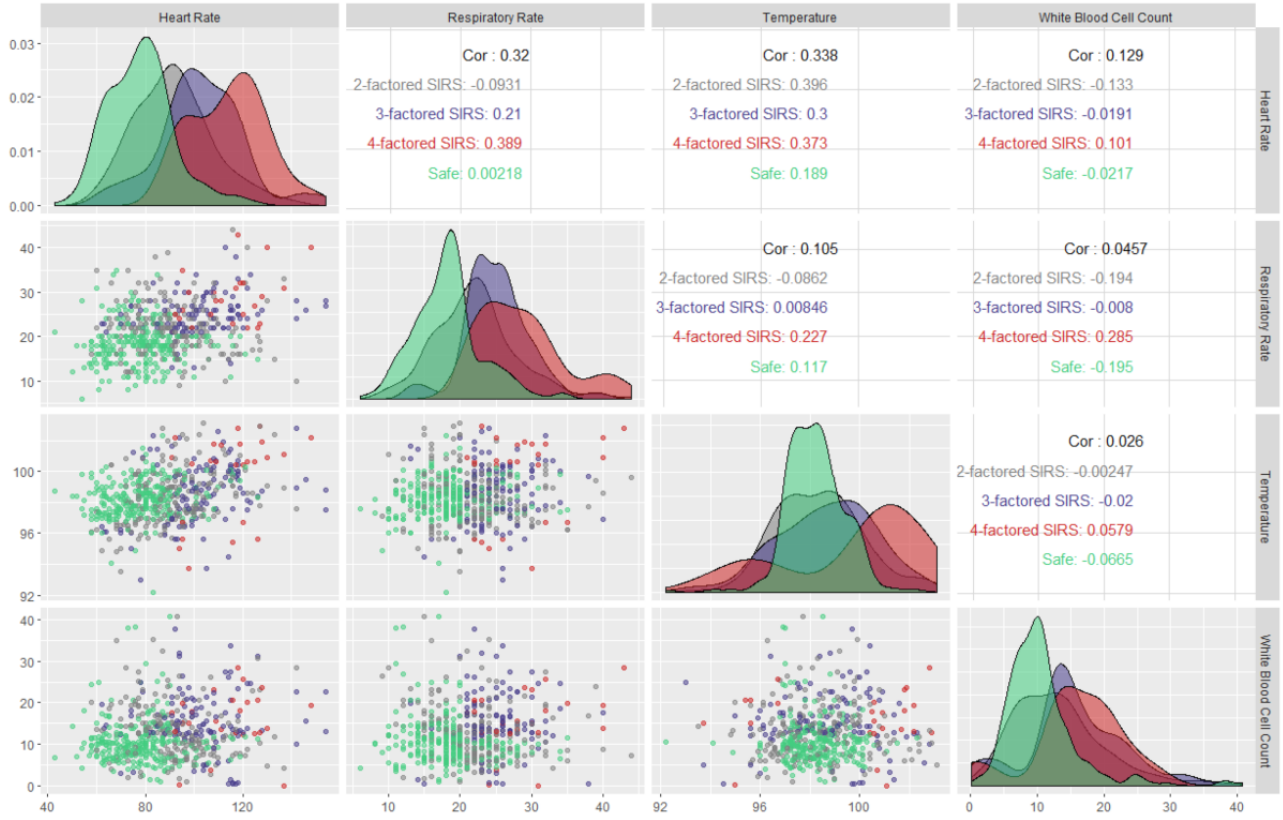


Figure 6 shows a facet grid plot of SIRS (sepsis-2) parameters to capture the most prevalent SIRS scenario and the underlying dichotomy among the parameters. The absolute Pearson correlation coefficients for heart rate-respiratory rate, heart rate-temperature, heart rate-white blood cell count, respiratory rate-temperature, respiratory rate-white blood cell count, and temperature-white blood cell count were 0.32, 0.34, 0.13, 0.11, 0.05, and 0.03, respectively. These insignificant absolute

correlation coefficients invalidate the possibility of any correlation among the critical parameters, thereby ensuring that multicollinearity does not exist between the parameters and further advocates for the dichotomy among them. However, despite being not statistically significant, the absolute correlation coefficients were not negligible in the case of heart rate-respiratory rate and heart rate-temperature pairs. Understanding this relationship can help in developing predictive models as it implies that the overdetermined system involved in the modeling is a full-ranked matrix (ie, not rank-deficient). However, the lack of multicollinearity cannot guarantee that two random variables are statistically independent.

Figure 6. Facet grid illustration of sepsis-2 (SIRS) parameters to capture the underlying relationship between parameters and the most prevalent sepsis scenario in the intensive care unit. SIRS: systemic inflammatory response syndrome.



Patients' First Observation Only for qSOFA

In the second phase of this study, we dissected data for only the first observations of each hospital admission. This may address two possible selection biases, including the opportunity to test the intrageneralizability of the result of this observational study. First, it is intuitive that the longer the patient stays in the ICU, there will be more observations available for that particular patient. This may influence the results of our study to some extent if there is considerable disproportion between the length of stay among patients. Second, when evaluating the respiratory rate for ICU patients, there may be a possible blend in the data between patients under intubated breathing and those naturally breathing. The kernel density estimation distribution of qSOFA parameters for both safe and qSOFA criterion–met observations are presented in Figure 7 to understand the prevalent qSOFA parameters. We found that 32.58% of the systolic arterial blood pressure observations, 44.54% of the Glasgow Coma Scale measurements, and 40.53% of the respiratory rate observations met the respective qSOFA criterion. This observational study entirely relied on passive retrospective observation without assigning any further treatment. The results suggest that altered mental status is the most prevalent qSOFA criterion experienced in the ICU. In addition, 18.25% of the observations had a two-factored qSOFA of high respiratory rate and altered mental status (among $3C_3+3C_2=4$ possibilities), resulting in this pair as the most prevalent qSOFA (sepsis-3) scenario in the ICU, although the no-sepsis scenario is also possible.

Figure 7. Kernel density estimation distribution of quick Sequential Organ Failure Assessment (qSOFA) parameters for both safe and qSOFA criterion–met patients at first observations to identify the prevalent qSOFA parameters.

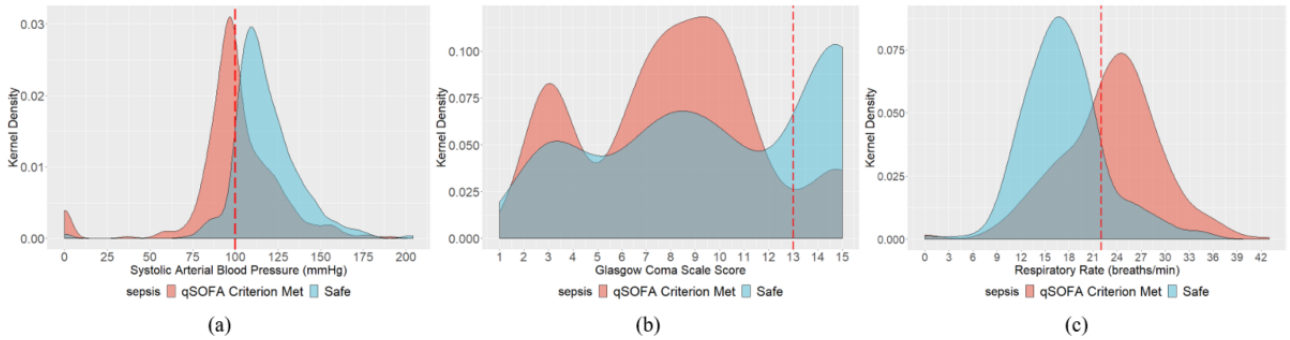
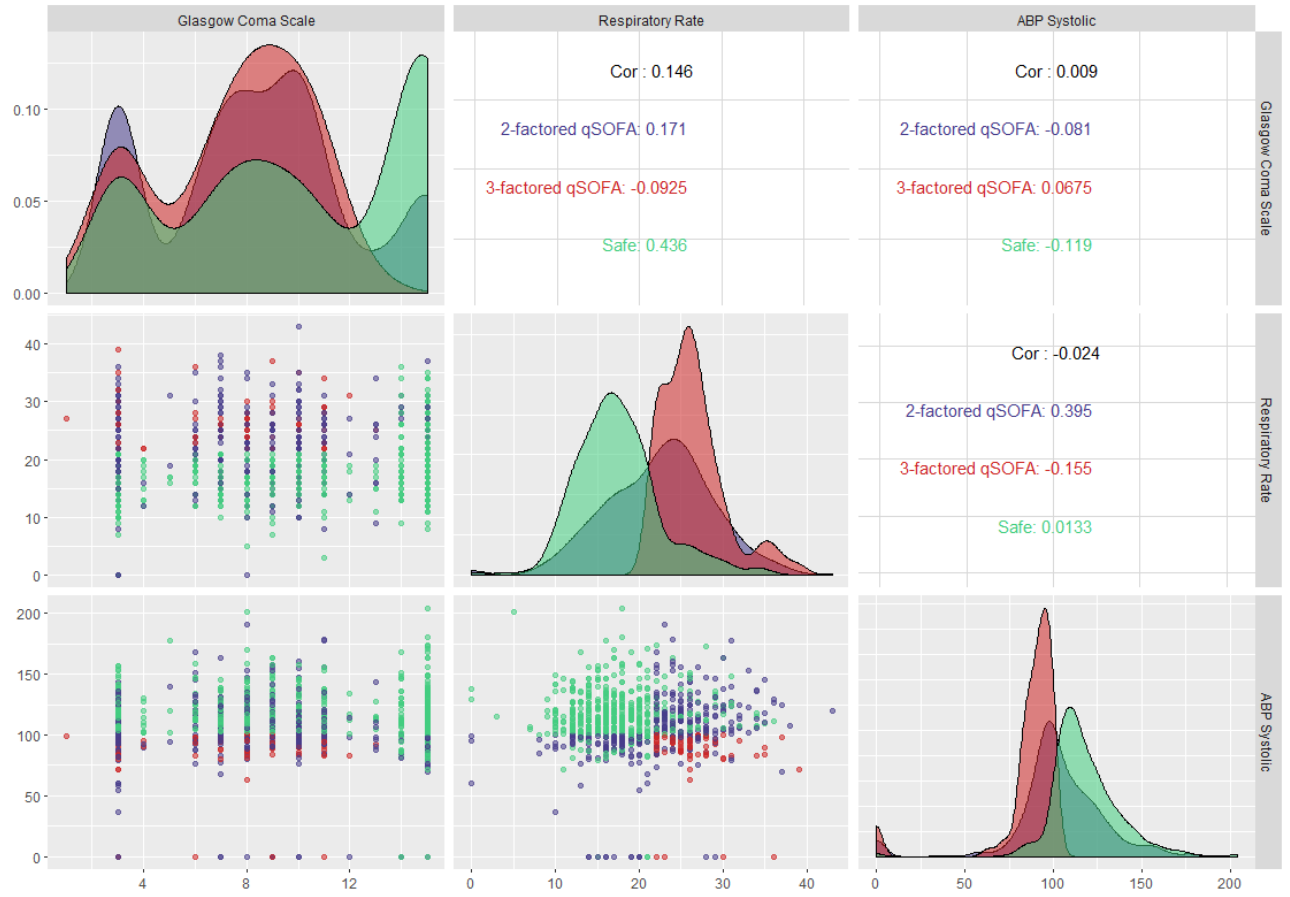


Figure 8 shows the facet grid on qSOFA parameters to understand the most prevalent qSOFA scenario and the underlying dichotomy among the parameters. The absolute Pearson correlation coefficients for respiratory rate-Glasgow Coma Scale measurement, Glasgow Coma Scale measurement-systolic arterial blood pressure, and respiratory rate-systolic arterial blood pressure pairs were 0.15, 0.01, and 0.02, respectively. These insignificant correlation coefficients invalidate the possibility of any correlation among the critical parameters, ensuring that multicollinearity does not exist between the parameters and further bolsters the dichotomy among them. However, the lack of multicollinearity cannot guarantee that two random variables are statistically independent.

Figure 8. Facet grid illustration of sepsis-3 (qSOFA) parameters to capture the underlying relationship between parameters and the most prevalent sepsis scenario of patients at first observations in the intensive care unit. qSOFA: quick Sequential Organ Failure Assessment.



Patients' First Observation Only for SIRS

Figure 9 shows the kernel density estimation distribution of SIRS parameters for both safe and SIRS criterion–met observations using only the first observations. We found that 57.03% of the heart rate observations, 45.89% of the respiratory rate observations, 33.93% of the body temperature observations, and 60.57% of the white blood cell count observations met the respective SIRS criterion. These results suggest that white blood cell count is the most prevalent criterion experienced in the ICU, albeit considering that both the white blood cell count and respiratory rate had significant prevalence. In addition, 11.38% of the SIRS criteria–met sepsis patients showed a three-factored SIRS of tachypnea-high heart rate-high white blood cell count

(among $4C_4+4C_3+4C_2=11$ possibilities), resulting in this trio as the most prevalent sepsis (SIRS) scenario in the ICU. It is important to consider that there are 6 possible pairs of combinations, 4 possible trios of combinations, and 1 combination considering all of the parameters as the possible sepsis scenarios in the ICU, and that no sepsis is another possible scenario. Determining the most prevalent SIRS criterion and sepsis scenario at the first observation upon hospitalization can help practitioners and researchers in diagnosis, treatment, and further factorial experiments.

Figure 9. Kernel density estimation distribution of systemic inflammatory response syndrome (SIRS) parameters for both safe and sepsis criterion–met patients at first observations to identify the prevalent SIRS parameters.

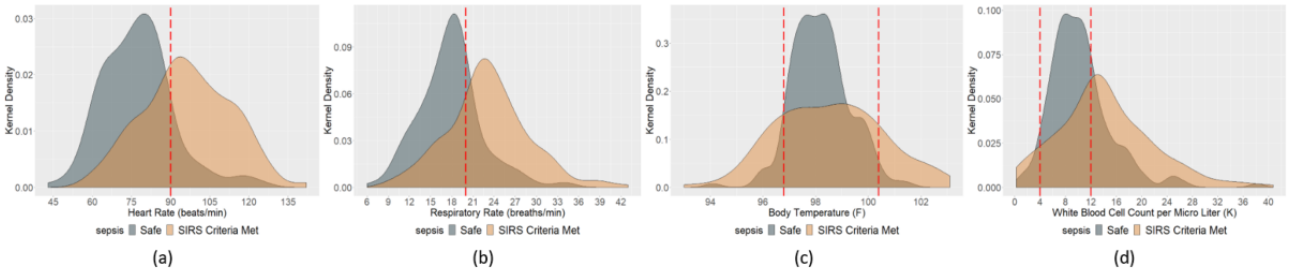
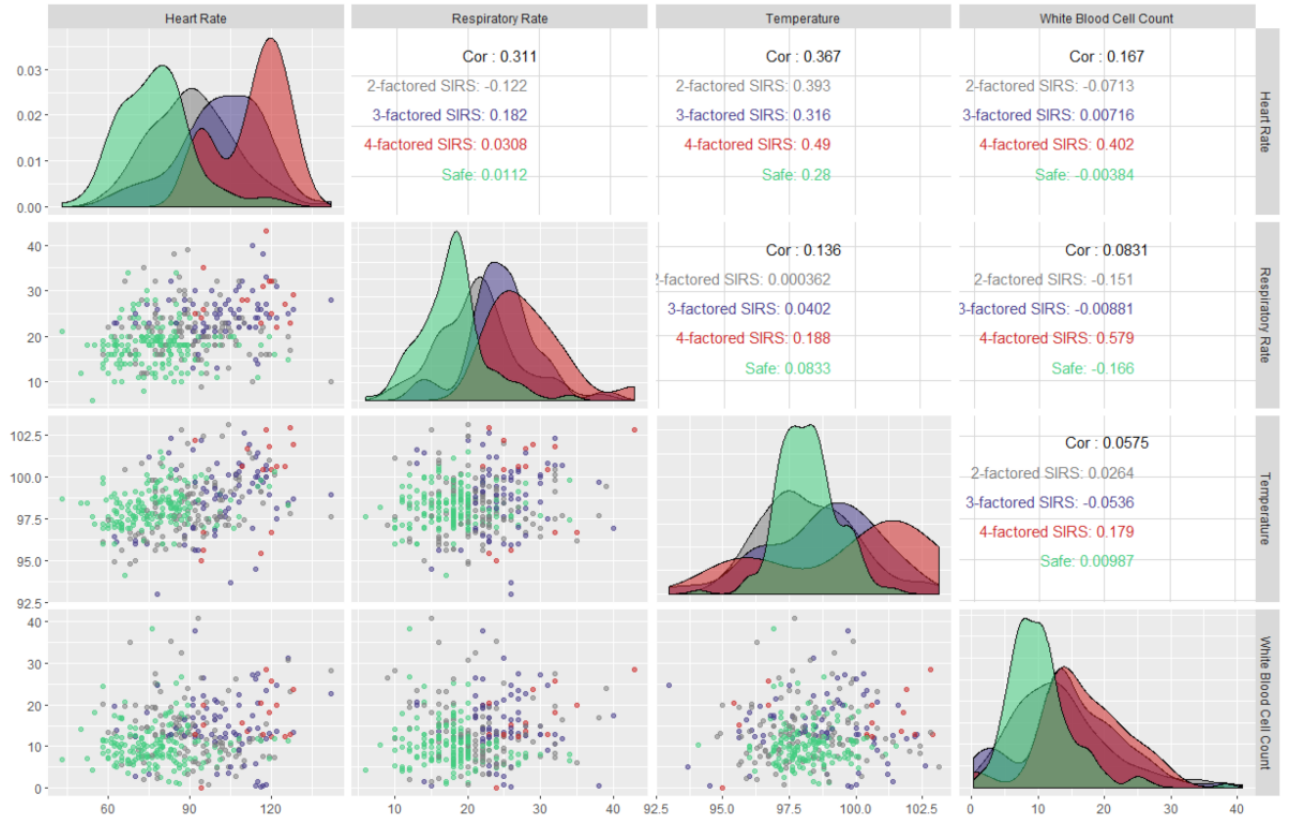


Figure 10 shows the facet grid illustration for SIRS parameters at the first observation. The insignificant absolute Pearson correlation coefficients invalidate the possibility of any correlation among the critical parameters, thereby ensuring that multicollinearity does not exist between the parameters and further bolsters the dichotomy among them. However, similar to the case for all observations, the absolute correlation coefficients were not negligible in the case of heart rate-respiratory rate and heart rate-temperature pairs.

Figure 10. Facet grid illustration of sepsis-2 (SIRS) parameters to capture the underlying relationship between parameters and the most prevalent sepsis scenario of patients at first observations in the intensive care unit. SIRS: systemic inflammatory response syndrome.



Discussion

Theoretical Reasoning

This study reveals that altered mental status and systolic arterial blood pressure are the most and least prevalent qSOFA criteria, respectively, observed in the ICU. Mathematically, blood pressure is the product of systemic vascular resistance and cardiac output. Hence, with the decrease in systemic vascular resistance due to vasodilation, blood pressure will drop down if the cardiac output remains the same. However, in practice, when the systemic vascular resistance drops down, the human body immediately tries to maintain the equilibrium for a few moments and compensates with the cardiac output. Cardiac output depends on the respiratory rate in a nonlinear and proportionate manner; hence, the increase in the respiratory rate

increases the cardiac output and maintains the equilibrium of the blood pressure initially. However, over time, that equilibrium breaks down, although the cardiac output (and consequently respiratory rate) continually tries to reach a stable state. This fact advocates the possibility of respiratory rate to be a more prevalent criterion compared to systolic arterial blood pressure as a symptom. From the aspect of SIRS criteria, the reason for the white blood cell count to emerge as the most prevalent criterion is intuitive. When a microorganism invades, the body's immune response is triggered and white blood cells appear immediately. Heart rate, respiratory rate, and temperature are consequential symptoms associated with an increase in white blood cells and the immune response. As sepsis is a spectrum disease, one predictor may influence another during disease development and progression to septic shock, although they are not linearly correlated. The findings of this observational study support the established pathophysiology of sepsis described in the literature.

Research Opportunities

Although MIMIC-III is an extensive critical care database, it is a single-center database comprising critical care unit electronic health record data of Beth Israel Deaconess Medical Center in Boston. Regardless of the myriad amount of patient data, the findings that are valid for the Beth Israel Deaconess Medical Center in Boston may not be useful for other medical centers and critical care units. The epidemiology and treatment facilities vary among the hospitals, states, and infrastructures of countries. Epidemiology and treatment facilities have a significant impact on patient outcome, as well as on patients' symptom distributives. On the flip side, this observational study entirely relied on passive retrospective observation, and the dynamics of the treatment and medicine advance with time and research. In addition, the prevalence of the physiological parameters, along with time and resource variability, may also affect the interrelation nature among parameters. The results may also vary if considering the analysis from an individual aspect. Although a collective analysis infers the dichotomy among parameters, there may be a possibility that data from even one patient show strong multicollinearity. Again, the parameters measured may vary according to the therapeutics undertaken in the ICU. For instance, the Glasgow Coma Scale score may become low due to sedation, catecholamines may be responsible for healthy blood pressure, or mechanical ventilation may affect the respiratory rate. Any predictive modeling and treatment plan should take this variability and uncertainty into account.

This uncertainty around generalizability opens up new research opportunities in the health informatics domain in three possible directions: (1) Does this finding hold its generalizability while integrating data from multiple electronic health records? (2) How can we study confounding variables induced by numerous groups of people with different characteristics? (3) How can these findings address the confounding medical interventions in sepsis treatment?

Moreover, the comparison between qSOFA and SIRS can be extended to comparing SOFA and qSOFA, SIRS and SOFA, or all the three criteria available to better understand the underlying interrelations between the parameters.

Conclusion

This study indicates that altered mental status (as assessed with the Glasgow Coma Scale) is the most prevalent qSOFA criterion and white blood cell count is the most prevalent SIRS criterion for patients in the ICU. Besides, two-factored sepsis comprising altered mental status and high respiratory rate (≥22 breaths/minute) is the most prevalent sepsis-3 (qSOFA) scenario, and two-factored sepsis of white blood cells and tachypnea is the most prevalent sepsis-2 (SIRS) scenario confronted in the ICU among patients screened for sepsis. In addition, the Pearson correlation coefficients advocate for the dichotomy among the sepsis parameters (for both qSOFA and SIRS). This study implies that sepsis diagnosis and treatment should be pertinent to its type, and in this regard, these multifactored attributes should be taken into account. Machine-learning predictive models should consider the most prevalent criterion pair, which would allow for a faster diagnosis. Moreover, the reasoning backed by the sepsis pathophysiology assures the interpretability that these results require. These findings can help obtain a better understanding of the algorithmic, as well as contextual challenges that influence predictive decisions in the ICU.

Acknowledgments

Partial support for this study was provided by grants from the Regenstrief Center for Healthcare Engineering at Purdue University, Northwestern Mutual Data Science Institute in Milwaukee, RB Annis School of Engineering at University of Indianapolis, and Ubicomp Lab at Marquette University.

Conflicts of Interest

The authors affirm that there are no known personal relationships and competing financial interests that could influence the scientific research reported in this article.

References

1. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016 Feb 23;315(8):801-810 [FREE Full text] [doi: 10.1001/jama.2016.0287] [Medline: 26903338]

2. Angus DC, Linde-Zwirble WT, Lidicker J, Clermont G, Carcillo J, Pinsky MR. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. Crit Care Med 2001 Jul;29(7):1303-1310. [doi: 10.1097/00003246-200107000-00002] [Medline: 11445675]
3. Liu V, Escobar GJ, Greene JD, Soule J, Whippy A, Angus DC, et al. Hospital deaths in patients with sepsis from 2 independent cohorts. JAMA 2014 Jul 02;312(1):90-92. [doi: 10.1001/jama.2014.5804] [Medline: 24838355]
4. Rhee C, Dantes R, Epstein L, Murphy DJ, Seymour CW, Iwashyna TJ, CDC Prevention Epicenter Program. Incidence and Trends of Sepsis in US Hospitals Using Clinical vs Claims Data, 2009-2014. JAMA 2017 Oct 03;318(13):1241-1249 [FREE Full text] [doi: 10.1001/jama.2017.13836] [Medline: 28903154]
5. Rhee C, Jones TM, Hamad Y, Pande A, Varon J, O'Brien C, Centers for Disease ControlPrevention (CDC) Prevention Epicenters Program. Prevalence, Underlying Causes, and Preventability of Sepsis-Associated Mortality in US Acute Care Hospitals. JAMA Netw Open 2019 Feb 01;2(2):e187571 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.7571] [Medline: 30768188]
6. van Wyk F, Khojandi A, Kamaleswaran R. Improving Prediction Performance Using Hierarchical Analysis of Real-Time Data: A Sepsis Case Study. IEEE J Biomed Health Inform 2019 May;23(3):978-986. [doi: 10.1109/JBHI.2019.2894570] [Medline: 30676988]
7. Blecker S, Pandya R, Stork S, Mann D, Kuperman G, Shelley D, et al. Interruptive Versus Noninterruptive Clinical Decision Support: Usability Study. JMIR Hum Factors 2019 Apr 17;6(2):e12469 [FREE Full text] [doi: 10.2196/12469] [Medline: 30994460]
8. Aakre CA, Kitson JE, Li M, Herasevich V. Iterative User Interface Design for Automated Sequential Organ Failure Assessment Score Calculator in Sepsis Detection. JMIR Hum Factors 2017 May 18;4(2):e14 [FREE Full text] [doi: 10.2196/humanfactors.7567] [Medline: 28526675]
9. Lehman LH, Adams RP, Mayaud L, Moody GB, Malhotra A, Mark RG, et al. A physiological time series dynamics-based approach to patient monitoring and outcome prediction. IEEE J Biomed Health Inform 2015 May;19(3):1068-1076 [FREE Full text] [doi: 10.1109/JBHI.2014.2330827] [Medline: 25014976]
10. Celi LA, Davidzon G, Johnson AE, Komorowski M, Marshall DC, Nair SS, et al. Bridging the Health Data Divide. J Med Internet Res 2016 Dec 20;18(12):e325 [FREE Full text] [doi: 10.2196/jmir.6400] [Medline: 27998877]
11. Hall M, Williams S, DeFrances C, Golosinskiy A. NCHS Data Brief No. 62: Inpatient care for septicemia or sepsis: a challenge for patients and hospitals. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics. 2011 Jun. URL: https://www.cdc.gov/nchs/data/databriefs/db62.pdf [accessed 2020-11-25]
12. Celi LA, Marshall JD, Lai Y, Stone DJ. Disrupting Electronic Health Records Systems: The Next Generation. JMIR Med Inform 2015 Oct 23;3(4):e34 [FREE Full text] [doi: 10.2196/medinform.4192] [Medline: 26500106]
13. Connell A, Raine R, Martin P, Barbosa E, Morris S, Nightingale C, et al. Implementation of a Digitally Enabled Care Pathway (Part 1): Impact on Clinical Outcomes and Associated Health Care Costs. J Med Internet Res 2019 Jul 15;21(7):e13147 [FREE Full text] [doi: 10.2196/13147] [Medline: 31368447]
14. Rockville M. Healthcare Cost and Utilization Project (HCUP) National Inpatient Sample (NIS). Agency for Healthcare Research and Quality. URL: https://www.hcup-us.ahrq.gov/news/exhibit_booth/nis_brochure.jsp [accessed 2020-11-25]
15. Celeste MT, Brian JM. National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2013. In: Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. Rockville, MD: Agency for Healthcare Research and Quality; May 2016.
16. Deisz R, Rademacher S, Gilger K, Jegen R, Sauerzapfe B, Fitzner C, et al. Additional Telemedicine Rounds as a Successful Performance-Improvement Strategy for Sepsis Management: Observational Multicenter Study. J Med Internet Res 2019 Jan 15;21(1):e11161 [FREE Full text] [doi: 10.2196/11161] [Medline: 30664476]
17. 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]
18. O'Brien J. The Cost of Sepsis. CDC Safe Healthcare Blog. 2015 Sep 9. URL: https://blogs.cdc.gov/safehealthcare/the-cost-of-sepsis/ [accessed 2020-11-25]
19. Fleischmann C, Scherag A, Adhikari NKJ, Hartog CS, Tsaganos T, Schlattmann P, International Forum of Acute Care Trialists. Assessment of Global Incidence and Mortality of Hospital-treated Sepsis. Current Estimates and Limitations. Am J Respir Crit Care Med 2016 Feb 01;193(3):259-272. [doi: 10.1164/rccm.201504-0781OC] [Medline: 26414292]
20. Russo C, Elixhauser A, Steiner C, Wier L. Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. Rockville, MD: Agency for Healthcare Research and Quality; 2006.
21. Sakib N, Saxena D, He L, Griffin P, Ahamed S, Haque M. Unpacking Prevalence and Dichotomy in qSOFA Parameters: A Step towards Multi-parameter Intelligent Sepsis Prediction in ICU. 2019 Presented at: 2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI); 2019; 2019; Chicago p. 1-4 URL: https://ieeexplore.ieee.org/document/8834547 [doi: 10.1109/BHI.2019.8834547]

22. Hwang SY, Shin J, Jo IJ, Park JE, Yoon H, Cha WC, et al. Delayed Antibiotic Therapy and Organ Dysfunction in Critically Ill Septic Patients in the Emergency Department. J Clin Med 2019 Feb 08;8(2):222 [FREE Full text] [doi: 10.3390/jcm8020222] [Medline: 30744073]
23. Kalil AJ, Dias VMDCH, Rocha CDC, Morales HMP, Fressatto JL, Faria RAD. Sepsis risk assessment: a retrospective analysis after a cognitive risk management robot (Robot Laura®) implementation in a clinical-surgical unit. Res Biomed Eng 2018 Nov 22;34(4):310-316. [doi: 10.1590/2446-4740.180021]
24. Feldman SS, Buchalter S, Hayes LW. Health Information Technology in Healthcare Quality and Patient Safety: Literature Review. JMIR Med Inform 2018 Jun 04;6(2):e10264 [FREE Full text] [doi: 10.2196/10264] [Medline: 29866642]
25. Lamichhane S, Manandhar N, Dhakal S, Shakya YL. Management and Outcome of Severe Sepsis and Septic Shock Patients Admitted to the Emergency Department in a Tertiary Hospital. J Nepal Health Res Council 2018 Jul 05;16(2):165-171. [doi: 10.3126/jnhrc.v16i2.20304]
26. Cecconi M, Evans L, Levy M, Rhodes A. Sepsis and septic shock. Lancet 2018 Jul 07;392(10141):75-87. [doi: 10.1016/S0140-6736(18)30696-2] [Medline: 29937192]
27. Khazaei H, McGregor C, Eklund JM, El-Khatib K. Real-Time and Retrospective Health-Analytics-as-a-Service: A Novel Framework. JMIR Med Inform 2015 Nov 18;3(4):e36 [FREE Full text] [doi: 10.2196/medinform.4640] [Medline: 26582268]
28. Ho K, Marsden J, Jarvis-Selinger S, Novak Lauscher H, Kamal N, Stenstrom R, et al. A collaborative quality improvement model and electronic community of practice to support sepsis management in emergency departments: investigating care harmonization for provincial knowledge translation. JMIR Res Protoc 2012 Jul 12;1(2):e6 [FREE Full text] [doi: 10.2196/resprot.1597] [Medline: 23611816]
29. Downey C, Randell R, Brown J, Jayne DG. Continuous Versus Intermittent Vital Signs Monitoring Using a Wearable, Wireless Patch in Patients Admitted to Surgical Wards: Pilot Cluster Randomized Controlled Trial. J Med Internet Res 2018 Dec 11;20(12):e10802 [FREE Full text] [doi: 10.2196/10802] [Medline: 30538086]
30. Johnson AEW, Aboab J, Raffa JD, Pollard TJ, Deliberato RO, Celi LA, et al. A Comparative Analysis of Sepsis Identification Methods in an Electronic Database. Crit Care Med 2018 Apr;46(4):494-499 [FREE Full text] [doi: 10.1097/CCM.0000000000002965] [Medline: 29303796]
31. Machado SM, Wilson EH, Elliott JO, Jordan K. Impact of a telemedicine eICU cart on sepsis management in a community hospital emergency department. J Telemed Telecare 2017 Feb 13;24(3):202-208. [doi: 10.1177/1357633x17691862]
32. Wellner B, Grand J, Canzone E, Coarr M, Brady PW, Simmons J, et al. Predicting Unplanned Transfers to the Intensive Care Unit: A Machine Learning Approach Leveraging Diverse Clinical Elements. JMIR Med Inform 2017 Nov 22;5(4):e45 [FREE Full text] [doi: 10.2196/medinform.8680] [Medline: 29167089]
33. Watkinson PJ, Barber VS, Young JD. Outcome of Critically ill Patients Undergoing Mandatory Insulin Therapy Compared to Usual Care Insulin Therapy: Protocol for a Pilot Randomized Controlled Trial. JMIR Res Protoc 2018 Mar 08;7(3):e44 [FREE Full text] [doi: 10.2196/resprot.5912] [Medline: 29519778]
34. Madrigal L, Escoffery C. Electronic Health Behaviors Among US Adults With Chronic Disease: Cross-Sectional Survey. J Med Internet Res 2019 Mar 05;21(3):e11240 [FREE Full text] [doi: 10.2196/11240] [Medline: 30835242]
35. Poncette A, Spies C, Mosch L, Schieler M, Weber-Carstens S, Krampe H, et al. Clinical Requirements of Future Patient Monitoring in the Intensive Care Unit: Qualitative Study. JMIR Med Inform 2019 Apr 30;7(2):e13064 [FREE Full text] [doi: 10.2196/13064] [Medline: 31038467]
36. Tseng Y, Wu J, Lin H, Chen M, Ping X, Sun C, et al. A Web-Based, Hospital-Wide Health Care-Associated Bloodstream Infection Surveillance and Classification System: Development and Evaluation. JMIR Med Inform 2015 Sep 21;3(3):e31 [FREE Full text] [doi: 10.2196/medinform.4171] [Medline: 26392229]
37. 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(1):160035. [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
38. Tsoukalas A, Albertson T, Tagkopoulos I. From data to optimal decision making: a data-driven, probabilistic machine learning approach to decision support for patients with sepsis. JMIR Med Inform 2015 Feb 24;3(1):e11 [FREE Full text] [doi: 10.2196/medinform.3445] [Medline: 25710907]
39. Johnson A, Stone D, Celi L, Pollard T. The MIMIC Code Repository: enabling reproducibility in critical care research. J Am Med Inform Assoc 2018 Jan 01;25(1):32-39 [FREE Full text] [doi: 10.1093/jamia/ocx084] [Medline: 29036464]
40. Bone RC, Balk RA, Cerra FB, Dellinger RP, Fein AM, Knaus WA, et al. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Consensus Conference Committee. American College of Chest Physicians/Society of Critical Care Medicine. Chest 1992 Jun;101(6):1644-1655. [doi: 10.1378/chest.101.6.1644] [Medline: 1303622]
41. Levy MM, Fink MP, Marshall JC, Abraham E, Angus D, Cook D, International Sepsis Definitions Conference. 2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference. Intensive Care Med 2003 Apr 28;29(4):530-538. [doi: 10.1007/s00134-003-1662-x] [Medline: 12664219]
42. Vincent J, Opal SM, Marshall JC, Tracey KJ. Sepsis definitions: time for change. Lancet 2013 Mar 02;381(9868):774-775 [FREE Full text] [doi: 10.1016/S0140-6736(12)61815-7] [Medline: 23472921]
43. Hotchkiss RS, Karl IE. The pathophysiology and treatment of sepsis. N Engl J Med 2003 Jan 09;348(2):138-150. [doi: 10.1056/NEJMra021333] [Medline: 12519925]

44. Remick DG. Pathophysiology of sepsis. Am J Pathol 2007 May;170(5):1435-1444 [FREE Full text] [doi: 10.2353/ajpath.2007.060872] [Medline: 17456750]
45. Gotts JE, Matthay MA. Sepsis: pathophysiology and clinical management. BMJ 2016 May 23;353:i1585. [doi: 10.1136/bmj.i1585] [Medline: 27217054]
46. Dewitte A, Lepreux S, Villeneuve J, Rigothier C, Combe C, Ouattara A, et al. Blood platelets and sepsis pathophysiology: A new therapeutic prospect in critically [corrected] ill patients? Ann Intensive Care 2017 Dec 01;7(1):115 [FREE Full text] [doi: 10.1186/s13613-017-0337-7] [Medline: 29192366]
47. Finkelsztein EJ, Jones DS, Ma KC, Pabón MA, Delgado T, Nakahira K, et al. Comparison of qSOFA and SIRS for predicting adverse outcomes of patients with suspicion of sepsis outside the intensive care unit. Crit Care 2017 Mar 26;21(1):73 [FREE Full text] [doi: 10.1186/s13054-017-1658-5] [Medline: 28342442]
48. Khwannimit B, Bhurayanontachai R, Vattanavanit V. Comparison of the performance of SOFA, qSOFA and SIRS for predicting mortality and organ failure among sepsis patients admitted to the intensive care unit in a middle-income country. J Crit Care 2018 Apr;44:156-160. [doi: 10.1016/j.jcrc.2017.10.023] [Medline: 29127841]
49. Haydar S, Spanier M, Weems P, Wood S, Strout T. Comparison of QSOFA score and SIRS criteria as screening mechanisms for emergency department sepsis. Am J Emerg Med 2017 Nov;35(11):1730-1733. [doi: 10.1016/j.ajem.2017.07.001] [Medline: 28712645]
50. Donnelly JP, Safford MM, Shapiro NI, Baddley JW, Wang HE. Application of the Third International Consensus Definitions for Sepsis (Sepsis-3) Classification: a retrospective population-based cohort study. Lancet Infect Dis 2017 Jun;17(6):661-670 [FREE Full text] [doi: 10.1016/S1473-3099(17)30117-2] [Medline: 28268067]
51. Hwang SY, Jo IJ, Lee SU, Lee TR, Yoon H, Cha WC, et al. Low Accuracy of Positive qSOFA Criteria for Predicting 28-Day Mortality in Critically Ill Septic Patients During the Early Period After Emergency Department Presentation. Ann Emerg Med 2018 Jan;71(1):1-9. [doi: 10.1016/j.annemergmed.2017.05.022] [Medline: 28669551]
52. Fernando SM, Tran A, Taljaard M, Cheng W, Perry JJ. Prognostic Accuracy of the Quick Sequential Organ Failure Assessment for Mortality in Patients With Suspected Infection. Ann Intern Med 2018 Aug 21;169(4):264-265. [doi: 10.7326/L18-0291] [Medline: 30128520]
53. Armitage DW, Ober HK. A comparison of supervised learning techniques in the classification of bat echolocation calls. Ecol Informat 2010 Nov;5(6):465-473. [doi: 10.1016/j.ecoinf.2010.08.001]

Abbreviations

ICU: intensive care unit
MIMIC-III: Medical Information Mart for Intensive Care
qSOFA: quick Sequential Organ Failure Assessment
SIRS: systematic inflammatory response syndrome
SOFA: Sepsis-related Organ Failure Assessment

Edited by G Eysenbach; submitted 21.02.20; peer-reviewed by L Santacroce, S Manaktala, B Popoff, D Clifton; comments to author 29.06.20; revised version received 10.08.20; accepted 15.09.20; published 03.12.20.

Please cite as:
Sakib N, Ahamed SI, Khan RA, Griffin PM, Haque MM
Unpacking Prevalence and Dichotomy in Quick Sequential Organ Failure Assessment and Systemic Inflammatory Response Syndrome Parameters: Observational Data–Driven Approach Backed by Sepsis Pathophysiology
JMIR Med Inform 2020;8(12):e18352
URL: https://medinform.jmir.org/2020/12/e18352
doi:10.2196/18352
PMID:33270030

Original Paper

Comprehensive Computer-Aided Decision Support Framework to Diagnose Tuberculosis From Chest X-Ray Images: Data Mining Study

Muhammad Owais[1], MSc; Muhammad Arsalan[1], PhD; Tahir Mahmood[1], MSc; Yu Hwan Kim[1], MSc; Kang Ryoung Park[1], PhD

Division of Electronics and Electrical Engineering, Dongguk University, Seoul, Republic of Korea

Corresponding Author:
Kang Ryoung Park, PhD
Division of Electronics and Electrical Engineering
Dongguk University
30 Pildong-ro 1-gil, Jung-gu
Seoul, 04620
Republic of Korea
Phone: 82 10 3111 7022
Fax: 82 2 2277 8735
Email: parkgr@dgu.edu

Abstract

Background: Tuberculosis (TB) is one of the most infectious diseases that can be fatal. Its early diagnosis and treatment can significantly reduce the mortality rate. In the literature, several computer-aided diagnosis (CAD) tools have been proposed for the efficient diagnosis of TB from chest radiograph (CXR) images. However, the majority of previous studies adopted conventional handcrafted feature-based algorithms. In addition, some recent CAD tools utilized the strength of deep learning methods to further enhance diagnostic performance. Nevertheless, all these existing methods can only classify a given CXR image into binary class (either TB positive or TB negative) without providing further descriptive information.

Objective: The main objective of this study is to propose a comprehensive CAD framework for the effective diagnosis of TB by providing visual as well as descriptive information from the previous patients' database.

Methods: To accomplish our objective, first we propose a fusion-based deep classification network for the CAD decision that exhibits promising performance over the various state-of-the-art methods. Furthermore, a multilevel similarity measure algorithm is devised based on multiscale information fusion to retrieve the best-matched cases from the previous database.

Results: The performance of the framework was evaluated based on 2 well-known CXR data sets made available by the US National Library of Medicine and the National Institutes of Health. Our classification model exhibited the best diagnostic performance (0.929, 0.937, 0.921, 0.928, and 0.965 for F1 score, average precision, average recall, accuracy, and area under the curve, respectively) and outperforms the performance of various state-of-the-art methods.

Conclusions: This paper presents a comprehensive CAD framework to diagnose TB from CXR images by retrieving the relevant cases and their clinical observations from the previous patients' database. These retrieval results assist the radiologist in making an effective diagnostic decision related to the current medical condition of a patient. Moreover, the retrieval results can facilitate the radiologists in subjectively validating the CAD decision.

(JMIR Med Inform 2020;8(12):e21790) doi:10.2196/21790

KEYWORDS

tuberculosis; computer-aided diagnosis; chest radiograph; lung disease; neural network; classification-based retrieval

Introduction

According to a World Health Organization (WHO) report, tuberculosis (TB) is a major global health problem that causes severe medical conditions among millions of people annually. It ranks along with the HIV as a leading cause of mortality worldwide [1]. In 2014, approximately 9.6 million new TB cases were reported as per the WHO report, which ultimately caused 1.5 million deaths [1]. Today, early diagnosis and proper

treatment can cure almost all the TB cases. Various types of laboratory tests have been developed to diagnose TB [2,3]. Among these tests, sputum smear microscopy is the most common, in which bacteria are examined from sputum samples using a microscope [2]. Developed in the last few years, molecular diagnostics [3] are the new techniques to diagnose TB. However, they may not be suitable in real-time screening applications. Currently, chest radiography is the most common test to detect pulmonary TB worldwide [4]. It has become cheaper and easier to use with the advent of digital chest radiography [5]. However, all these diagnostic tests are assessed by specialized radiologists, who must expend significant time and effort to make an accurate diagnostic decision. Therefore, such subjective methods may not be suitable for real-time screening.

Over the past few years, researchers have made a significant contribution to the development of computer-aided diagnosis (CAD) tools related to chest radiography [6,7]. Such automated tools can detect the various type of chest abnormalities within seconds and can aid in population screening applications, particularly in scenarios which lack medical expertise. Fortunately, the recent development in artificial intelligence has presented a remarkable breakthrough in the performance of these tools. Deep learning algorithms, specifically artificial neural networks [8], are the state-of-the-art achievement in the artificial intelligence domain. These algorithms offer more reliable methods to distinguish positive and negative TB cases from chest radiographs (CXR) images in a fully automated manner. In recent decades, several ground-breaking CAD methods have been proposed for TB diagnosis [9-24]. Most of the previous studies used segmentation-, detection-, and classification-based approaches to make the ultimate diagnostic decisions. All these methods indicated a binary decision (either TB positive or TB negative) without providing further descriptive information that may assist medical experts to validate the CAD decision. As the CAD decision can also be erroneous in some scenarios, a method to perform its cross-validation is necessary. Therefore, further research is required to achieve the practical performance and usability of such diagnostic systems in the real world. A comprehensive analysis of these existing studies [9-24] in comparison with our proposed method can be found in Multimedia Appendix 1.

Recently, various types of artificial neural networks have been proposed in the domain of general image processing to achieve the maximum performance in terms of accuracy (ACC) and computational cost. Among these models, convolutional neural networks (CNNs) [25] attract special attention because of their outstanding performance in many general and medical image recognition applications [26,27]. The entire structure of a CNN model consists of an input layer, hidden layers, and a final output layer. Among all these layers, hidden layers are considered the main components of the CNN model and primarily consist of a series of convolutional layers that include trainable filters of different sizes and depths. These filters are trained by performing a training procedure to extract the deep features from a training data set. When the training procedure is completed, the trained network can analyze the given testing data and generate the desired output.

In this paper, a novel CAD framework is proposed to diagnose TB from a given CXR image and provide the appropriate visual and descriptive information from a previous database, which can further assist radiologists to subjectively validate the computer decision. Thus, both subjective and CAD decisions will complement each other and ultimately result in effective diagnosis and treatment. The performance of our proposed framework was evaluated using 2 well-known CXR data sets [9,28]. The overall performance of our method is substantially higher than that of various state-of-the-art methods. The main contributions of our work can be summarized as follows:

1. To the best of our knowledge, this is the first comprehensive CAD framework in chest radiography based on multiscale information fusion that effectively diagnoses TB by providing visual and descriptive information based on a previous patients' database.
2. We propose an ensemble classification model obtained by integrating 2 CNNs named shallow CNN (SCNN) to capture the low-level features such as edge information and a deep CNN (DCNN) to extract high-level features such as TB patterns.
3. Furthermore, a multilevel similarity measure (MLSM) algorithm is proposed based on multiscale information fusion to retrieve the best-matched cases from a previous database by computing a weighted structural similarity (SSIM) score of multilevel features.
4. The cross-data analysis (trained with one data set and tested with another data set, and vice versa) is a key measure to access the generalizability of a CAD tool. However, in the medical image analysis domain, most of the existing studies [9-15,18,19,21-24] did not analyze the performance of their methods in cross data set. Therefore, to further highlight the discriminative power of the proposed model in real-world scenarios, we also analyzed its performance in a cross data set.

The remainder of the paper is structured as follows. In the "Methods" section, we describe our proposed framework. Subsequently, the experimental results along with the data set, the experimental setup, and the performance evaluation metrics are provided in the "Results" section. Finally, the "Discussion" section presents the comprehensive discussions of our paper including the principal findings.

Methods

This section presents a comprehensive description of our proposed framework in the following sequential order. First, we provide a brief overview of the proposed method to describe its end-to-end workflow. Subsequently, a detailed explanation of our proposed classification model and similarity measuring algorithm is presented in subsequent subsections.

Overview of Our Proposed Framework

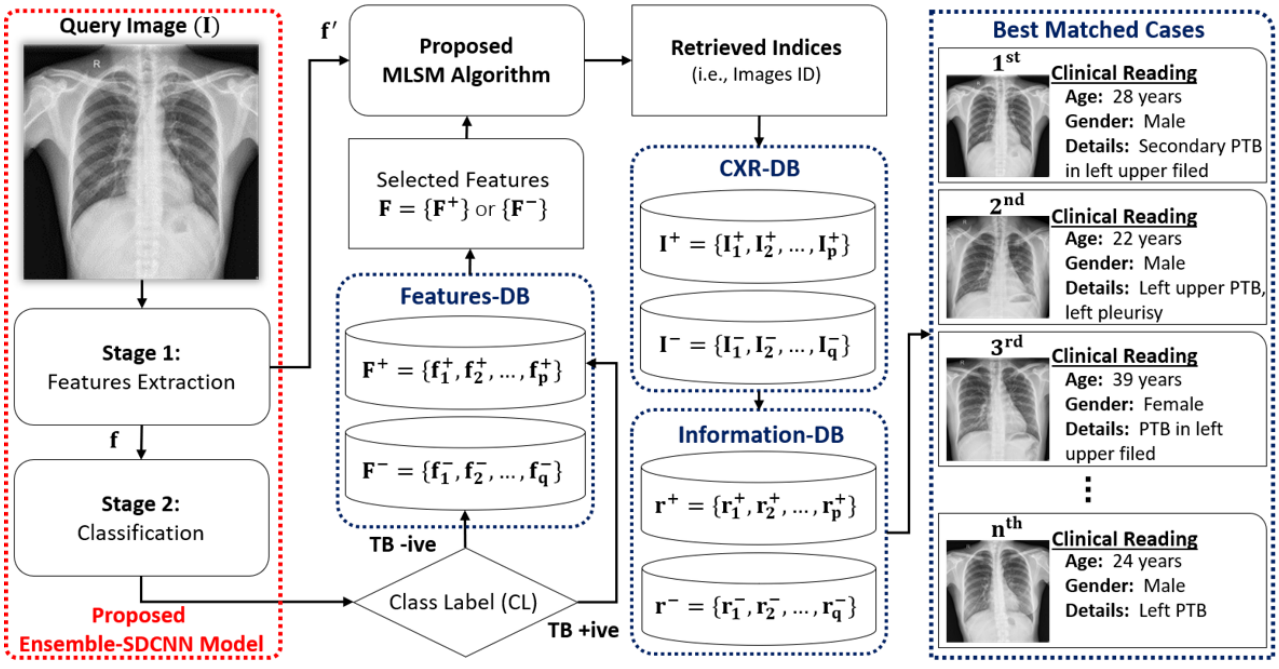
In general, the overall performance of the image classification and retrieval framework is directly related to the mechanism of feature extraction, which is adopted to transform the visual data from high-level semantics to low-level features. These low-level features incorporate the distinctive information that can easily

distinguish the instances of multiple classes. Recently, deep learning methods provide a fully automated means to extract the optimal features from available training data sets and lead to a substantial performance gain. In this study, we used the strengths of such deep learning methods to develop a comprehensive CAD tool to diagnose TB from CXR images. A comprehensive representation of the proposed framework is shown in Figure 1. The complete framework comprised a classification stage, a retrieval phase to perform the diagnostic decision, and retrieval of the descriptive evidence, respectively. In the first phase, our proposed ensemble-shallow–deep CNN (ensemble-SDCNN) model was trained to make the diagnostic decision for the given CXR image I by predicting its class label (CL) as either TB positive or TB negative. Such a diagnostic decision was made into 2 stages: feature extraction and classification. The detailed explanation of the proposed ensemble-SDCNN model and its workflow is provided in the subsequent subsection.

In the second phase, a classification-driven retrieval was performed for the input query image. The ultimate objective of this phase was to retrieve the relevant cases (such as CXR images) corresponding to the given CXR image with the inclusion of clinical observations (such as textual description) from the previous patients' database. Such retrieval results can assist radiologists to subjectively validate the computer diagnostic decision, which ultimately results in an effective diagnostic decision. Initially, based on the predicted CL (in the first phase), a set of positive or negative feature vectors was selected from features database based on the following predefined criteria: $F = F^+$, if CL = TB positive; otherwise $F = F^-$, where $F^+$and F^-present the set of positive ($F^+ = \{f_1^+, f_2^+, ..., f_p^+\}$) and negative features maps ($F^- = \{f_1^-, f_2^-, ..., f_q^-\}$) in the features database, respectively, and p and q are the total numbers of positive and negative cases, respectively.

Both F^+ and F^- were extracted from TB-positive and TB-negative CXR-database (previously collected CXR images of different patients), respectively, and stored as a features database. In the subsequent step, our proposed MLSM algorithm was applied to select a subset of n best-matched features from this selected set of positive or negative features maps (ie, $F=\{F^+\}$ or $\{F^-\}$) in the first phase. Such feature matching was performed for the extracted multilevel features f′ of input query image I (as explained in a later subsection). Finally, the selected subset of n best-matched features was used to select the corresponding CXR images and their clinical readings from CXR-database and information database, respectively.

Figure 1. Comprehensive flow diagram of the proposed classification and retrieval framework. In the first stage, the given input CXR image is categorized as either TB positive or TB negative. In the second stage, the n best relevant cases are retrieved from the previous database based on our proposed MLSM algorithm. The parameter n is a user given input and controls the total number of retrieved cases from the previous record related to a current medical condition. CXR: chest radiograph; DB: database; MLSM: multilevel similarity measure; SDCNN: shallow–deepCNN; TB: tuberculosis.



Classification Network

The first phase of our proposed framework involved classifying the given CXR image as either TB positive or TB negative by predicting its CL. To accomplish this task, we proposed a jointly connected ensemble-SDCNN model by performing a features-level fusion of 2 different networks, SCNN and DCNN (Figure 2). In general, a shallow network captures low-level features such as edge information while a deep model is used to exploit high-level information such as overall shape patterns. In our radiograph image analysis study, the experimental results prove that the combination of low- and high-level features results in better performance compared with using only high-level features. Therefore, both networks were combined in parallel (by connecting their input and last output layers with each other; Figure 2) to create a single end-to-end trainable network. An existing DCNN model called a residual network (ResNet18) [29] was selected based on its substantial

classification performance and the optimal number of parameters in comparison with the other CNN models. After selecting an optimal DCNN model, we further enhanced its performance by connecting our proposed SCNN model in parallel to it. Several experiments were performed to select the optimal number of convolutional and fully connected (FC) layers (and their hyper parameters) for the SCNN. The ultimate objective of these experiments was to construct an optimal shallow network (according to the number of parameters) that could maximize the overall classification performance of the complete network.

A complete layer-wise configuration of these models is shown in Table 1. This information can assist in exploring the parametric configuration of these models more precisely. Moreover, Figure 2 shows the overall architecture of the proposed ensemble-SDCNN model based on shallow and deep networks. Both SCNN and DCNN models processed the given CXR image in a parallel order to extract low- and high-level features, respectively. In the SCNN, the Conv1 layer (first convolutional layer with a total of 128 filters of size 7 × 7) explored the input image I in both horizontal and vertical directions and generated the output feature map, F_{SN1} of size 73 × 73 × 128. This output feature map was further processed through the Conv2 layer (second convolutional layer with a total of 64 filters of size 5 × 5) and converted into a new features map F_{SN2} of size 35 × 35 × 64. Thereafter, the FC1 layer (first fully connected layer including a total of 32 output nodes) identified the significant hidden patterns in F_{SN2} by combining all the learned features into a single features vector f_{SN} of size 1 × 1 × 32. Thus, we obtained a low-dimension features vector f_{SN}that held a more diverse representation of the low-level features compared with F_{SN2}.

Figure 2. Overall architecture of our ensemble-SDCNN model by connecting 2 different networks, SCNN and DCNN. Both networks process the input image I simultaneously (in the testing phase) and extract 2 different feature vectors, which are concatenated and finally used to make a diagnostic decision by predicting the CL. CL: class label; CNN: convolutional neural network; DCNN: deep CNN; SCNN: shallow CNN; SDCNN: shallow–deep CNN.

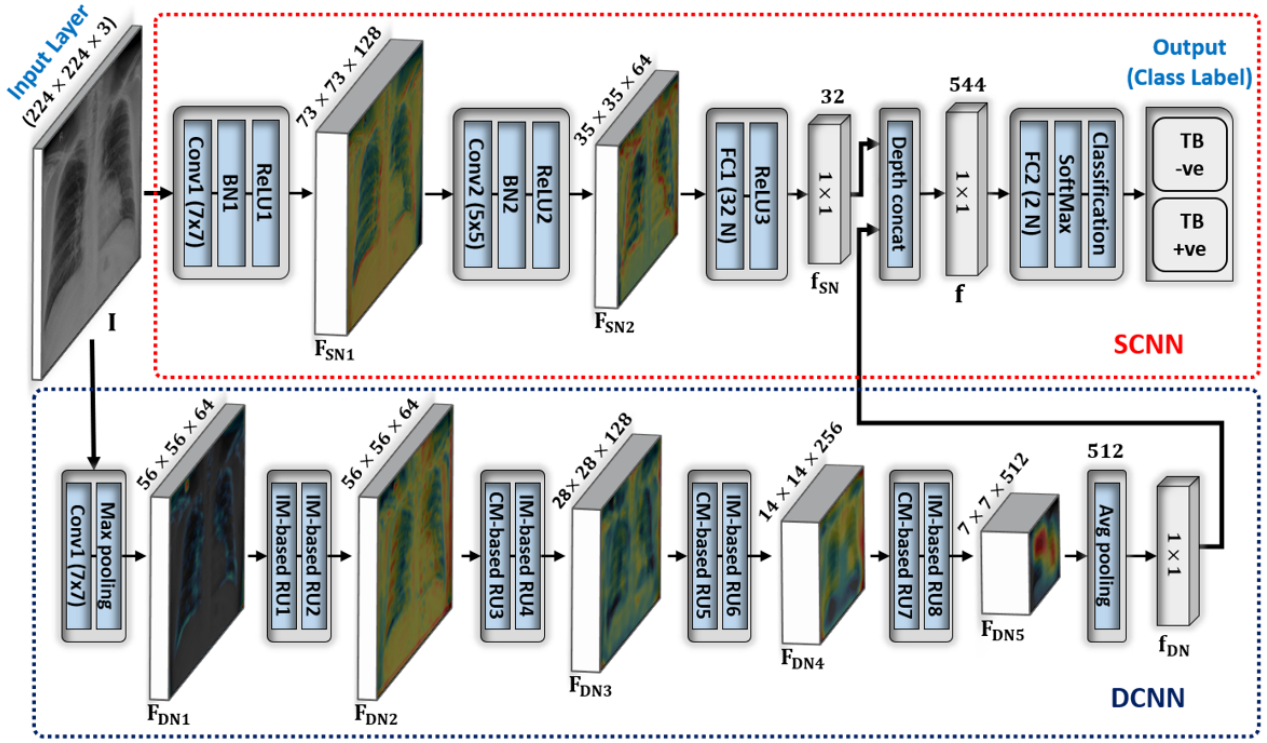


Table 1. Layer-wise configuration details of the proposed ensemble-SDCNN[a] model.[b]

Layer name	Output size[c]	Filter size[d]	Iterations	Parameters
DCNN[e] model				
Input	(224,224,3)	N/A[f]	—[g]	—
Conv1	(112,112,64)	(7,7,64)	1	9600
Max pooling	(56,56,64)	(3,3)	1	—
IM[h]-based RU1[i]	(56,56,64)	(3,3,64)	2	74,112
IM-based RU2	(56,56,64)	(3,3,64)	2	74,112
CM[j]-based RU3	(28,28,128)	(3,3,128); (1,1,128)	2; 1	230,528
IM-based RU4	(28,28,128)	(3,3,128)	2	295,680
CM-based RU5	(14,14,256)	(3,3,256); (1,1,256)	2; 1	919,808
IM-based RU6	(14,14,256)	(3,3,256)	2	1,181,184
CM-based RU7	(7,7,512)	(3,3,512); (1,1,512)	2; 1	3,674,624
IM-based RU8	(7,7,512)	(3,3,512)	2	4,721,664
Avg pooling	(1,1,512)	(7,7)	1	—
SCNN[k] model				
Conv1	(112,112,128)	(7,7,128)	1	19,200
Conv2	(35,35,64)	(5,5,64)	1	204,992
FC1	(1,1,32)	(5,5,64)	1	2,508,832
Depth concat	(1,1,544)	—	1	—
FC2	(1,1,2)	—	1	1090
SoftMax	(1,1,2)	—	1	—
Classification	2	—	1	—

[a]SDCNN: shallow–deep CNN.

[b]Total learnable parameters: 13,915,426.

[c]Output size (image width, image height, # of channels),

[d]Kernel size (kernel width, kernel height, # of filters), Max pooling (kernel width, kernel height), Avg pooling (kernel width, kernel height).

[e]DCNN: deep CNN.

[f]N/A: not applicable.

[g]—: not available.

[h]IM: identity mapping.

[i]RU: residual unit.

[j]CM: convolutional mapping.

[k]SCNN: shallow CNN.

Similarly, for the DCNN, the input image I passes through a large number of convolutional layers (as compared with the SCNN) to exploit the high-level features. Our selected DCNN model was composed of multiple residual units (RUs) that consisted of identity mapping–based or convolutional mapping–based shortcut connections to each pair of 3 × 3 filters [29]. These shortcut connections caused the network to converge more efficiently compared with other sequential networks without including any shortcut connection. Moreover, a detailed explanation of these RUs is provided in [30]. Figure 2 also depicts an abstract representation of our selected DCNN model.

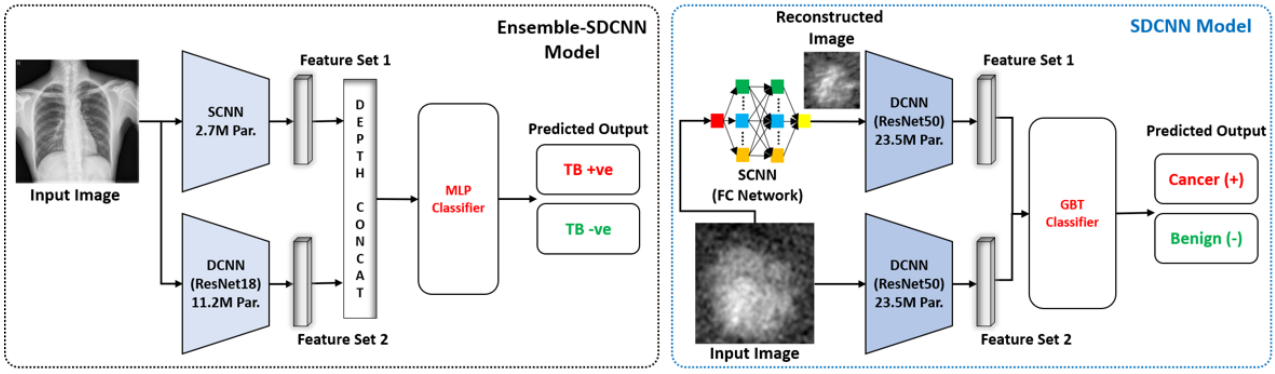
Primarily, the input image I underwent the first convolutional layer, Conv1, with a total 64 filters of size 7 × 7. Subsequently, a Max pooling layer (with a window size 3 × 3) further down sampled the output of Conv1 and generated an intermediate features map F_{DN1} of size 56 × 56 × 64. Thereafter, a stack of 8 consecutive RUs (including 5 identity mapping–based RUs and 3 convolutional mapping–based RUs, as shown in Figure 2) further exploited high-level features. Furthermore, each RU converted the preceding features map into a new one by exploiting much deeper features in comparison with the previous layer. In Figure 2, all the intermediate features maps (ie, F_{DN2},

F_{DN3}, F_{DN4}, and F_{DN5}) after each pair of RU show the progressive effect of different RUs. We observed that the depth of these features maps increased progressively, and the spatial size decreased after passing through the RUs. Ultimately, a low-dimension feature vector, f_{DN}, of size 1 × 1 × 512 was obtained after processing the final features map, F_{DN5} (obtained from the last RU), through an average pooling layer. This low-dimension feature vector exhibited a high-level abstraction of the input image I and substantially contributed, together with f_{SN}, to the prediction of the final CL.

After extracting both low- and high-level features, a depth concatenation layer (labeled as Depth concat in Figure 2 and Table 1) performed the feature-level fusion by combining both f_{SN} and f_{DN} along the depth direction and generated a final features vector, f, of size 1 × 1 × 544. Finally, a stack of the FC2, SoftMax, and the classification layers (Figure 2) acted as a multilayer perceptron classifier and predicted the CL for the given image I using the ultimate features vector f. In this stack, the FC2 layer (including the number of nodes equal to the total number of classes) identified the larger patterns in fby combining all the features values. It multiplied f by a weight matrix W, and then added a bias vector b, where y = W·f + b, with y = $[y_i|_{i=1,2}]$. Subsequently, the SoftMax layer converted the output of FC2 in terms of probability by applying the softmax function as $y'_i=e^{yi}/\Sigma^2_{i=1}$ [8]. Ultimately, the classification layer obtained (y'_i)from the SoftMax layer was assigned each input to one of the 2 mutually exclusive classes (ie, TB positive and TB negative) using a cross-entropy (CE) loss function as $Loss_{CE}(W,b) = \Sigma^2_{i=1}\ c_i \ln(y'_i)$ [8]. Here, (W, b) are the network trainable parameters and c_i is the indicator of the actual class label of the *i*th class during the training procedure. Meanwhile, in the testing phase, the network generated a single CL (as either TB positive or TB negative) corresponding to each input image I.

There is also an existing SDCNN model [31] (proposed for effective breast cancer diagnosis). However, there is a substantial difference between our proposed and the existing model [31] in terms of architecture, application, and computational complexity. In [31], the authors proposed an ensemble of 2 existing ResNet50 [29] models to extract the deep features and then used a gradient boosted tree classifier to make the diagnostic decision. In addition, a 4-layer FC network, namely SCNN (which includes FC convolutional layers), was proposed for image reconstruction to increase the data samples in the preprocessing stage. By contrast, in our work, we proposed an ensemble of SCNN (which includes 2 convolutional layers [no FC] and 1 FC layer) and DCNN models as shown in Figure 2 to extract low- and high-level features, respectively. Then, an FC classifier (also known as a multilayer perceptron) was used to make the final diagnostic decision using both low- and high-level features. Furthermore, the SCNN [31] is an image reconstruction network (ie, both input and output are images), whereas our proposed SCNN is a classification network (ie, input is image, and output is feature vector). Therefore, the architecture of both SCNN models is completely different from each other. In addition, our DCNN model is based on ResNet18 that includes a substantially lower number of trainable parameters than ResNet50 as used in [31], that is, 11.2M (ResNet18) << 23.5M (ResNet50). In this way, the total number of trainable parameters of the proposed ensemble-SDCNN is substantially lower than the existing SDCNN [31], that is, 13.9M (proposed) << 47M [31]. Figure 3 further highlights the overall structural difference between our proposed and the existing model [31].

Figure 3. Overall structural comparison of our proposed ensemble-SDCNN (left) and existing SDCNN model (right). MLP: multilayer perceptron; GBT: gradient boosted tree.



Multilevel Similarity Measure Algorithm

In the medical domain, the visually correlated images occasionally depict different illnesses, whereas the images for a similar ailment have distinctive appearances. Therefore, estimating the similarity by contemplating the multilevel features is more advantageous in content-based medical image retrieval systems rather than using single-level features. Most of the existing systems often use a single-level similarity measure (SLSM) method to perform the content-based medical image retrieval task. However, it can miss the potentially useful information that is required in discriminating the different diseases in visually correlated images. To overcome these challenges, we proposed an MLSM algorithm to retrieve the best-matched cases from the previous patients' database by fusing multilevel features starting from a low-level visual to a high-level semantic scale. The similarity at multiple features levels was calculated using a well-known matching algorithm called SSIM [32], as it quantified the visibility of errors (differences) between 2 samples more appropriately compared with other simple matching schemes such as mean square error, peak signal-to-noise ratio (PSNR), and Euclidean distance. A

generalized mathematical expression to calculate the SSIM score between 2 samples (x and y) is given as follows:

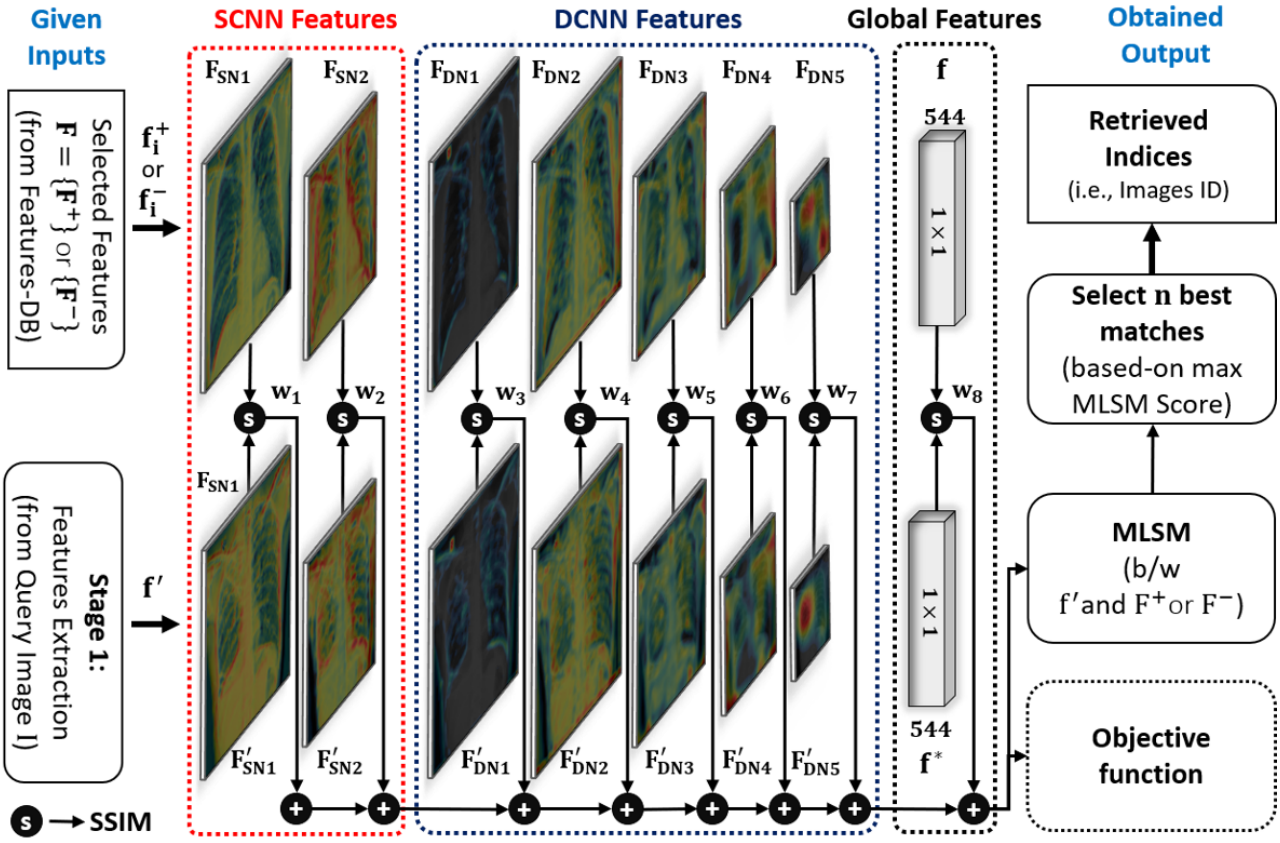
$$SSIM(x,y) = ([2\mu_x\mu_x + c1][2\sigma_{xy} + c_2])/[\mu^2_x + \mu^2_y + c_1][\sigma^2_x + \sigma^2_y + c_2] \quad \textbf{(1)}$$

where $[\mu_x \cdot \mu_y]$, $[\sigma_x \cdot \sigma_y]$, and σ_{xy} are the local mean, standard deviation, and cross-covariance of the given samples, respectively; and c_1 and c_2 are constants to avoid instabilities such as infinity errors and undefined solutions.

In our MLSM algorithm, multilevel features were extracted from the 8 different locations of the ensemble-SDCNN model (Figure 4). Each features map in Figure 4 was obtained by calculating the depth-wise averaging of each stack of feature maps (extracted from a particular location). Moreover, this newly obtained feature map corresponding to each specific location was further presented with a pseudocolor scheme to highlight the activated regions more appropriately. In Figure 4, f′ presents a set of these multilevel features maps (ie, {F'_{SN1}, F'_{SN2}, F'_{DN1}, F'_{DN2}, F'_{DN3}, F'_{DN4}, F'_{DN5}, f*}) corresponding to the given query image I. Similarly, f^+_i or f^-_i notates a set of multilevel features maps (ie, {F_{SN1}, F_{SN2}, F_{DN1}, F_{DN2}, F_{DN3}, F_{DN4}, F_{DN5}, f}) for the *i*th positive or negative sample image in CXR-database, respectively. The selection of f^+_i or f^-_i was conducted based on the CL prediction, which was performed by our proposed network in the first phase. For example, in a positive prediction (ie, CL = TB positive) for the input query image I, the MLSM score between the query image I and set of p positive sample images I^+ (stored in CXR-database) is calculated as follows:

$$MLSM = \Sigma^8_{k=1} w_k SSIM(f'\{k\}, f^+_i\{k\})_{i=1, 2, \ldots, p} \quad \textbf{(2)}$$

Figure 4. Complete workflow diagram of our proposed MLSM algorithm using the multilevel features (extracted from the different parts of the proposed ensemble-SDCNN model) in retrieving the best-matched cases from a previous patients' database. DCNN: deep convolutional neural network; MLSM: multilevel similarity measure; SCNN: shallow convolutional neural network; SSIM: structure similarity.



Similarly, in a negative prediction (ie, CL = TB negative), the MLSM score between the query image I and set of q negative sample images I^- (also stored in CXR-database) is calculated as follows:

$$MLSM = \Sigma^8_{k=1} w_k SSIM(f'\{k\}, f^-_i\{k\})_{i=1, 2, \ldots, q} \quad \textbf{(3)}$$

In both mathematical expressions, w_1, w_2, w_3, …, w_8 are the weights of SSIM measured at different levels and their total sum is equal to one (ie, $\Sigma^8_{i=1} w_i = 1$). The optimal weights were obtained by maximizing the intraclass SSIM score for some selected pairs of positive CXR images. Each pair (I^+_i, I^+_j) was selected from the positive data samples based on the highly correlated clinical observations between 2 CXR images. These observations were provided in our selected data sets as a text file for each data sample. As our main objective was to diagnose TB by retrieving similar abnormal cases from a previous database, we only considered positive CXR images in calculating the optimal weights rather than using normal images.

Finally, the overall objective function to maximize the intraclass similarity is defined as follows:

$$w^* = \max\left(\Sigma_{i,j}\varepsilon TB_{positive}\Sigma_{k=1}^{8} w_k \, SSIM[f^+_i\{k\}, f^+_j\{k\}]\right) / N^+ \quad \textbf{(4)}$$

where N^+ is the total number of pair images selected from the positive data samples. In our experiment, the total number of pairs was 30 (ie, $N^+ = 30$). After performing the optimization according to Equation (4), we obtained the optimal values of w_1, w_2, w_3, …, w_8 as 0.069, 0.179, 0.087, 0.133, 0.071, 0.123, 0.299, and 0.039, respectively. Finally, these optimal weights were used to calculate the MLSM scores between f′ and F^+ (set of positive features maps in features database) or F^- (set of negative features maps in features database) depending on the predicted CL in the classification stage. Thereafter, the indices of n best-matched features were selected based on the maximum MLSM scores. These indices were eventually used to select the corresponding CXR images and their clinical readings from CXR-database and information database, respectively. Thus, n best-matched cases were retrieved from the previous patients' database, which could assist radiologists in making an effective diagnostic decision after performing the subjective validation of the computer decision.

Results

Data Set and Preprocessing

Our proposed diagnostic framework was validated using 2 publicly available data sets: Montgomery County (MC) and Shenzhen (SZ) [9,28]. The MC data set is a subset of a larger CXR repository collected within the TB control program of the Department of Health and Human Services of Montgomery County, Maryland, USA. All these images are in 12-bit grayscale, captured using a Eureka stationary X-ray machine. This data set comprises a total of 138 posteroanterior CXR images, among which there are 80 normal and 58 abnormal images with the manifestations of TB disease. The abnormal images encompass a vast range of abnormalities related to pulmonary TB. The SZ data set is collected from the Shenzhen No. 3 People's Hospital in Shenzhen, Guangdong Providence, China. This data set includes a total of 326 normal and 336 abnormal CXR images, which include different types of abnormalities related to pulmonary TB. All these images are also in 12-bit grayscale and were captured using the Philips DR DigitalDiagnost system. In both data sets, a radiologist report is also provided for each CXR image as a text file, containing the clinical observation related to chest abnormalities along with the patient's age and gender information. After collecting both data sets, we resized all the images to a spatial dimension of 224 × 224 (according to the fixed input layer size of our ensemble-SDCNN model).

Implementation Details

The proposed framework was implemented using a standard deep learning toolbox available in the MATLAB R2019a (MathWorks, Inc.) framework [33]. It provides a complete framework for developing and testing different types of artificial neural networks and using existing pretrained networks. All the experiments were performed on a desktop computer with a 3.50-GHz Intel Core i7-3770K CPU [34], 16-GB RAM, an NVIDIA GeForce GTX 1070 graphics card [35], and Windows 10 operating system (Microsoft). Our proposed and other baseline models were trained through back-propagation (a procedure to determine the optimal parameters of a model) using a well-known optimization algorithm called the stochastic gradient descent [36]. It iteratively trains the network by computing the optimal learnable parameters (such as filter weights and biases) that are included in different layers of the network. The following hyper-parameters were selected for our proposed and all the comparative CNN-based methods: learning rate as 0.001 with a drop factor of 0.1. Moreover, the min-batch size was selected as 10 (ie, feeding a stack of 10 images per gradient update in each iteration), L2-regularization as 0.0001, and a momentum factor as 0.9.

Evaluation Metrics and Protocol

After the training, the quantitative performance of our proposed framework was evaluated based on the following metrics: ACC, average precision (AP), average recall (AR), F1 score (F1), and finally the area under the curve (AUC) [37]. These well-known metrics can quantify the overall performance of a deep learning model from many perspectives. The mathematical definition of all these metrics is provided in Table 2.

In our binary classification problem, true positive (TP) and true negative (TN) were the outcomes of our model for correctly predicted positive and negative cases, respectively, whereas false positive (FP) and false negative (FN) could be interpreted as the incorrectly predicted positive and negative cases, respectively. Finally, these 4 different outcomes were further used in assessing the overall performance of a model in terms of ACC, AP, AR, F1, and AUC. We performed a fivefold cross-validation in all the experiments by randomly selecting 80% of data (110/138 [79.7%] of MC data and 530/662 [80.0%] SZ data) for training and the remaining 20% (28/138 [20.2%] of MC data and 132/662 [19.9%] SZ data) for testing. As most of the previous studies considered fivefold cross-validations, we followed a similar data splitting protocol. However, the fivefold cross-validation was not possible for the evaluation of the cross–data set performance, as a complete data set was used for training and others for testing. However, we performed cross-data validation using the MC data set as training and the SZ data set as testing, and vice versa.

Table 2. Mathematical definition of our selected performance evaluation metrics.

Metric name	Mathematical equation
Accuracy (ACC)	$(TP^a + TN^b)/(TP + TN + FP^c + FN^d)$
Average precision (AP)	TP/(TP + FP)
Average recall (AR)	TP/(TP + FN)
F1 score (F1)	2×([AP × AR]/[AP + AR])
Area under the curve (AUC)	0.5 × (TP/[TP + FP] + TN/[TN + FP])

[a]TP: true positive.

[b]TN: true negative.

[c]FP: false positive.

[d]FN: false negative.

Our Results and an Ablation Study

The overall performance of our diagnostic framework was directly related to the classification performance of the proposed ensemble-SDCNN model. As in our classification-driven retrieval framework, the first step was to predict the CL for the given query image and then explore that predicted class database to retrieve the relevant cases. Consequently, the correct prediction would ultimately result in correct retrieval and the incorrect prediction in incorrect retrieval. Therefore, we comprehensively assessed the classification performance of the proposed model for both data sets and their combinations. Table 3 shows the performance of our classification model along with an ablation study to highlight the significance of each submodel in enhancing the overall performance. Therefore, the individual performance of both SCNN and DCNN models was also computed as an ablation study. The experimental results indicated that the combination of SCNN and DCNN resulted in a substantial performance gain (ie, 8.8%, 8.12%, 9.42%, 8.76%, and 5.68% for the average F1, AP, AR, ACC, and AUC, respectively) compared with their individual performances. We further performed a *t* test [38] and Cohen *d* [39] analysis to signify the performance gain of our SDCNN model in contrast to the DCNN (second-best model). In these 2 performance analysis measures, a large number of experimental results appropriately discriminated the performances of 2 systems.

Therefore, the detailed performance results of both ensemble-SDCNN and DCNN for all the different folds were used to perform the *t* test and Cohen *d* analysis. In the *t* test analysis, all the *P*-values (ie, .012, .011, .015, .014, and .012 in the case of average F1, AP, AR, ACC, and AUC, respectively) were less than .05. These results implied the discriminative performance of our ensemble-SDCNN against the SCNN with a 95% confidence score. In the Cohen *d* analysis, the performance difference between 2 systems was measured in terms of effect size [40], which is generally categorized as small (approximately 0.2-0.3), medium (approximately 0.5), and large (≥0.8). The large effect size indicated a substantial performance difference between the 2 systems. In this analysis, all the effect sizes (ie, 0.6, 0.6, 0.6, 0.5, and 0.5 for the average F1, AP, AR, ACC, and AUC, respectively) were greater than and equal to 0.5, which also indicated the substantial performance difference between the ensemble-SDCNN and SCNN models.

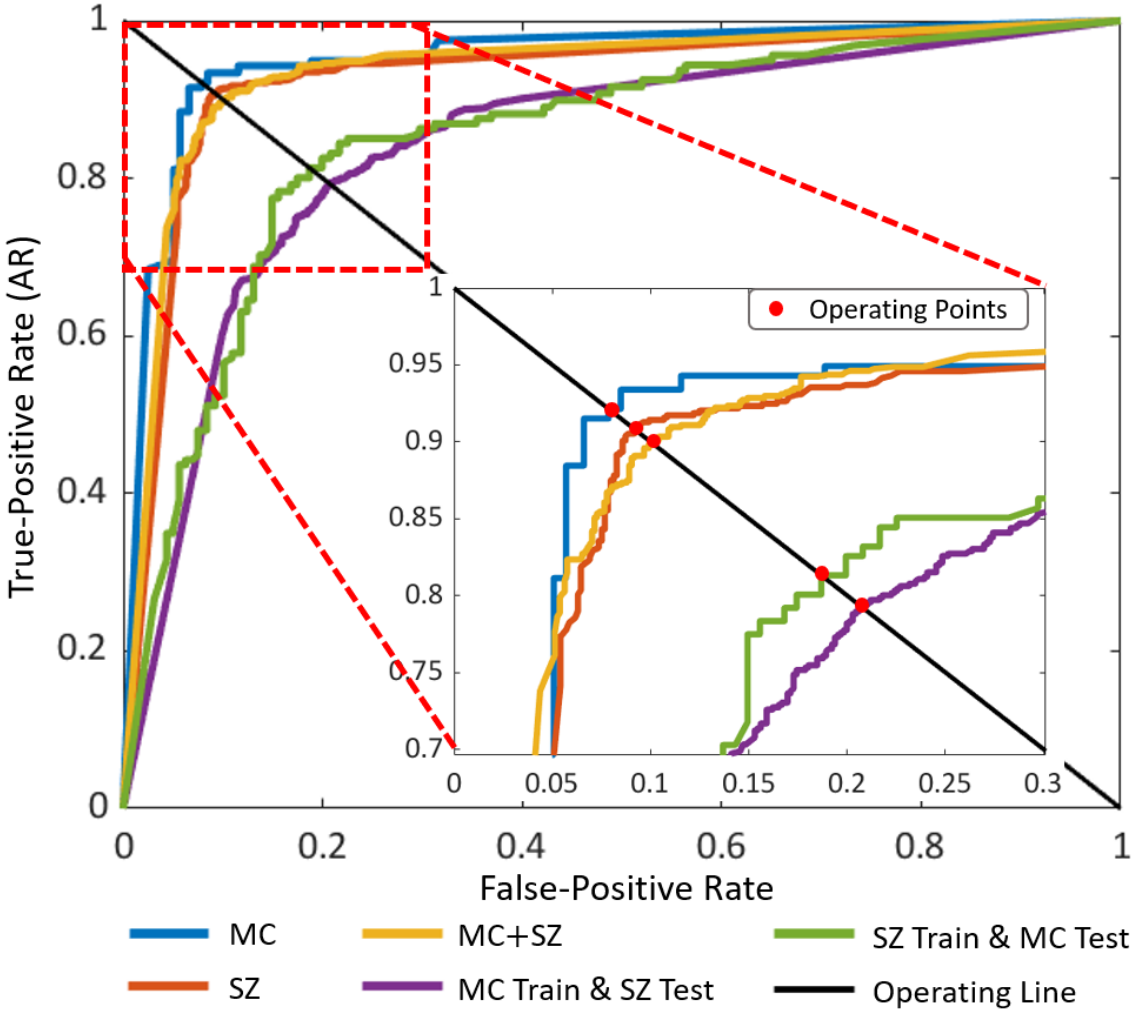
Table 3. Classification performance of our proposed ensemble-SDCNN[a] model including the submodels as an ablation study.

Data sets and models	F1	AP[b]	AR[c]	ACC[d]	AUC[e]
MC[f]					
SCNN[g,h]	0.765	0.775	0.757	0.769	0.817
DCNN[i,j]	0.88	0.888	0.872	0.878	0.932
ensemble-SDCNN	0.929	0.937	0.921	0.928	0.965
SZ[k]					
SCNN	0.802	0.803	0.802	0.802	0.868
DCNN	0.892	0.892	0.892	0.891	0.939
ensemble-SDCNN	0.908	0.909	0.908	0.908	0.948
MC + SZ					
SCNN	0.79	0.793	0.788	0.789	0.841
DCNN	0.891	0.892	0.89	0.89	0.943
ensemble-SDCNN	0.9	0.902	0.898	0.899	0.95
MC train and SZ test					
SCNN	0.557	0.559	0.555	0.557	0.541
DCNN	0.54	0.574	0.51	0.517	0.737
ensemble-SDCNN	0.795	0.798	0.793	0.792	0.853
SZ train and MC test					
SCNN	0.625	0.624	0.626	0.616	0.601
DCNN	0.7	0.702	0.698	0.71	0.754
ensemble-SDCNN	0.811	0.808	0.813	0.797	0.873

[a]SDCNN: shallow–deep CNN.
[b]AP: average precision.
[c]AR: average recall.
[d]ACC: accuracy.
[e]AUC: area under the curve.
[f]MC: Montgomery County.
[g]Ablation study performance by only considering SCNN for classification.
[h]SCNN: shallow CNN.
[i]Ablation study performance by only considering DCNN for classification.
[j]DCNN: deep CNN.
[k]SZ: Shenzhen.

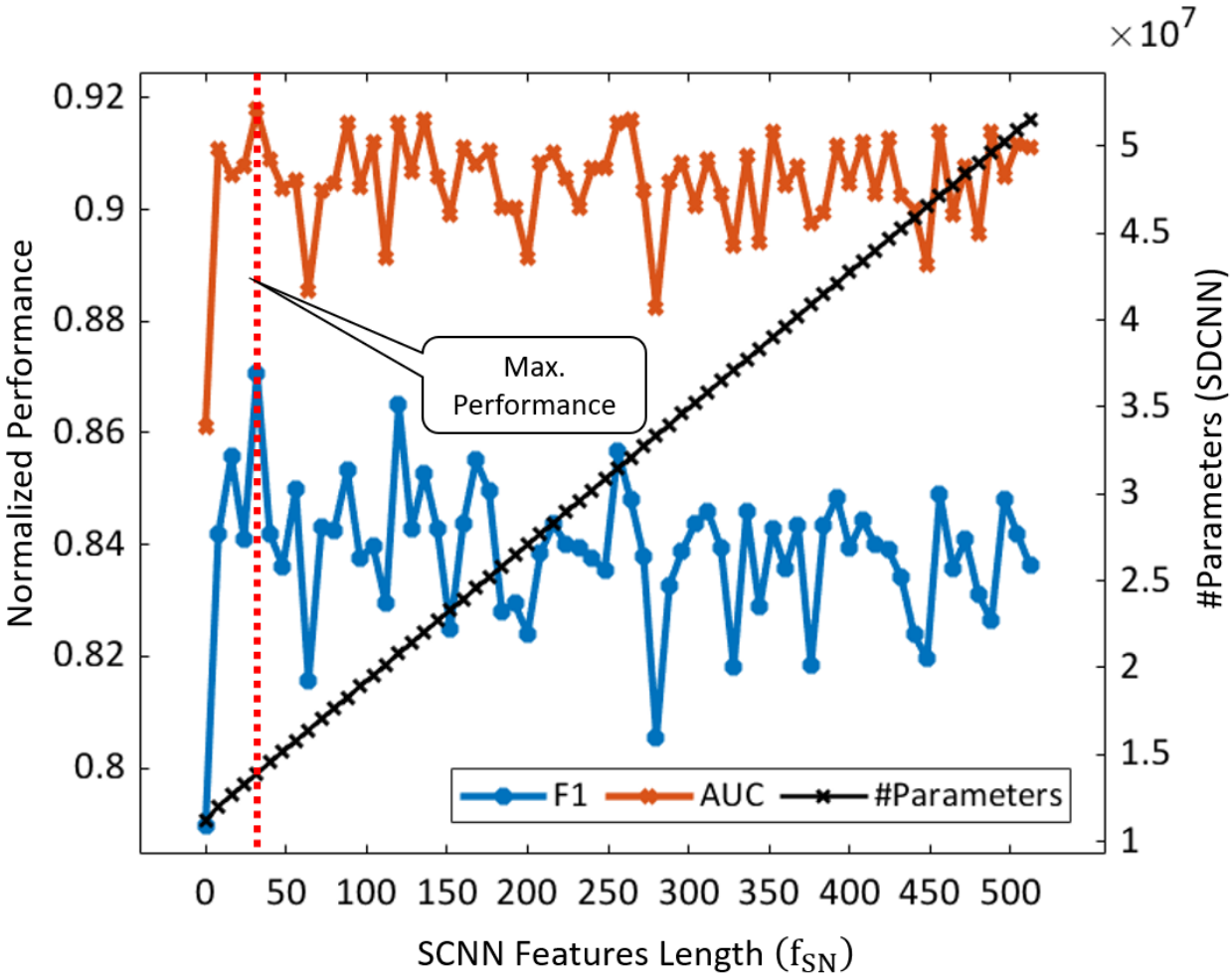
Figure 5 depicts the receiver operating characteristic curves of the proposed model for all the data sets. Each curve plots the TPR versus the FPR of our model at different classification thresholds beginning from 0 to 1 at 0.001 increments. Among all the classification thresholds, the optimal threshold was obtained based on the operating points (as highlighted with red closed circles) existing over the operating line. We attained the optimal threshold value of 0.507 for all the data sets. This implied that any CXR image with a classification probability larger than .507 was reported as a positive case. Finally, based on these receiver operating characteristic curves, we calculated the AUC results of our model for each data set (Table 3). We observed that the MC, SZ, and MC + SZ data sets had comparable AUCs of 0.965, 0.948, and 0.95, respectively. However, the performance of the cross–data set AUC was lower than that of the MC and SZ because of high intraclass and interclass variances between 2 different data sets, but the comparative performance (as reported in the subsequent section) of our model was still greater than the existing state-of-the-art methods for all the data sets.

Figure 5. Receiver operating characteristic curves of our ensemble-SDCNN model for all the datasets. Each curve plots true-positive rate (TPR) vs false-positive rate (FPR) of our model at different classification thresholds beginning from 0 to 1 in 0.001 increments. MC: Montgomery County; SDCNN: shallow–deep convolutional neural network; SZ: Shenzhen.



To determine the optimal ratio of the SCNN features with the DCNN, we performed several experiments for all the data sets by considering the different feature lengths of f_{SN} concatenated with f_{DN}. In this analysis, the feature lengths began from 0 to 512 with the increment of 8 features per experiment. Figure 6 shows the F1 and AUC results (average performance of all the data sets) according to different features length of f_{SN}. In addition, the black line depicts the growing number of the total parameters of our proposed model with the increasing length of f_{SN}. The figure indicates that our model exhibited the best performance (ie, maximum F1 of 0.871 and AUC of 0.918 as indicated by the vertical red line) and required the optimal number of total parameters as 1.39×10^7 for f_{SN}=32. Although the total number of trainable parameters of our model was slightly higher (approximately 2.7 million) than that of the DCNN, a substantial performance difference was observed, particularly for the cross data set (Table 3).

Figure 6. Average performance of the proposed ensemble-SDCNN model by considering different lengths of SCNN features with DCNN features (beginning from 0 to 512 with the increment of eight features in each experiment). AUC: area under the curve; DCNN: deep convolutional neural network; SDCNN: shallow–deep convolutional neural network; SCNN: shallow convolutional neural network.



In our classification-driven framework, both classification and retrieval performances were similar. However, we also evaluated the retrieval performance without performing the class prediction to validate the superiority of our classification-driven approach. In Table 4, the experimental results indicate that our classification-driven approach exhibited higher retrieval accuracies than the retrieval without class prediction. Moreover, our retrieval approach was computationally more efficient than that without class prediction as feature matching was performed using only the predicted class database rather than the entire database as in the retrieval without class prediction. In conclusion, these comparative results (Tables 3 and 4) implied that our jointly connected model exhibited superior performance in making the effective diagnostic decision and retrieving the best-matched cases from the previous database.

Table 4. Comparative retrieval performance with and without predicting the class label (CL).

Retrieval and data sets	F1	AP[a]	AR[b]	ACC[c]
Without class prediction				
MC[d]	0.844	0.861	0.828	0.847
SZ[e]	0.891	0.892	0.89	0.89
MC + SZ	0.88	0.882	0.878	0.879
MC train and SZ test	0.534	0.538	0.53	0.533
SZ train and MC test	0.729	0.737	0.72	0.739
With class prediction				
MC	0.929	0.937	0.921	0.928
SZ	0.908	0.909	0.908	0.908
MC + SZ	0.9	0.902	0.898	0.899
MC train and SZ test	0.795	0.798	0.793	0.792
SZ train and MC test	0.811	0.808	0.813	0.797

[a]AP: average precision.

[b]AR: average recall.

[c]ACC: accuracy.

[d]MC: Montgomery County.

[e]SZ: Shenzhen.

Comparative Analysis

Several CAD methods are presented in the literature for diagnosing pulmonary TB in CXR images. To make a fair comparison, we considered the following state-of-the-art methods [14,15,17,21,22,41,42], because these approaches selected the same data sets and experimental protocols as considered in our study. Moreover, in some recent studies [21], the authors adopted existing CNN models to classify the different types of pulmonary abnormalities including TB. However, these studies considered different data sets and experimental protocols. For a fair and detailed comparison, we evaluated the performance of these methods for our selected data sets and experimental protocol. Additionally, we calculated the performance of other CNN models [29,43-45] proposed for the general image-classification domain rather than radiology. The objective of this comparative analysis was to estimate the performance of the existing state-of-the-art CNN models in CXR image analyses. All these comparative analysis results are shown in Table 5.

Table 5. Comparative performance analysis of the proposed ensemble-SDCNN[a] model with various state-of-the-art methods.

Comparative methods	MC[b]					SZ[c]					MC + SZ				
	F1	AP[d]	AR[e]	ACC[f]	AUC[g]	F1	AP	AR	ACC	AUC	F1	AP	AR	ACC	AUC
LBP[h] and SVM[i,j] [46]	0.537	0.58	0.5	0.58	0.675	0.76	0.76	0.76	0.76	0.83	0.729	0.729	0.729	0.729	0.763
HoG[k] and SVM[i] [47]	0.797	0.796	0.798	0.797	0.863	0.85	0.85	0.85	0.85	0.90	0.822	0.823	0.821	0.821	0.882
ShuffleNet[i] [43]	0.747	0.771	0.727	0.748	0.84	0.875	0.876	0.873	0.873	0.937	0.884	0.885	0.883	0.884	0.936
InceptionV3[i] [44]	0.739	0.773	0.711	0.74	0.828	0.882	0.883	0.881	0.881	0.942	0.887	0.89	0.884	0.885	0.944
MobileNetV2[i] [45]	0.762	0.769	0.755	0.769	0.833	0.876	0.878	0.875	0.875	0.941	0.886	0.888	0.883	0.884	0.946
Santosh et al [41]	—[l]	—	—	0.79	0.88	—	—	—	0.86	0.93	—	—	—	—	—
Hwang et al [17]	—	—	—	0.674	0.884	—	—	—	0.837	0.926	—	—	—	—	—
ResNet50[i] [29]	0.788	0.796	0.78	0.79	0.886	0.877	0.877	0.877	0.876	0.94	0.88	0.881	0.878	0.879	0.921
ResNet101[i] [29]	0.8	0.821	0.782	0.798	0.895	0.864	0.865	0.862	0.861	0.934	0.859	0.862	0.857	0.858	0.923
Alfadhli et al [14]	—	0.81	0.79	0.791	0.89	—	—	—	—	—	—	—	—	—	—
GoogLeNet[i] [20,21]	0.834	0.851	0.818	0.834	0.902	0.852	0.853	0.851	0.851	0.921	0.843	0.846	0.84	0.84	0.914
Lopes and Valiati [21]	—	—	—	0.826	0.926	—	—	—	0.847	0.904	—	—	—	—	—
Vajda et al [42]	—	—	—	0.783	0.87	—	—	—	—	—	—	—	—	—	—
Pasa et al [22]	—	—	—	0.79	0.811	—	—	—	0.844	0.9	—	—	—	0.862	0.925
Govindarajan and Swaminathan [15]	0.876	—	0.877	0.878	0.94	—	—	—	—	—	—	—	—	—	—
Proposed	0.929	0.937	0.921	0.928	0.965	0.908	0.909	0.908	0.908	0.948	0.9	0.902	0.898	0.899	0.95

[a]SDCNN: shallow–deep CNN.

[b]MC: Montgomery County.

[c]SZ: Shenzhen.

[d]AP: average precision.

[e]AR: average recall.

[f]ACC: accuracy.

[g]AUC: area under the curve.

[h]LBP: local binary pattern.

[i]We evaluated the performance of these models using our selected data sets and experimental protocol.

[j]SVM: support vector machine.

[k]HoG: histogram of oriented gradients.

[l]—: not available. These results were not reported in some existing studies.

We observed that our method exhibited a superior performance (in terms of all the performance measures and data sets) compared with all the other baseline methods. In addition to deep learning–based methods, we evaluated and compared the performance of 2 known handcrafted feature-based methods [46,47]. To evaluate the performance of these 2 methods [46,47], we used the following default parameters as provided by the MATLAB framework [33]: size of histogram of oriented gradients cell as 8 × 8 with block size of 2 × 2 and number of overlapping cells between adjacent blocks as 1 block and the number of orientation bins as 9. In local binary patterns (LBPs) [46], the number of neighbor pixels considered was 8, with the linear interpolation method applied to compute pixel neighbors. Whereas in LBP histogram parameters, cell size was selected as 1 × 1 by applying L2-normalization to each LBP cell histogram. Thus, our comparative analysis was more detailed than the various existing studies [14,17,21,22]. For the MC data set, the performance gain of our model in contrast to Govindarajan and Swaminathan [15] (second-best) was greater than 4.4%, 5%, and 2.5% for AR, ACC, and AUC, respectively. Similarly, the difference in the performance of our model from a second-best model called InceptionV3 [44] (for the SZ data set) was more than 2.6%, 2.6%, 2.7%, 2.7%, and 0.6% for F1, AP, AR, ACC, and AUC, respectively. Moreover, for the combined data set (MC + SZ), the performance gain of our model in contrast to InceptionV3 [44] (second-best) was equal to 2.1%, 1.9%, 2.4%, 2.3%, and 0.4% for F1, AP, AR, ACC, and AUC, respectively. Hence, the performance of all these existing baseline methods validated the superiority of our proposed model with a substantial performance difference.

Moreover, comparative studies on the analysis of the cross–data set performance are rare. The majority of the studies only considered a similar data set for training and testing. Cross–data

set testing is an important analysis to demonstrate the general capability of a model and its potential applicability in a real-world environment. Therefore, similar comparative results are also evaluated (in a cross data set) for different baseline models for a detailed performance comparison with the proposed ensemble-SDCNN model. In this analysis, the MC data set was used to train the model and SZ was used to test, and vice versa. Table 6 shows the results of these cross–data set analyses along with comparative studies.

These comparative results indicated that our model had outperformed the various deep learning and handcrafted feature-based TB diagnostic methods. For the SZ data set, which was used for training, the accuracies were slightly higher than those for the MC data set. The main reason for this was the presence of more training data samples compared with the MC data set. For the scenario in which the MC data set was the training set and the SZ the testing set, the performance of our model in contrast to that of Santosh and Antani [16] (second best) was higher than 3.3%, 3.2%, and 3.3% for AR, ACC, and AUC, respectively, and the comparative performance difference of our model with that of Santosh and Antani [16] (for SZ as training and MC as testing data sets) was also higher than 2.3%, 1.7%, and 2.3% for AR, ACC, and AUC, respectively. All these experimental results highlighted the potential applicability of our model in real-world diagnostics related to chest abnormalities.

Table 6. Results of comparative performance analysis of our proposed method with various baseline methods for cross data sets.

Data sets and our methods	F1	AP[a]	AR[b]	ACC[c]	AUC[d]
MC[e] train and SZ[f] test					
LBP[g] and SVM[h,i] [46]	0.496	0.492	0.5	0.492	0.69
HoG[j] and SVM[i] [47]	0.664	0.695	0.635	0.639	0.762
ShuffleNet[i] [43]	0.661	0.715	0.615	0.61	0.709
InceptionV3[i] [44]	0.708	0.717	0.7	0.698	0.761
MobileNetV2[i] [45]	0.613	0.678	0.559	0.565	0.78
ResNet50[i] [29]	0.686	0.707	0.667	0.663	0.77
ResNet101[i] [29]	0.674	0.677	0.671	0.672	0.772
GoogLeNet[i] [20,21]	0.592	0.595	0.589	0.591	0.65
Santosh and Antani [16]	—[k]	—	0.76	0.76	0.82
Proposed	0.795	0.798	0.793	0.792	0.853
SZ train and MC test					
LBP and SVM[i] [46]	0.537	0.58	0.5	0.58	0.552
HoG and SVM[i] [47]	0.559	0.573	0.546	0.594	0.601
ShuffleNet[i] [43]	0.633	0.643	0.624	0.652	0.683
InceptionV3[i] [44]	0.681	0.722	0.644	0.688	0.748
MobileNetV2[i] [45]	0.668	0.772	0.589	0.652	0.797
ResNet50[i] [29]	0.64	0.642	0.638	0.616	0.787
ResNet101[i] [29]	0.641	0.726	0.574	0.638	0.698
GoogLeNet[i] [20,21]	0.648	0.691	0.609	0.659	0.754
Santosh and Antani [16]	—	—	0.79	0.78	0.85
Proposed	0.811	0.808	0.813	0.797	0.873

[a]AP: average precision.

[b]AR: average recall.

[c]ACC: accuracy.

[d]AUC: area under the curve.

[e]MC: Montgomery County.

[f]SZ: Shenzhen.

[g]LBP: local binary pattern.

[h]SVM: support vector machine.

[i]We also evaluated the performance of these models (for the cross data set) using our selected data sets and experimental protocol.

[j]HoG: histogram of oriented gradients.

[k]—: not available. The results were not provided in this comparative study for these performance metrics.

Discussion

This article presents an interactive CAD framework based on multiscale information fusion to diagnose TB in CXR images and retrieve the relevant cases (CXR images) from a previous patients' database including clinical observations. In this framework, a classification model is primarily proposed to classify the given CXR image as either a positive or a negative sample. Subsequently, classification-based retrieval is performed to retrieve the relevant cases and corresponding clinical readings based on our newly proposed MLSM algorithm. The proposed model substantially improves diagnostic performance by performing the fusion of both low- and high-level features. The network processes the input image through different layers and finally activates the class-specific discriminative region [48] as key-features maps. Figure 7 shows such activation maps extracted from the 7 different layers (ie, F_{SN1}, F_{SN2}, F_{DN1}, F_{DN2}, F_{DN3}, F_{DN4}, and F_{DN5} as labeled in Figure 2) of our model for

both positive and negative sample images. As Figure 7 shows, each activation map is generated by calculating the average of all the extracted maps from a specific location. All the activation maps overlay on their corresponding input image after resizing and applying a pseudo-color scheme (blue to red, equivalent to lower to higher activated region) to produce a better visualization of the activated regions.

Figure 7. Extracted features maps from the different parts of the proposed ensemble-SDCNN model for both TB positive and negative cases. DCNN: deep convolutional neural network; SDCNN: shallow–deep convolutional neural network; SCNN: shallow convolutional neural network; TB: tuberculosis.

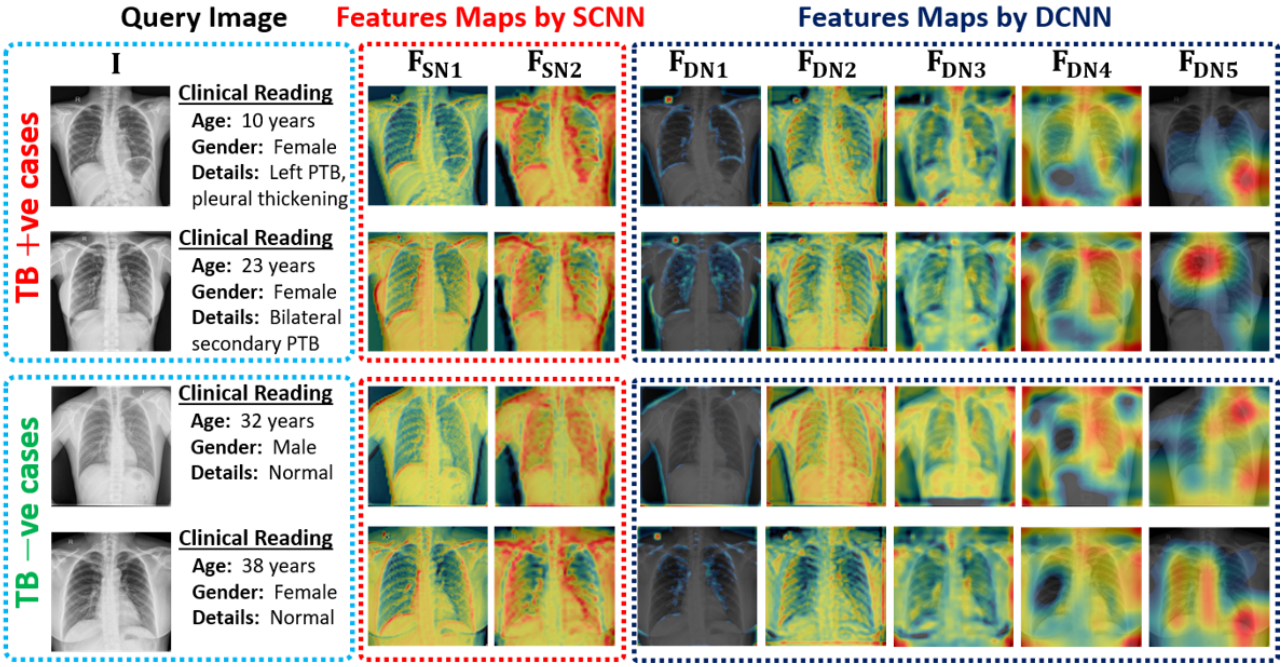
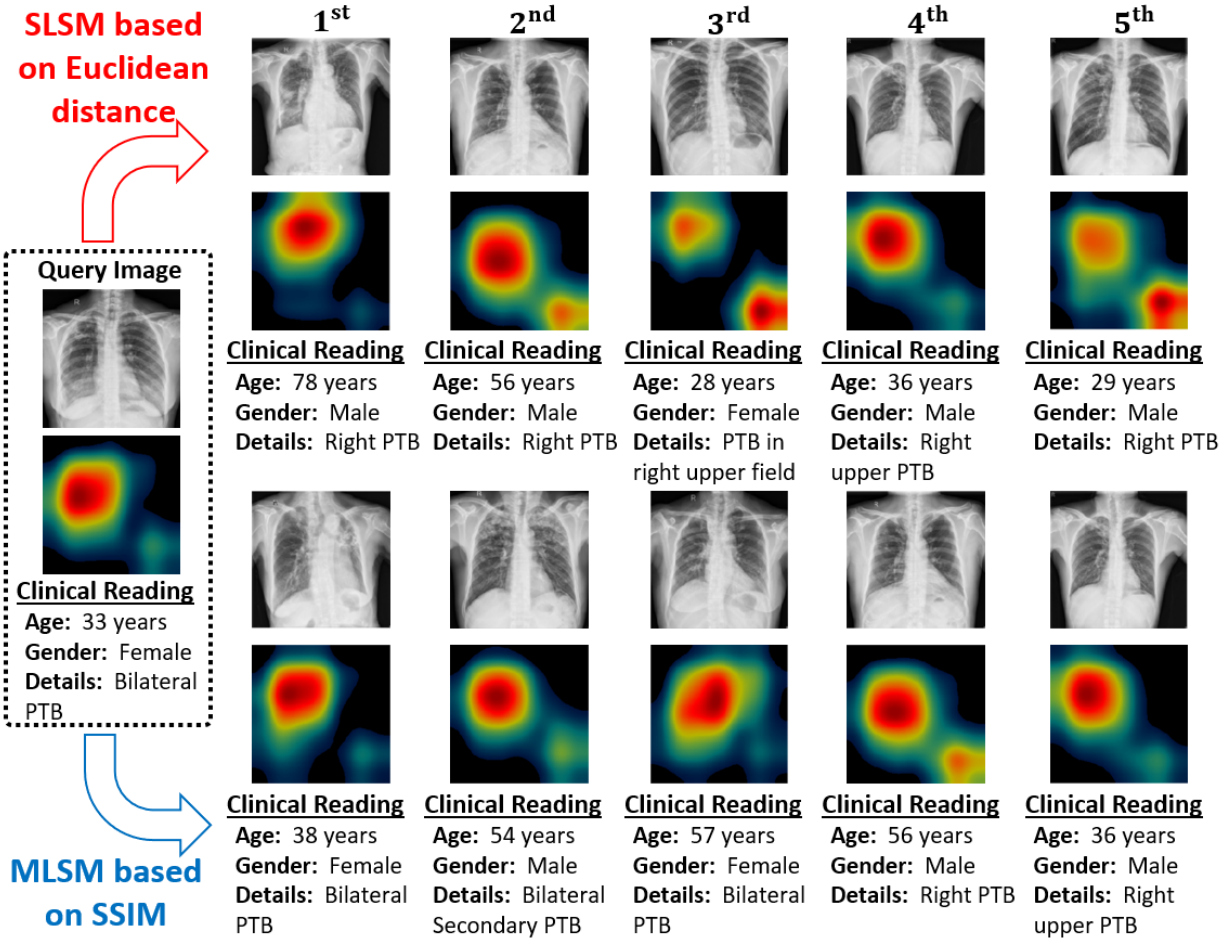


Figure 7 indicates that the class-specific discriminative regions of the given input image become more prominent after processing through the successive layers of the network. A semilocalized activation map (labeled as F_{DN5} in Figure 7) is obtained from the last convolutional layer of the DCNN model, which includes the more distinctive high-level features for each class. Moreover, for the SCNN, the obtained activation map from the last convolutional layer (labeled as F_{SN2} in Figure 7) encompasses the low-level features such as edge information. Finally, both low- and high-level features are used in making an effective diagnostic decision for the given CXR image. The experimental results (also provided in Multimedia Appendix 2) proved that the diagnostic performance of our ensemble-SDCNN model is more effective than the various CNN models where only single-level features are used for class prediction.

After an effective diagnostic decision, we can further retrieve the relevant cases based on our proposed MLSM algorithm, which considers the multilevel features in retrieving the best matches. Figure 8 depicts the retrieval results of our proposed MLSM algorithm in comparison with the conventional Euclidean distance–based SLSM scheme. In Figure 8, these results comprise the 5 best-matched CXR images along with their corresponding high-level activation maps (labeled as F_{DN5} in Figure 7) and clinical readings. Generally, a high correlation between the high-level activation maps (as F_{DN5} in our study) of the query image and retrieved image implies the optimal performance of a retrieval system. With our MLSM algorithm, these activation maps (corresponding to retrieved cases) were more analogous (in terms of shape and location) to that of query image compared with the conventional SLSM scheme. This implied that our algorithm retrieved the highly correlated cases in terms of TB patterns, location, and clinical observation.

Figure 8. Visualization of retrieval performance for the given input query image by considering SLSM and MLSM (our proposed model). MLSM: multilevel similarity measure; SLSM: single-level similarity measure.



In addition, we evaluated the objective similarity score in terms of the PSNR between the activation maps of the input query and 20 best-matched cases for both algorithms (MLSM and SLSM). The main purpose of this analysis was to quantitatively evaluate such feature-level similarities of both algorithms. A total of 28 images (28/138, 20.2% of the MC data set) from the MC data set and 132 images (132/662, 19.9% of the SZ data set) from the SZ data set were selected as the query database to perform this analysis. Using each query image one at a time, we retrieved the 20 best-matched cases corresponding to each algorithm. Thus, 20 different PSNR values were computed corresponding to these retrieved images for each matching algorithm. After these results for the entire selected query database were evaluated, an average PSNR performance was calculated to present the average performance of a single query image for each algorithm. Figure 9 shows the comparative performance results of our proposed MLSM algorithm and the conventional SLSM scheme. We observed that our matching algorithm exhibited the higher features-level similarity scores in terms of the PSNR (for all the retrieved images and both data sets) in contrast to the SLSM scheme. Thus, our algorithm resulted in an optimal retrieval performance because of the significant correlation of high-level activation maps. All these results (Figures 8 and 9) were computed based on our selected classification-driven retrieval method. The experimental results provided in Table 4 have already proved that our selected class prediction–based retrieval method outperforms the retrieval method without class prediction.

In addition to the numerical results provided in Table 4, Figure 10 further distinguishes the retrieved results of these 2 different approaches (ie, with and without class prediction) figuratively. Figure 10 indicates that all the retrieved cases (for the given query image) were TPs in our class prediction–based retrieval method.

However, in the retrieval without class prediction, the first and third best matches were FPs (highlighted by the red bounding box) while the remaining three cases were TPs. Such FP cases may lead to a vague diagnostic decision. Additionally, the numerical results (Table 4) indicated that the average number of FPs in retrieval without class prediction was substantially higher than our class-prediction retrieval method. Therefore, in this study, we considered a classification-driven retrieval by performing the class prediction in the first step and then retrieving the best-matched cases from the predicted class database rather than exploring the entire database. Ultimately, the classification results can aid in making a diagnostic decision and the retrieved CXR images can assist radiologists to further validate the computer decision. Furthermore, if the wrong prediction is made by the computer, the medical expert can check other relevant cases (ie, second-, third-, or fourth-best matches) that can be more relevant than the first best match.

Thus, both classification and retrieval results can aid radiologists in making an effective diagnostic decision even in scenarios of small TB patterns that remain undetectable in the early stage. Such a comprehensive CAD framework may assist radiologists in clinical practices and alleviate the burden of an increasing number of patients by providing an effective and timely diagnostic decision. Our trained model and the training and testing data splitting information are publicly available [49] to enable other researchers to evaluate and compare its performance.

Figure 9. PSNR-based objective similarity measures between the high-level activation maps of the query image and retrieved images to evaluate feature-level similarities of both algorithms (ie, MLSM and SLSM). MLSM: multilevel similarity measure; PSNR: peak signal-to-noise ratio; SLSM: single-level similarity measure.

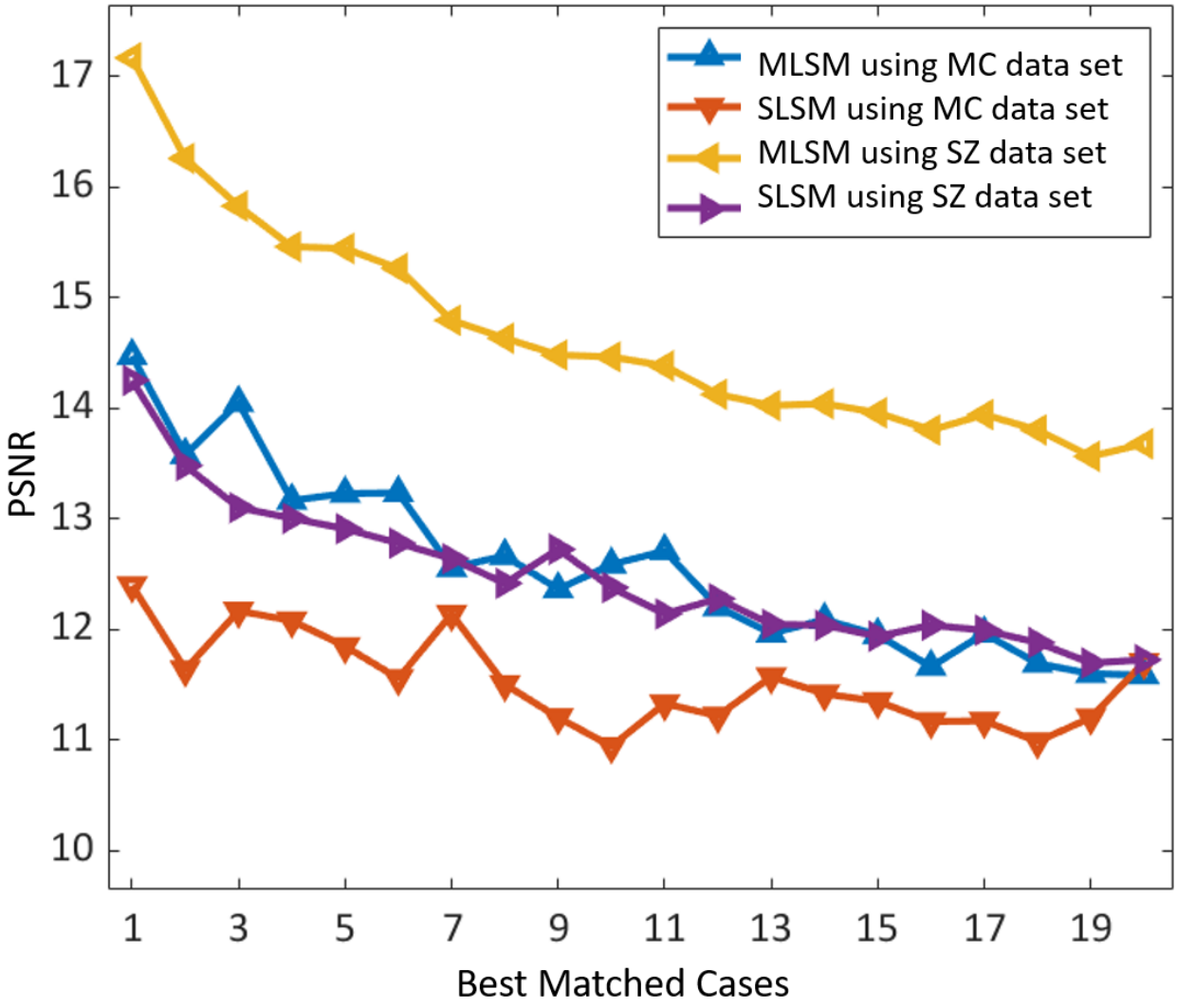
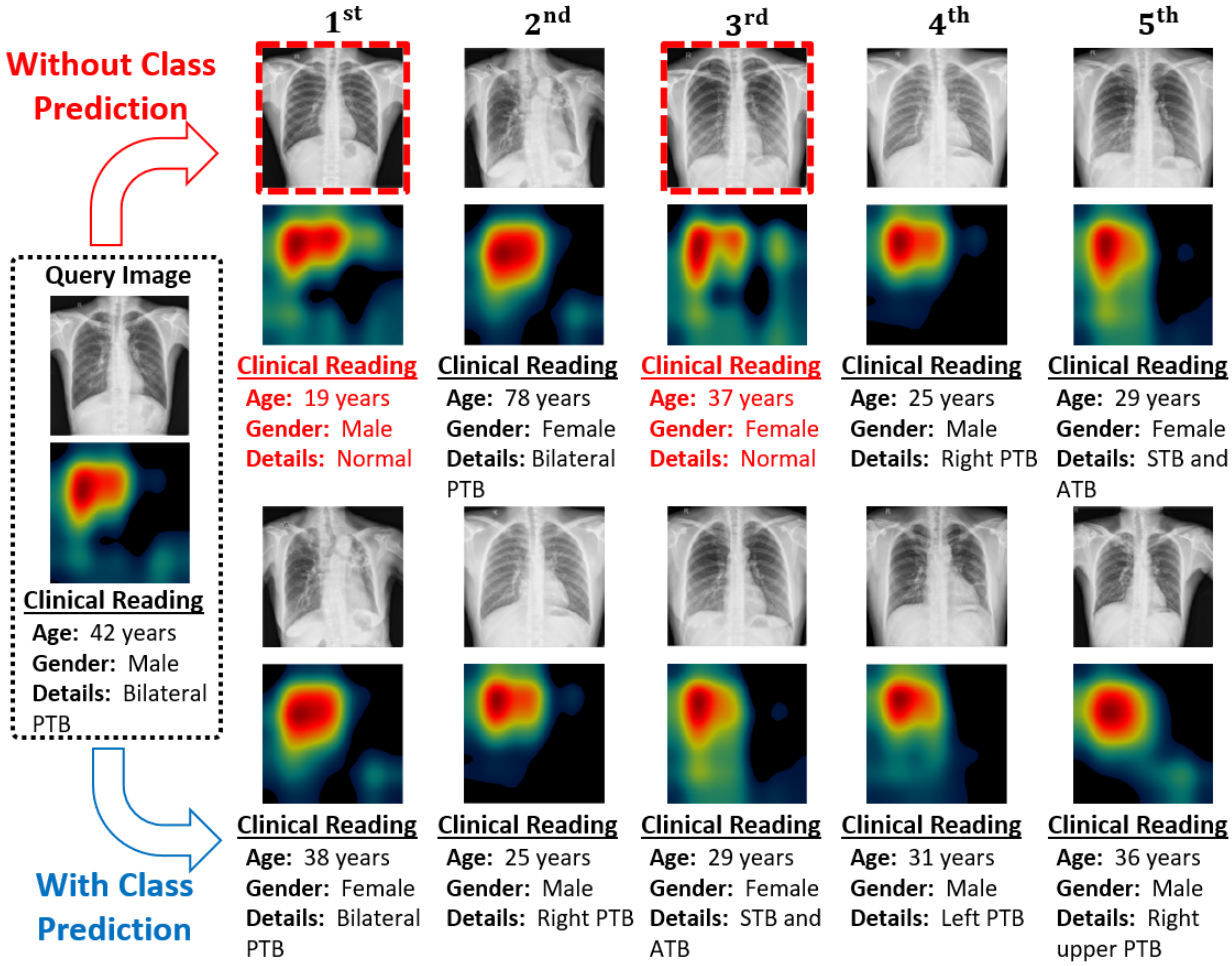


Figure 10. Visualization of retrieval performance for the given input query image by considering both retrieval methods with class prediction and without class prediction.



Acknowledgments

This work was supported in part by the Ministry of Science and ICT (MSIT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2020-2020-0-01789) supervised by the IITP (Institute for Information & Communications Technology Promotion) and in part by the Bio and Medical Technology Development Program of the National Research Foundation of Korea (NRF) funded by the Korean government, the MSIT (NRF-2016M3A9E1915855).

Authors' Contributions

MO and KP designed the overall framework. Moreover, they wrote and revised the complete paper. MA, TM, and YK facilitated in designing comparative analysis and experiments.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Other supplementary material is provided in the attached word file [DOCX file (MS Word), 44 KB].
[DOCX File , 44 KB - medinform_v8i12e21790_app1.docx]

Multimedia Appendix 2
All the experimental results are provided in the attached excel file [XLSX file (MS Excel), 226 KB].
[XLSX File (Microsoft Excel File), 226 KB - medinform_v8i12e21790_app2.xlsx]

References

1. Global tuberculosis report. World Health Organization. Geneva, Switzerland: World Health Organization; 2015. URL: http://www.tbonline.info/posts/2015/10/28/global-tuberculosis-report-2015/ [accessed 2020-11-26]
2. Cheon SA, Cho HH, Kim J, Lee J, Kim HJ, Park TJ. Recent tuberculosis diagnosis toward the end TB strategy. J Microbiol Methods 2016 Apr;123:51-61. [doi: 10.1016/j.mimet.2016.02.007] [Medline: 26853124]
3. Casela M, Cerqueira SMA, Casela TDO, Pereira MA, Santos SQD, Pozo FAD, et al. Rapid molecular test for tuberculosis: impact of its routine use at a referral hospital. J Bras Pneumol 2018 Apr;44(2):112-117 [FREE Full text] [doi: 10.1590/s1806-37562017000000201] [Medline: 29791546]
4. Panteix G, Gutierrez MC, Boschiroli ML, Rouviere M, Plaidy A, Pressac D, et al. Pulmonary tuberculosis due to Mycobacterium microti: a study of six recent cases in France. J Med Microbiol 2010 Aug;59(8):984-989. [doi: 10.1099/jmm.0.019372-0] [Medline: 20488936]
5. Schaefer-Prokop C, Neitzel U, Venema HW, Uffmann M, Prokop M. Digital chest radiography: an update on modern technology, dose containment and control of image quality. Eur Radiol 2008 Sep;18(9):1818-1830 [FREE Full text] [doi: 10.1007/s00330-008-0948-3] [Medline: 18431577]
6. Lee Y, Raviglione MC, Flahault A. Use of Digital Technology to Enhance Tuberculosis Control: Scoping Review. J Med Internet Res 2020 Feb 13;22(2):e15727. [doi: 10.2196/15727]
7. Gardezi SJS, Elazab A, Lei B, Wang T. Breast Cancer Detection and Diagnosis Using Mammographic Data: Systematic Review. J Med Internet Res 2019 Jul 26;21(7):e14464 [FREE Full text] [doi: 10.2196/14464] [Medline: 31350843]
8. Nielsen M. Neural Networks and Deep Learning. San Francisco, CA: Determination Press; 2015.
9. Jaeger S, Karargyris A, Candemir S, Folio L, Siegelman J, Callaghan F, et al. Automatic Tuberculosis Screening Using Chest Radiographs. IEEE Trans. Med. Imaging 2014 Feb;33(2):233-245. [doi: 10.1109/tmi.2013.2284099]
10. Kumar A, Wang YY, Liu KC, Tsai IC, Huang CC, Hung N. Distinguishing normal and pulmonary edema chest x-ray using Gabor filter and SVM. 2014 Presented at: IEEE International Symposium on Bioelectronics and Bioinformatics (IEEE ISBB); 3-6 April 2014; Chung Li, Taiwan p. 117-120. [doi: 10.1109/isbb.2014.6820918]
11. Hogeweg L, Sanchez CI, Maduskar P, Philipsen R, Story A, Dawson R, et al. Automatic Detection of Tuberculosis in Chest Radiographs Using a Combination of Textural, Focal, and Shape Abnormality Analysis. IEEE Trans. Med. Imaging 2015 Dec;34(12):2429-2442. [doi: 10.1109/tmi.2015.2405761]
12. Carrillo-de-Gea JM, García-Mateos G, Fernández-Alemán JL, Hernández-Hernández JL. A Computer-Aided Detection System for Digital Chest Radiographs. J Healthc Eng 2016;2016(1):8208923 [FREE Full text] [doi: 10.1155/2016/8208923] [Medline: 27372536]
13. Karargyris A, Siegelman J, Tzortzis D, Jaeger S, Candemir S, Xue Z, et al. Combination of texture and shape features to detect pulmonary abnormalities in digital chest X-rays. Int J Comput Assist Radiol Surg 2016 Jan 20;11(1):99-106. [doi: 10.1007/s11548-015-1242-x] [Medline: 26092662]
14. Alfadhli FHO, Mand AA, Sayeed MD, Sim KS, Al-Shabi M. Classification of tuberculosis with SURF spatial pyramid features. 2017 Presented at: IEEE International Conference on Robotics, Automation and Sciences (ICORAS); 27-29 November 2017; Melaka, Malaysia. [doi: 10.1109/icoras.2017.8308044]
15. Govindarajan S, Swaminathan R. Analysis of Tuberculosis in Chest Radiographs for Computerized Diagnosis using Bag of Keypoint Features. J Med Syst 2019 Mar 28;43(4):87. [doi: 10.1007/s10916-019-1222-8] [Medline: 30820678]
16. Santosh KC, Antani S. Automated Chest X-Ray Screening: Can Lung Region Symmetry Help Detect Pulmonary Abnormalities? IEEE Trans. Med. Imaging 2018 May;37(5):1168-1177. [doi: 10.1109/tmi.2017.2775636]
17. Hwang S, Kim HE, Jeong J, Kim HJ. A novel approach for tuberculosis screening based on deep convolutional neural networks. In: Proceedings of SPIE 9785, Medical Imaging 2016: Computer-Aided Diagnosis. 2016 Presented at: SPIE 9785, Medical Imaging 2016: Computer-Aided Diagnosis; 24 March 2016; San Diego, CA, USA. [doi: 10.1117/12.2216198]
18. Shin HC, Roberts K, Lu L, Demner-Fushman D, Yao J, Summers RM. Learning to read chest x-rays: recurrent neural cascade model for automated image annotation. 2016 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; 27-30 June 2016; Las Vegas, NV, USA p. 2497-2506. [doi: 10.1109/cvpr.2016.274]
19. Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology 2017 Aug;284(2):574-582. [doi: 10.1148/radiol.2017162326] [Medline: 28436741]
20. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. 2015 Presented at: IEEE Conference Computer Vision Pattern Recognition; 7-12 June 2015; Boston, MA, USA p. 1-9. [doi: 10.1109/cvpr.2015.7298594]
21. Lopes UK, Valiati JF. Pre-trained convolutional neural networks as feature extractors for tuberculosis detection. Computers in Biology and Medicine 2017 Oct;89(1):135-143. [doi: 10.1016/j.compbiomed.2017.08.001]
22. Pasa F, Golkov V, Pfeiffer F, Cremers D, Pfeiffer D. Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. Sci Rep 2019 Apr 18;9(1):6268 [FREE Full text] [doi: 10.1038/s41598-019-42557-4] [Medline: 31000728]
23. Qin ZZ, Sander MS, Rai B, Titahong CN, Sudrungrot S, Laah SN, et al. Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. Sci Rep 2019 Oct 18;9(1):15000 [FREE Full text] [doi: 10.1038/s41598-019-51503-3] [Medline: 31628424]

24. Nash M, Kadavigere R, Andrade J, Sukumar CA, Chawla K, Shenoy VP, et al. Deep learning, computer-aided radiography reading for tuberculosis: a diagnostic accuracy study from a tertiary hospital in India. Sci Rep 2020 Jan 14;10(1):210 [FREE Full text] [doi: 10.1038/s41598-019-56589-3] [Medline: 31937802]
25. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. 2012 Presented at: 25th International Conference on Neural Information Processing Systems; 3-6 December 2012; Lake Tahoe, NV, USA p. 1097-1105. [doi: 10.1145/3065386]
26. Triantafyllidis AK, Tsanas A. Applications of Machine Learning in Real-Life Digital Health Interventions: Review of the Literature. J Med Internet Res 2019 Apr 05;21(4):e12286 [FREE Full text] [doi: 10.2196/12286] [Medline: 30950797]
27. Chen S, Wu S. Identifying Lung Cancer Risk Factors in the Elderly Using Deep Neural Networks: Quantitative Analysis of Web-Based Survey Data. J Med Internet Res 2020 Mar 17;22(3):e17695 [FREE Full text] [doi: 10.2196/17695] [Medline: 32181751]
28. Candemir S, Jaeger S, Palaniappan K, Musco JP, Singh RK, Zhiyun X, et al. Lung Segmentation in Chest Radiographs Using Anatomical Atlases With Nonrigid Registration. IEEE Trans. Med. Imaging 2014 Feb;33(2):577-590. [doi: 10.1109/tmi.2013.2290491]
29. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 Presented at: IEEE Conference on Computer Vision Pattern Recognition; 26 June-01 July 2016; Las Vegas, NV, USA p. 770-778. [doi: 10.1109/cvpr.2016.90]
30. He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. 2016 Presented at: European Conference on Computer Vision; 8-16 October 2016; Amsterdam, The Netherlands p. 630-645. [doi: 10.1007/978-3-319-46493-0_38]
31. Gao F, Wu T, Li J, Zheng B, Ruan L, Shang D, et al. SD-CNN: A shallow-deep CNN for improved breast cancer diagnosis. Comput Med Imaging Graph 2018 Dec;70:53-62. [doi: 10.1016/j.compmedimag.2018.09.004] [Medline: 30292910]
32. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 2004 Apr;13(4):600-612. [doi: 10.1109/tip.2003.819861] [Medline: 15376593]
33. Deep Learning Toolbox. URL: https://in.mathworks.com/products/deeplearning.html [accessed 2020-05-01]
34. Intel® Core i7-3770K Processor. URL: https://ark.intel.com/content/www/us/en/ark/products/65523/intel-core-i7-3770k-processor-8m-cache-up-to-3-90-ghz.html [accessed 2020-05-01]
35. GeForce GTX 1070. URL: https://www.geforce.com/hardware/desktop-gpus/geforce-gtx1070/specifications [accessed 2020-05-01]
36. Li XL. Preconditioned Stochastic Gradient Descent. IEEE Trans. Neural Netw. Learning Syst 2018 May;29(5):1454-1466. [doi: 10.1109/tnnls.2017.2672978]
37. M H, M.n S. A Review on Evaluation Metrics for Data Classification Evaluations. IJDKP 2015 Mar 31;5(2):01-11. [doi: 10.5121/ijdkp.2015.5201]
38. Livingston EH. Who was student and why do we care so much about his t-test? J Surg Res 2004 May 01;118(1):58-65. [doi: 10.1016/j.jss.2004.02.003] [Medline: 15093718]
39. Cohen J. A power primer. Psychological Bulletin 1992;112(1):155-159. [doi: 10.1037/0033-2909.112.1.155]
40. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev Camb Philos Soc 2007 Nov;82(4):591-605. [doi: 10.1111/j.1469-185X.2007.00027.x] [Medline: 17944619]
41. Santosh KC, Vajda S, Antani S, Thoma GR. Edge map analysis in chest X-rays for automatic pulmonary abnormality screening. Int J Comput Assist Radiol Surg 2016 Sep 19;11(9):1637-1646. [doi: 10.1007/s11548-016-1359-6] [Medline: 26995600]
42. Vajda S, Karargyris A, Jaeger S, Santosh KC, Candemir S, Xue Z, et al. Feature Selection for Automatic Tuberculosis Screening in Frontal Chest Radiographs. J Med Syst 2018 Jun 29;42(8):146. [doi: 10.1007/s10916-018-0991-9] [Medline: 29959539]
43. Zhang X, Zhou X, Lin M, Sun J. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. 2018 Presented at: IEEE/CVF Conference on Computer Vision Pattern Recognition; 18-23 June 2018; Salt Lake City, UT, USA p. 6848-6856. [doi: 10.1109/cvpr.2018.00716]
44. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. 2016 Presented at: IEEE Conference on Computer Vision Pattern Recognition; 27-30 June 2016; Las Vegas, NV, USA p. 2818-2826. [doi: 10.1109/cvpr.2016.308]
45. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. Mobilenetv2: Inverted residualslinear bottlenecks. 2018 Presented at: IEEE/CVF Conference on Computer Vision Pattern Recognition; 18-23 June 2018; Salt Lake City, UT, USA p. 4510-4520. [doi: 10.1109/cvpr.2018.00474]
46. Subrahmanyam M, Maheshwari RP, Balasubramanian R. Local maximum edge binary patterns: A new descriptor for image retrieval and object tracking. Signal Processing 2012 Jun;92(6):1467-1479. [doi: 10.1016/j.sigpro.2011.12.005]
47. Velmurugan K, Baboo SS. Image Retrieval using Harris Corners and Histogram of Oriented Gradients. IJCA 2011 Jun 30;24(7):6-10. [doi: 10.5120/2968-3968]
48. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. 2016 Presented at: IEEE Conference on Computer Vision Pattern Recognition; 27-30 June 2016; Las Vegas, NV, USA p. 2921-2929. [doi: 10.1109/cvpr.2016.319]
49. Dongguk CAD framework for effective diagnosis of tuberculosis. URL: http://dm.dgu.edu/link.html [accessed 2020-05-01]

Abbreviations

ACC: accuracy
AP: average precision
AR: average recall
AUC: area under the curve
CAD: computer-aided diagnosis
CL: class label
CNN: convolutional neural network
CXR: chest radiograph
DCNN: deep convolutional neural network
FN: false negatives
FP: false positives
FPR: false-positive rate
F1: F1 score
HoG: histogram of oriented gradients
LBP: local binary pattern
MC: Montgomery County
MLSM: multilevel similarity measure
PSNR: peak signal-to-noise ratio
ROC: receiver operating characteristic (curve)
SDCNN: shallow–deep convolutional neural network
SCNN: shallow convolutional neural network
SLSM: single-level similarity measure
SSIM: structure similarity
SVM: support vector machine.
SZ: Shenzhen
TB: tuberculosis
TN: true negative
TP: true positive
TPR: true-positive rate
WHO: World Health Organization

Edited by G Eysenbach; submitted 25.06.20; peer-reviewed by W Sun, MO Kaya; comments to author 03.11.20; revised version received 05.11.20; accepted 09.11.20; published 07.12.20.

Please cite as:
Owais M, Arsalan M, Mahmood T, Kim YH, Park KR
Comprehensive Computer-Aided Decision Support Framework to Diagnose Tuberculosis From Chest X-Ray Images: Data Mining Study
JMIR Med Inform 2020;8(12):e21790
URL: http://medinform.jmir.org/2020/12/e21790/
doi:10.2196/21790
PMID:33284119

Original Paper

Cystic Fibrosis Point of Personalized Detection (CFPOPD): An Interactive Web Application

Christopher Wolfe[1*], MSc; Teresa Pestian[1*], BSc; Emrah Gecili[1], PhD; Weiji Su[1,2], PhD; Ruth H Keogh[3], DPhil; John P Pestian[4,5], PhD; Michael Seid[4,6,7], PhD; Peter J Diggle[8,9], PhD; Assem Ziady[4,6], PhD; John Paul Clancy[4,6,10], MD; Daniel H Grossoehme[11,12,13], MS, DMin; Rhonda D Szczesniak[1,4,6*], PhD; Cole Brokamp[1,4*], PhD

[1]Division of Biostatistics & Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States
[2]Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH, United States
[3]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, United Kingdom
[4]Department of Pediatrics, University of Cincinnati, Cincinnati, OH, United States
[5]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States
[6]Division of Pulmonary Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States
[7]James M Anderson Center for Health Systems Excellence, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States
[8]Centre for Health Informatics, Computing, and Statistics, Lancaster Medical School, Lancaster University, Lancaster, United Kingdom
[9]Health Data Research UK, London, United Kingdom
[10]Cystic Fibrosis Foundation, Bethesda, MD, United States
[11]Haslinger Family Pediatric Palliative Care Center, Akron Children's Hospital, Akron, OH, United States
[12]Rebecca D Considine Research Institute, Akron Children's Hospital, Akron, OH, United States
[13]Division of Family & Community Medicine, Akron Children's Hospital, Akron, OH, United States
[*]these authors contributed equally

Corresponding Author:
Rhonda D Szczesniak, PhD
Department of Pediatrics
University of Cincinnati
3333 Burnet Avenue
Cincinnati, OH, 45229
United States
Phone: 1 513 803 0563
Email: rhonda.szczesniak@cchmc.org

Abstract

Background: Despite steady gains in life expectancy, individuals with cystic fibrosis (CF) lung disease still experience rapid pulmonary decline throughout their clinical course, which can ultimately end in respiratory failure. Point-of-care tools for accurate and timely information regarding the risk of rapid decline is essential for clinical decision support.

Objective: This study aims to translate a novel algorithm for earlier, more accurate prediction of rapid lung function decline in patients with CF into an interactive web-based application that can be integrated within electronic health record systems, via collaborative development with clinicians.

Methods: Longitudinal clinical history, lung function measurements, and time-invariant characteristics were obtained for 30,879 patients with CF who were followed in the US Cystic Fibrosis Foundation Patient Registry (2003-2015). We iteratively developed the application using the R Shiny framework and by conducting a qualitative study with care provider focus groups (N=17).

Results: A clinical conceptual model and 4 themes were identified through coded feedback from application users: (1) ambiguity in rapid decline, (2) clinical utility, (3) clinical significance, and (4) specific suggested revisions. These themes were used to revise our application to the currently released version, available online for exploration. This study has advanced the application's potential prognostic utility for monitoring individuals with CF lung disease. Further application development will incorporate additional clinical characteristics requested by the users and also a more modular layout that can be useful for care provider and family interactions.

Conclusions: Our framework for creating an interactive and visual analytics platform enables generalized development of applications to synthesize, model, and translate electronic health data, thereby enhancing clinical decision support and improving care and health outcomes for chronic diseases and disorders. A prospective implementation study is necessary to evaluate this tool's effectiveness regarding increased communication, enhanced shared decision-making, and improved clinical outcomes for patients with CF.

(JMIR Med Inform 2020;8(12):e23530) doi:10.2196/23530

KEYWORDS

application programming interface; chronic disease; clinical decision rules; clinical decision support; medical monitoring

Introduction

Background

Cystic fibrosis (CF) is a life-limiting, recessively inherited disease resulting from mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) gene. Irregular functioning of the CFTR protein, which controls the transport of water and salt across epithelial cells in different organ systems, primarily affects the lungs [1]. Forced expiratory volume in 1 second (FEV_1), expressed as a percentage of an individual's predicted value based on normative standards for age, race, height, and sex (percent predicted FEV_1), is a measure of airway obstruction and a primary indicator of CF disease progression, severity, and efficacy of therapeutic interventions [2]. Acute decreases in FEV_1, clinically termed *rapid decline*, occur throughout adolescence and adulthood. Early prediction of FEV_1 decline is critical in order to initiate preventative interventions. Tools to predict rapid decline are crucial for clinical decision support and timely intervention. Various statistical models have been proposed and applied to understand and predict CF lung function over time [3,4]. Linear mixed-effects models with random intercepts and slopes are commonly employed but are problematic because lung function data are correlated within an individual over time in a potentially more complex and nonlinear manner [5]. CF studies show that lung function decline is nonlinear and heterogeneous; using an exponential correlation structure can improve predictions of lung function decline [5,6]. We recently used a nonstationary Gaussian linear mixed-effects model [7] to predict rapid FEV_1 decline using data from the US Cystic Fibrosis Foundation Patient Registry (CFFPR) [8]. Specifically, we applied a nonlinear model to simultaneously fit both population- and individual-level FEV_1 decline. We used integrated Brownian motion instead of random slopes to account for longitudinal correlation in each patient's lung function trajectory. We provided risk prediction of rapid decline in the form of predictive probabilities.

Objective

This study's objective was to translate our predictive algorithm into an interactive web-based graphical user interface that can be integrated with electronic health record systems and utilized by CF care providers. Over a 3-year period, we codeveloped the application with algorithm statisticians, programmers, and CF care providers. We have detailed our development process, including a multiphase study to acquire and incorporate clinician feedback, and our technical approach. The resulting application, Cystic Fibrosis Point of Personalized Detection (CFPOPD), is available online [9].

Methods

Application User Feedback

Participants

This study was conducted in the Cystic Fibrosis Care Center within the Division of Pulmonary Medicine of Cincinnati Children's Hospital Medical Center and was approved by the Cincinnati Children's Hospital Medical Center Institutional Review Board. Individuals involved in CF clinical care were eligible to participate; these included physicians, advanced practice nurses, social workers, dieticians, pharmacists, and respiratory therapists.

Procedures

Clinician feedback regarding the readability, feasibility, and perceptions of the CFPOPD application was collected in 2 phases. In the first phase, participants were encouraged to provide written feedback, drawings, and verbal comments. A semistructured interview guide was tailored to assess a given clinician's experience in using the application. Subsequent to the initial phase, additional feedback was gathered through either individual, semistructured interviews, or focus groups. Interview guides in the second phase were revised based on previously conducted clinician focus groups and revisions to the application. Clinician feedback was recorded and transcribed by MT-STAT, a commercial medical transcription company, and it was subsequently verified for accuracy and de-identified by study staff. When discussion prompted examples of specific patients or providers were referenced, names, places, family relationships, and other potentially identifying data were removed from the transcript [10].

Analysis

Initial interviews were analyzed using thematic analysis [11] in which transcribed data were used to generate codes based on participant feedback and were then grouped according to the arising motifs. These resulting themes and subthemes were used to advance application development.

Application Development

Data and Algorithm Development

The source of patient data used during CFPOPD development and the algorithm's development and validation has been described in detail elsewhere [8]. Briefly, we obtained data for

30,879 patients from the US CFFPR from 2003 to 2015 to train and validate our algorithm. Our model exhibited excellent predictive accuracy. Mean absolute percentage errors for the forecasted FEV_1 values in the validation sample for 6-month, 1-year, and 2-year intervals were within 5.6%, 6.9%, and 8.6% of patients' actual values, respectively. CFPOPD displays data from 4847 patients from the validation sample. Data within CFPOPD were de-identified by jittering demographic and clinical measurements and reassigning a separate identifier for the purpose of application development. Patients with CF contributed data to the registry at regular clinic visits that typically occurred at least once every 3 months and during suspected pulmonary exacerbations. The algorithm requires the input of a patient's longitudinal clinical history, including FEV_1, the number of pulmonary exacerbations in the last year, the number of clinic visits in the past year, the presence of CF-related diabetes, the presence of chronic *Pseudomonas aeruginosa* (Pa) infection, the presence of a persistent methicillin-resistant *Staphylococcus aureus* (MRSA) infection, and their utilization of public or private insurance. Furthermore, the algorithm takes as inputs time-invariant characteristics, including age and FEV_1 at entry, year of birth categorized into different cohorts, sex, and the number of F508del alleles.

Software Development

CFPOPD was built using R (version 3.6.1; R Core Team) [12] and R Shiny (version 1.4.0.2; RStudio) [13], a framework for interactive web applications and data visualization using R. Other packages used for development included emo, flexdashboard (version 0.5.1.1; RStudio), and plotly (version 4.9.2.1; Plotly) [14-16]. The software version control platform *git* was used to manage changes to the source code and implement modifications to CFPOPD functionality and features. The source code was hosted on GitHub, where multiple developers could track modifications to the source code, document software issues, and catalog major revisions through software releases. Each versioned release of the CFPOPD web application was deployed within a Docker container and stored on DockerHub to ensure a reproducible and automated workflow. A public version of the application suitable for interactive exploration is hosted online [9]. This paper describes version 7.1 of the software application.

Results

Initial Application Development

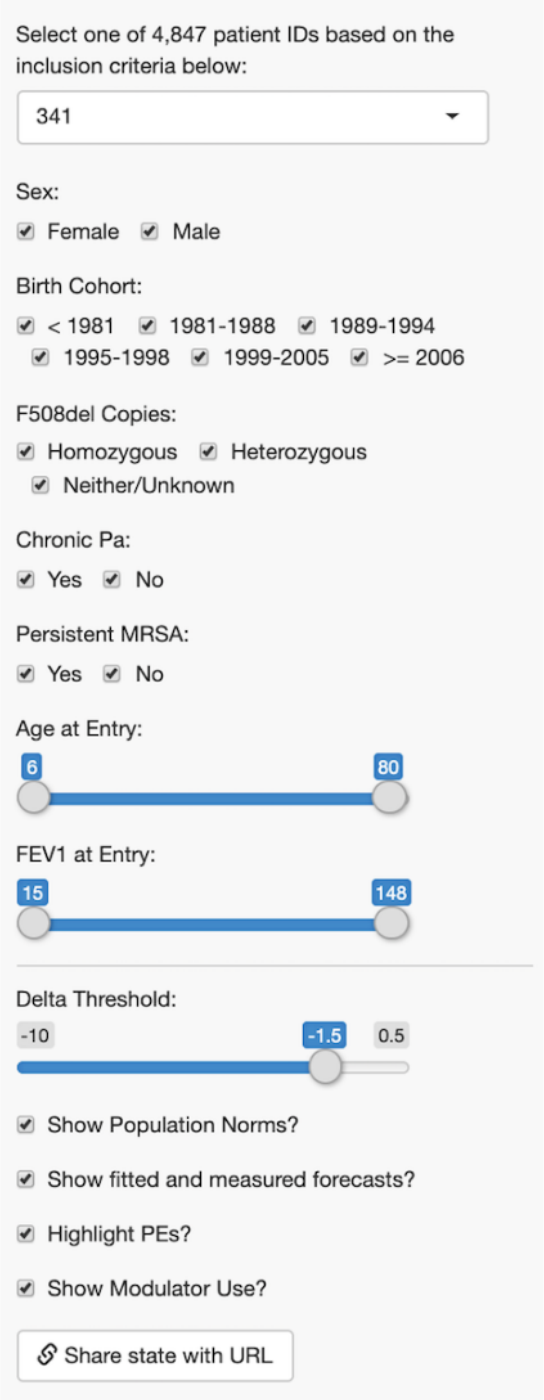
The progression of our application development is depicted in Figure 1 and shows screenshots of 4 CFPOPD versions (versions 1, 3, 5, and 7.1) in which significant revisions were implemented. Preliminary clinician feedback from CF chart and data conferences provided a blueprint for a bootstrap layout and structure, which was developed during the first 3 versions of CFPOPD [17]. The underlying layout and structure from CFPOPD (version 3) prior to formal clinician feedback remain the same in the current version, 7.1 (Figure 1). Clinician participants formally reviewed versions 3 and 7.1, and a subset of participants commented on intermittent updates to CFPOPD.

Figure 1. Progression of Cystic Fibrosis Point of Personalized Detection (CFPOPD) across multiple versioned releases. From versions 1 (top left) to 3 (top right), additional pulmonary function plots for the rate of forced expiratory volume in 1 second (FEV1) change and the risk of rapid decline was added. In version 5 (bottom left), users were given the ability to adjust the delta threshold to calculate the risk of rapid decline, and covariate information was moved to a table in the farthest right panel rather than a banner at the bottom of the application screen. The addition of a checkbox to visualize the initiation of modulator use was a key feature in version 7 (bottom right).

The leftmost sidebar of the application includes filtering options to enable a clinician to subset the data based on model covariates and other patient-level characteristics (Figure 2). Users can select a patient to explore via a drop-down list of identification

numbers. Patient data can also be filtered by toggling a sidebar checkbox and slider features for patient age at entry (coded as the first record available in the CFFPR registry data), FEV_1 at entry into the registry, patient sex, birth cohort group, F508del copies, chronic Pa, and persistent MRSA. The list of patient identification numbers is conditional on which features are selected and the available data. For example, if the user alters the minimum value for age at entry to 16 years of age, only patients 16 years of age or older will be available for selection. Similarly, a text box above the drop-down list displays changes dynamically and displays the number of patients available based on the selected filters.

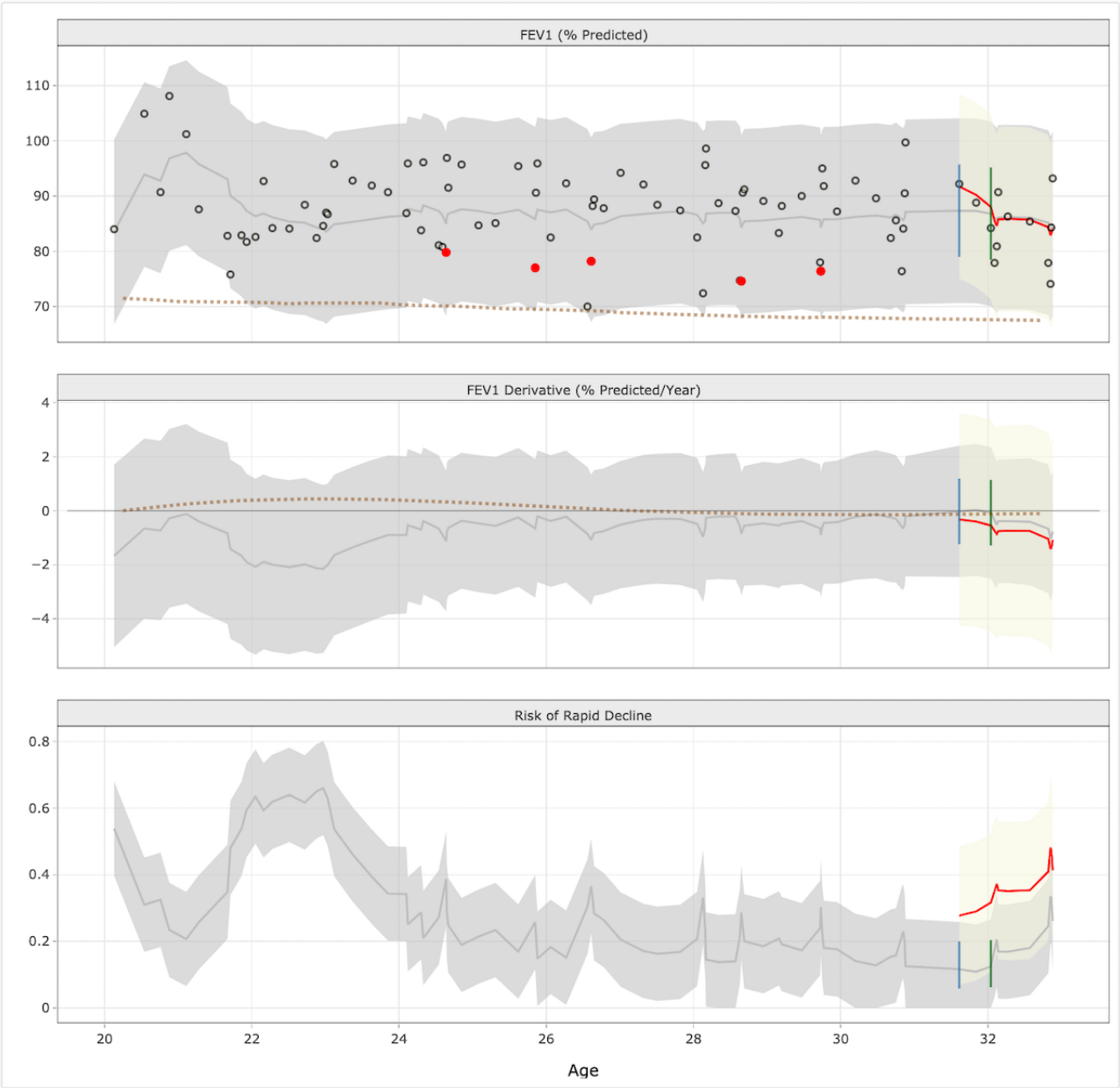
Figure 2. Leftmost panel of Cystic Fibrosis Point of Personalized Detection (CFPOPD). The drop-down menu shows patient 341 has been selected. Users can subset the patient sample by toggling options for sex, birth cohort, genotype (F508del copies), Pseudomonas aeruginosa (Pa) and Staphylococcus aureus (MRSA) infections, and forced expiratory volume in 1 second (FEV1) and age at entry into the US Cystic Fibrosis Foundation Patient Registry. A slider rule allows a user to select a delta threshold that is clinically relevant to a specific patient. Checkboxes allow users to select what data is viewable in the pulmonary function plots [ie, population norms, fitted and forecasted values, pulmonary exacerbations (PEs), and modulator use]. In the pictured instance, all subset and data viewing options have been selected.



CFPOPD has 2 main plot windows. The middle panel of our current application (Figure 3) displays pulmonary function data recorded over a patient's years of clinical follow-up via 3 faceted plots: observed percent predicted FEV_1 (top), predicted rate of

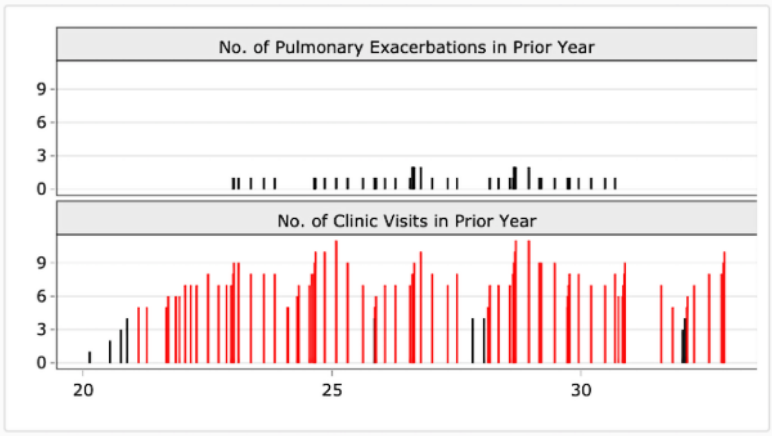
FEV_1 decline (middle), and predicted risk of rapid decline (bottom). Together, these 3 plots facilitate clinical interpretation of a patient's historical and future lung function trajectory. Bands surrounding each FEV_1 trajectory line show 95% confidence intervals to demonstrate the degree of uncertainty. Bands for fitted values are colored gray, and bands representing 2-year forecasted values are beige. For the 2-year forecasted period, we show the predictions holding this interval of data out of the model (the red trend line shown in each plot); the gray trend lines represent the predictions with the data included in the model. Both sets of trend lines were presented to clinician focus group participants in order to show model fit and transparency. In addition to filtering options, users can choose what underlying data is viewable in pulmonary function plots. In version 3, one toggle was made available that allowed users to view population norms for the FEV_1 rate of change and observed values. Normative data is generated through dynamic medians, which are computed based on the available patient data as specified by the filtering options; this was a suggestion from the aforementioned work soliciting informal feedback at chart review and data conference sessions [17].

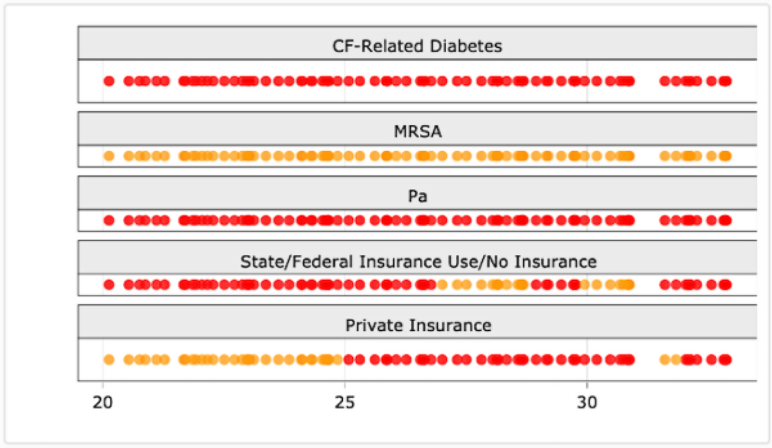
Figure 3. Middle panel of Cystic Fibrosis Point of Personalized Detection (CFPOPD). The 3 plots show pulmonary function data from patient 341. The top plot displays the patient's % predicted forced expiratory volume in 1 second (FEV1) values (circles) recorded during pulmonary function testing, as well as the patient's fitted (gray line) and forecasted (red line) values. Pulmonary function values recorded at the time a patient experienced a pulmonary exacerbation are colored red. A dotted line shows normative data (dynamic medians) respective to the patients % predicted values and rate of change in FEV1 (middle plot). The plot shows that the patient's rate of change in FEV1 fluctuated initially but has remained stable from ages 24 to 32 years. Compared to the overall norms, patient 341's rate of change is analogous to other patients. Similarly, the patient's risk of rapid decline initially fluctuated but declined and stabilized (bottom plot). All plots show that patient 341 was prescribed a modulator at 31 years of age (blue line; ivacaftor) and a second modulator at 32 years of age (green line; lumacaftor/ivacaftor).



The rightmost window in version 7.1 of CFPOPD (Figure 4) presents patient longitudinal covariate data and other disease status information such as the number of pulmonary exacerbations (denoted as "PEs" on the app) in the previous year, persistent MRSA, and CF-related diabetes. In version 3 of CFPOPD, this data was displayed using points plotted over time and colored to correspond to continuous and dichotomous variables, including the presence (red) or absence (gray) of clinical characteristics. Lastly, CFPOPD also displays time-invariant covariate information such as the selected patient's starting age, birth cohort, sex, and number of F508del copies. In version 3, these were shown in a horizontal table below the plotting windows.

Figure 4. Rightmost panel of Cystic Fibrosis Point of Personalized Detection (CFPOPD). The covariate table (bottom) shows that patient 341 is female, born between 1981 and 1988, enrolled in the Cystic Fibrosis Foundation Patient Registry at age 20, had a baseline of 84% predicted forced expiratory volume in 1 second (FEV1), and is homozygous for F508del copies. The top bar plot shows that she has had few pulmonary exacerbations (PEs) but numerous clinic visits throughout her clinical history. Binary covariate plots (middle) indicate that she has been diagnosed with cystic fibrosis (CF)-related diabetes and had not developed Staphylococcus aureus (MRSA) infection but has experienced chronic Pseudomonas aeruginosa (Pa) infection since age 20. The plots for insurance type indicate that she utilized public insurance at entry and transitioned between public and private insurance, beginning at around 25 years of age.





ID		341
Gender	♀	Female
Birth Cohort		1981-1988
Starting Age		20.13
Baseline FEV1	✅	84%
F508Del Copies		Homozygous
Mutation 1		F508del
Mutation 2		F508del

CFPOPD also features an 'About' tab in the top banner that describes the purpose of the application, defines application-specific terms, and instructs users how to use the data filtering and data viewing options. This section also provides a narrative of the clinical history and covariate information for an example patient (186) to illustrate CFPOPD's utility in clinical practice.

A key feature of CFPOPD is the interactivity of pulmonary function and covariate plots. Users have the capability to zoom in and pan across a specific year in a patient's clinical history. Faceted FEV_1 plots are also linked. For example, if a user zooms to a specific range of ages in the bottom pulmonary function plot where a patient's risk of rapid decline appears to change, the same period of interest will be displayed in the plots for FEV_1 derivative and observed FEV_1 values. The scales on the x- and y-axes also change dynamically. An additional interactive feature includes text hovering. When a user scans across the plots with the cursor, a text window will display the values of the underlying data.

Focus Group and Conceptual Model

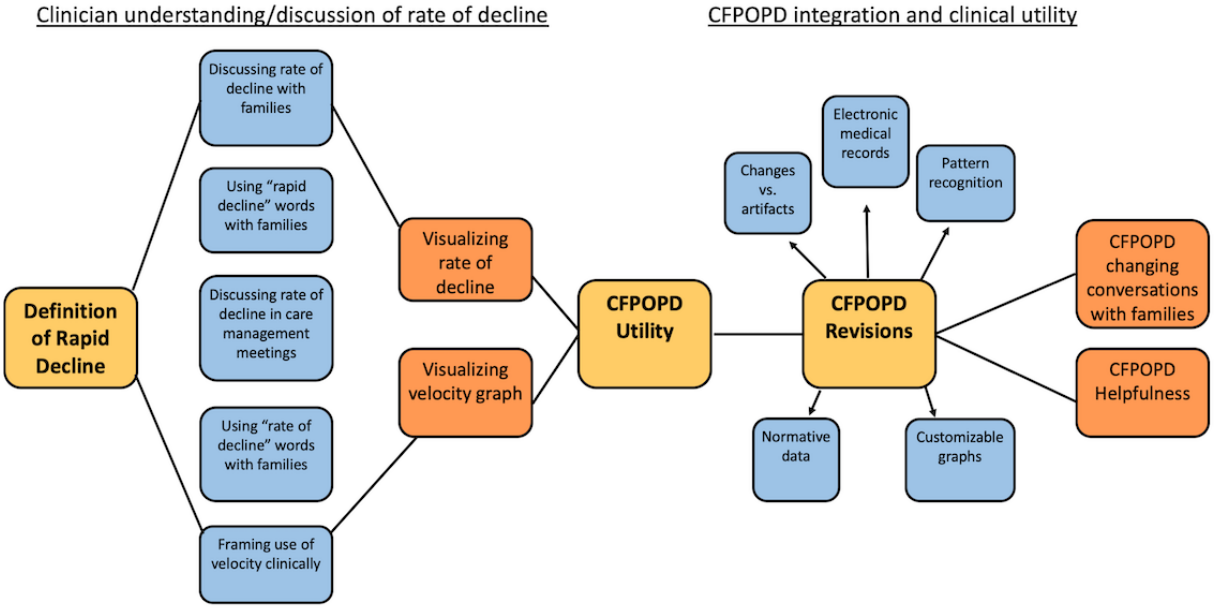
A total of 17 clinicians (6 attending pulmonologists, 1 nurse practitioner, and 10 pulmonary research fellows) participated in 2 formal focus group sessions (Table 1). The first session included attending physicians and a nurse practitioner, while the second session included fellows. Fellows were grouped separately from attending physicians and other standing members of the care teams, given their roles as trainees. Select participants from the attending and nurse practitioner session were followed up in individual interviews for additional feedback after CFPOPD updates were made based on the focus group. We followed up with a subset of 3 participants from the fellows' session. Participants were chosen for follow-up based on the salience of their feedback. Prior iterations garnering feedback through CF-specific chart and data conferences consisted of 35 members across the clinical care teams.

Table 1. Focus group participant (clinician) characteristics (n=17).

Clinician characteristics	Women (n=9), n (%)	Men (n=8), n (%)	Total (n=17), n (%)
Ethnicity			
Hispanic or Latino	0 (0%)	2 (25%)	2 (12%)
Not Hispanic or Latino	9 (100%)	6 (75%)	15 (88%)
Race			
Asian	1 (11%)	0 (0%)	1 (6%)
Black/African American	1 (11%)	1 (12%)	2 (12%)
White	7 (78%)	7 (88%)	14 (82%)

As a result of focus group sessions, we developed a conceptual model of clinician perceptions toward rapid decline and CFPOPD integration (Figure 5). Key a priori discussion points were the definition of rapid decline, challenges to CFPOPD utility, and revisions (yellow boxes). The first discussion point illuminated how clinicians use different communication techniques with families as opposed to care teams when referring to the rate of decline. Clinicians expressed hesitation with using the phrase "rapid decline." There was also difficulty expressed in the concept of "rate of decline" and how to conceptualize rate as velocity. Challenges to CFPOPD utility, which prompted ways to improve the application, focused on electronic health record (EHR) accessibility, distinguishing change in FEV_1 from artifacts, and the desire to have a decision support tool that could help reveal patterns in FEV_1 trajectories. Actionable revisions included the development of dynamic medians, which allowed for the use of normative data and customizable graphics. Participants described how CFPOPD could be used to strengthen conversations with patients and families, particularly in promoting adherence to therapies. Another identified area of potential clinical significance was its use in communicating rapid disease progression during inpatient settings, as CFPOPD could serve as a motivation to improve the clinical course or initiate antibiotic therapy to raise lung function levels. Targeted interviews prompted further CFPOPD developments.

Figure 5. Conceptual clinical model of rapid decline. Larger yellow-shaded boxes with bold text represent key discussion points during focus group sessions. The first half of the diagram summarizes clinician understanding and discussion with other care team members and patients/families. Clinicians translated concepts of rapid decline into optimizing Cystic Fibrosis Point of Personalized Detection (CFPOPD) data visualization/monitoring capabilities. The second half represents CFPOPD integration and clinical utility. Clinicians identified challenges to the user interface and suggested revisions. The rightmost boxes represent clinician feedback on how the use of CFPOPD would change the conversations with families or otherwise be helpful.



Based on coded feedback from focus group participants and semistructured interviews, 4 primary themes were identified, providing granularity to the conceptual model in Figure 5. Each theme and corresponding illustrative quotes from focus group participants are shown in Textbox 1. Clinicians expressed uncertainty regarding the definition of rapid lung function decline (list 1, *Ambiguity*). The other 3 themes focused on the CFPOPD application's utility, clinical significance, and suggested revisions (lists 2-4). CFPOPD facilitated the clinicians' ability to decipher trends in a patient's FEV_1, recognize when a patient may be at risk of rapid FEV_1 decline, and assist in determining the clinical impact of treatment interventions. Focus group participants stated that CFPOPD may be a useful educational tool (list 2, quotes a-d). Visualizing a patient's clinical history assisted clinical adjudication (list 2, quotes e-f). Still, some care providers expressed concern that the CFPOPD may cause confusion in patient and family interactions (list 2, quotes g-h). Clinician feedback demonstrated that our application had the potential to advance clinical practice by facilitating decision-making, discussions with patients, and identification of rapid decline. Care providers articulated that incorporating CFPOPD into previsit planning meetings would improve point-of-care decision-making and facilitate conversations between families and the care team (list 3, quote a). Physicians stated that visualizing a patient's risk of rapid decline may also be used as a motivator by eliciting treatment adherence (list 3, quote b). Clinicians recognized the value of CFPOPD and the capability to advance clinical practice (list 3, quotes c-f). Caution was expressed regarding its impact during inpatient visits, as it could serve as a demotivator (list 3, quote g). Provider feedback regarding revisions to CFPOPD has been critical to ensuring our instrument is translational, relevant, and impactful in clinical practice (list 4, quotes a-d).

Textbox 1. Emergent themes and accompanying quotes from clinician focus groups.

Defining Rapid Decline:

- 1. Ambiguity
  a. "[I am] more likely to refer to the curve in a clinical setting than to a threshold that is going to capture almost every patient."
  b. "Really hard to define."
  c. "If we were able to tweak [the definition] 'rapid decline'…go for minimal change in lung function over time as opposed to something that might be more realistic for the patient."

Cystic Fibrosis Point of Personalized Detection (CFPOPD) Application:

- 2. Utility
  a. "Oh my gosh, it's just what I wanted."
  b. "Yes, [I] would use graphs in preclinic meetings."
  c. "… helpful both on a sort of clinical decision-making side and describing it to families' side."
  d. "As a fellow trainee, I feel sometimes that it's really difficult for me to see that big picture."
  e. "Great that hovering gives you the exact numbers."
  f. "If you can show some improvement in the derivative, in the trajectory, it's more cause for optimism."
  g. "I don't think it would be helpful at all to show a family. I think it is complicated for families; it's complicated for me."
  h. "I like graphs in talking with families, but as a clinician, I think the only one I would feel comfortable using would be the top one."
- 3. Clinical Significance
  a. "If you have a visual representation like that, it would be substantially more helpful than me verbally saying, 'You're getting worse faster than we think you should.'"
  b. "I would definitely show a 16-year-old who is noncompliant…'if you don't step it up, this is where you are going.'"
  c. "10 years ago, we were just trying to look at random pieces of paper, and we never could see any of this whatsoever."
  d. "… put this in Epic."
  e. "These are things you can intervene on if you knew 5 years ago this trend was coming."
  f. "If you look at any clinical trial or any aspect of medicine, the more frequent your intervention is, the more frequent your clinic visits, the more frequent you're ahead of this data, the better your outcomes."
  g. "… billboard of death."
- 4. Revisions
  a. "Customize threshold for rapid decline…if you want to call rapid as 3% or as 6% or 10%...you can play with that."
  b. "Add mutation classes and modulator therapy use."
  c. "Categorize continuous covariates based on clinical severity."
  d. "Different dots and colors…what's bad and what's steady."

Further Application Development

Our collaborative approach to developing CFPOPD has allowed our team of programmers to prospectively track its evolution, as shown in Figure 1. Data filters, pulmonary function data-viewing options, covariate information, coloring according to values, and icon typography were added to the application based on feedback received from clinical application users. Subsequent to clinician feedback, we implemented a feature to enable users to adjust the threshold value for percent predicted FEV_1 loss or delta threshold, used to calculate a patient's risk of rapid decline (Textbox 1, list 4, a). This threshold can be modified by manipulating the slider to the desired value, which ranges from -10% to 0.5% (Figure 2). The default threshold of -1.5% predicted/year was chosen previously [17].

We incorporated CF registry data on modulator use and mutation type (Textbox 1, list 4, b) through a checkbox in the left sidebar ('Show Modulator Use?'). If a patient has been prescribed a modulator, vertical lines are shown on each pulmonary function graph at the age medication was first administered (Figure 3). When hovering over the vertical line, a window stating the name of the medication and age at administration is displayed. The names of each patient's CFTR gene mutations were added to the covariate table (Figure 4).

Clinician feedback (Textbox 1, list 4, c-d) to categorize covariate information and assign clinical severity based on color was applied to pulmonary exacerbation and visit frequency plots. Pulmonary exacerbations are acute respiratory events that can emerge from precipitous drops in lung function. We revised the color scheme according to a categorical designation versus the continuous scale from version 3. Occurrences greater than 5 are colored red to designate an exceedance of the clinical threshold (Figure 4). In order to enhance a clinician's ability to visualize pulmonary exacerbations and rapid decline, a checkbox option ('Highlight PEs?') was added to the left sidebar (Figure 2). When checked, a patient's FEV_1 value in the top pulmonary function plot will be colored red if a pulmonary exacerbation was observed (Figure 3).

Other CFPOPD revisions were based on informal feedback or implemented to optimize application functionality and comprehension. To maximize the space to visualize pulmonary function plots, we repositioned the covariate table underneath the covariate dot plot (Figure 4). We also increased the pixel width of the pulmonary function plots to improve readability and a checkbox that allowed users to toggle whether patient FEV_1 values are displayed in the top pulmonary function plot ('Show Fitted and Measured Forecasts?'). Depending on the number of spirometry results, removing FEV_1 values from the plot may facilitate a clinician's ability to decipher rapid decline (Figure 3). These revisions were completed under CFPOPD version 4.

We supplemented the covariate table with emojis to increase the ease of visual interpretation, implemented in version 5. Where applicable, emojis change dynamically according to the age and sex of the selected patient. The standard symbol for either male (♂) or female (♀) is shown to communicate the selected patient's sex, and depending on if the patient is younger or older than 18 years of age, either a girl, boy, woman, or man emoji is shown to communicate the starting age.

Lastly, binary dot plots of the number of PEs and clinic visits a patient experienced in the previous year were modified to bar plots in version 6. In addition to colored bars indicating clinical severity, this second dimension enhances a user's ability to visually evaluate a patient's clinical trajectory.

Discussion

Principal Findings

We developed and coproduced an interactive web application designed to facilitate clinical point-of-care decision-making by predicting acute pulmonary function decline in patients with CF. We conducted focus groups with clinicians and CF care providers to garner feedback on a prototype application [17] and used this feedback to further develop the application in order to advance its utility for clinical care.

Clinicians suggested insightful and actionable CFPOPD revisions, which we incorporated over the course of 4 versioned releases. A principal revision was to add a feature enabling care providers to tailor the delta threshold according to their clinical judgment and characteristics of an individual patient. Implementing this capability was paramount to ensure CFPOPD was applicable in clinical practice. Adding this feature also manifested in a related theme regarding uncertainty toward a single clinical definition of "rapid decline."

With the advent of modulator therapies, another requested modification was to include visualization of modulator use and descriptive text to communicate patient mutations. While numerous therapies exist to mitigate and treat acute symptoms in CF, modulator therapies act at a molecular level to restore function to CFTR protein [1,18]. By enabling care providers to detect when a patient is at risk for acute decline in pulmonary function, CFPOPD may facilitate clinical judgment and decision-making regarding the initiation of acute therapies, such as intravenous antibiotics. Previous research has shown that a treatment of acute drops in FEV_1 using intravenous antibiotics improved long-term pulmonary function [19]. Similarly, if a patient is currently prescribed a modulator, our application allows care providers to track a patient's lung function prospectively and assess the effectiveness of personalized treatment regimens. CFPOPD has implications for emerging studies involving patient withdrawal of maintenance therapies, given observed effectiveness for select combinations of mutations and modulators.

Technological advances in electronic data storage have transformed the management of medical records, greatly increasing the volume of data accessible to researchers, clinicians, and patients [20]. This abundance of information has yielded opportunities for novel development of interactive applications to synthesize, model, and translate EHR data [21]. Web-based applications have been employed across research and medical domains, ranging from infection management [22] to personalized mental health monitoring [23]. Likewise, others have leveraged visual analytics to translate results from complex statistical techniques used in EHR research, such as case-crossover design [24] and hierarchical clustering [25], into a comprehensible form. We sought to develop CFPOPD in order to improve point-of-care decision-making, and feedback from clinicians at our institution demonstrates our application has the potential to do so. Furthermore, clinician responses also indicate CFPOPD may promote communication and shared decision-making. Previous research indicates that participatory decision-making between physicians and their patients results in greater patient satisfaction [26]. Care providers noted that CFPOPD use may encourage adherence among patients with CF that are noncompliant, and there is empirical evidence to support this. Heisler et al [27] have shown that effective communication and shared decision-making are associated with positive diabetes self-management.

Limitations

Although our results indicate that CFPOPD has the potential to positively impact clinical care, some feedback suggests that care provider comprehension is not universal. Additional training may be necessary before our application is fully deployed for clinical practice. Some discord existed among physicians as to whether our application would facilitate conversations between patients/families and the clinical care team, as clinicians expressed differing opinions regarding

approaches to communicate a high risk of rapid decline. To accommodate this limitation, future revisions to CFPOPD could include additional options that allow care providers to customize CPOPD's layout by selecting only plots that are relevant to the patient-provider discourse. Currently, CFPOPD is limited to existing fields available from the CFFPR data. Risk calculations are not computed in real time; rather, values are pulled from precomputed lookup tables. While our application demonstrates the predictive accuracy of our algorithm, further development is needed to integrate CFPOPD into near real-time clinical practice. Lastly, our findings are based on a single-center study (Table 1). We anticipate drawing a larger, more diverse sample of care teams in future multicenter studies assessing CFPOPD feasibility and acceptability.

Future Work

Our future work will address CFPOPD limitations; chiefly, we will strive to implement CFOPD into an EHR system to provide "now-casting," or near real-time statistical predictions of rapid decline. In addition to rapid decline, a similar area of extension is to calculate risk probabilities for pulmonary exacerbation onset. Recently, a data-driven definition for pulmonary exacerbation has been proposed and is being tested by the Cystic Fibrosis Learning Network [28]. Making CFPOPD available for use in clinical practice will enable assessment of its impact on clinical practice and patient outcomes. It may be desirable for patients to access their longitudinal data as well, which could potentially be made available to patients through the medical institution's patient portal. Given emerging public health issues and a drastic increase in telehealth, integrating home spirometry into CFPOPD may become a critical priority. Combined with access to the CFPOPD application through a care provider's patient portal, this extension could facilitate home monitoring and diagnosis of acute drops in lung function among patients with CF being clinically followed via telemedicine. The developmental framework outlined herein is capable of adaptation to different clinical markers or chronic diseases, such as diabetes and asthma, for which longitudinal tracking is valuable.

Conclusions

We developed CFPOPD to translate a novel predictive algorithm into an interactive clinical tool to enhance early detection and forecasting of rapid pulmonary function decline in patients with CF. Our application was built through an iterative and collaborative process among programmers, statisticians, and clinicians. We have demonstrated that this framework of collaborative design between developers and end-users is successful, capable of delivering an impactful product, and may be generalized to other chronic diseases and disorders that rely on routinely collected clinical data for medical monitoring and decision-making.

Acknowledgments

This study was funded by the National Heart, Lung, and Blood Institute of the National Institutes of Health, grants K25 HL125954, R01 HL141286, R01 HL142210, and R61 HL154105; and Ohio Development Services, grant TECG2019-0159.

The authors would like to thank the clinicians and care providers for their feedback offered during focus group sessions and eagerness to improve our CFPOPD application. The authors acknowledge Elizabeth Shepherd's efforts to format the conceptual model diagram. The authors also thank the Cystic Fibrosis Foundation for supplying data used in model development for the CFPOPD, as well as the patients, care providers, and clinical coordinators for their contributions to the Registry.

Conflicts of Interest

AZ, JPC, RDS, and CB are co-inventors on a provisionally approved patent, Application No. 15/927,575, under disclosure D17-0021 and tech ID # 2017-0211. For all other authors, there are no conflicts of interest to declare.

References

1. De Boeck K, Amaral MD. Progress in therapies for cystic fibrosis. The Lancet Respiratory Medicine 2016 Aug;4(8):662-674. [doi: 10.1016/s2213-2600(16)00023-0]
2. Szczesniak R, Heltshe SL, Stanojevic S, Mayer-Hamblett N. Use of FEV1 in cystic fibrosis epidemiologic studies and clinical trials: A statistical perspective for the clinical researcher. Journal of Cystic Fibrosis 2017 May;16(3):318-326. [doi: 10.1016/j.jcf.2017.01.002]
3. Salvatore D, Buzzetti R, Mastella G. An overview of international literature from cystic fibrosis registries. Part 5: Update 2012-2015 on lung disease. Pediatr Pulmonol 2016 May 10;51(11):1251-1263. [doi: 10.1002/ppul.23473]
4. Harun SN, Wainwright C, Klein K, Hennig S. A systematic review of studies examining the rate of lung function decline in patients with cystic fibrosis. Paediatr Respir Rev 2016 Sep;20:55-66. [doi: 10.1016/j.prrv.2016.03.002] [Medline: 27259460]
5. Taylor-Robinson D, Whitehead M, Diderichsen F, Olesen HV, Pressler T, Smyth RL, et al. Understanding the natural progression in %FEV decline in patients with cystic fibrosis: a longitudinal study. Thorax 2012 May 03;67(10):860-866. [doi: 10.1136/thoraxjnl-2011-200953]
6. Szczesniak RD, McPhail GL, Duan LL, Macaluso M, Amin RS, Clancy JP. A semiparametric approach to estimate rapid lung function decline in cystic fibrosis. Annals of Epidemiology 2013 Dec;23(12):771-777. [doi: 10.1016/j.annepidem.2013.08.009]

7. Diggle PJ, Sousa I, Asar O. Real-time monitoring of progression towards renal failure in primary care patients. Biostatistics 2015 Jul;16(3):522-536. [doi: 10.1093/biostatistics/kxu053] [Medline: 25519432]
8. Szczesniak RD, Su W, Brokamp C, Keogh RH, Pestian JP, Seid M, et al. Dynamic predictive probabilities to monitor rapid cystic fibrosis disease progression. Stat Med 2020 Mar 15;39(6):740-756 [FREE Full text] [doi: 10.1002/sim.8443] [Medline: 31816119]
9. 2020;. Cystic Fibrosis Point of Personalized Detection. URL: http://clinic.predictfev1.com [accessed 2020-11-18]
10. Saunders B, Kitzinger J, Kitzinger C. Anonymising interview data: challenges and compromise in practice. Qual Res 2015 Oct;15(5):616-632 [FREE Full text] [doi: 10.1177/1468794114550439] [Medline: 26457066]
11. Lerra M. Transforming qualitative information: Thematic analysis and code development. In: Transforming qualitative information: Thematic analysis and code development. Thousand Oaks, CA: Sage; 1998.
12. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/ [accessed 2020-11-18]
13. Chang W, Cheng J, Allaire J, Xie Y, McPherson J. shiny: web application framework for R. R package version 132. URL: https://CRAN.R-project.org/package=shiny [accessed 2020-11-18]
14. Iannone R, Allaire J, Borges B. flexdashboard: R markdown format for flexible dashboards. R package version 0511. URL: https://CRAN.R-project.org/package=flexdashboard [accessed 2020-11-18]
15. plotly for R. Sievert C. 2018. URL: https://plotly-rcom [accessed 2020-11-18]
16. Wickham H, Francois R, D'Agostino ML. emo: easily insert 'emoji'. R package version 0009000. URL: https://github.com/hadley/emo [accessed 2020-11-18]
17. Szczesniak RD, Brokamp C, Su W, Mcphail GL, Pestian J, Clancy JP. Improving Detection of Rapid Cystic Fibrosis Disease Progression–Early Translation of a Predictive Algorithm Into a Point-of-Care Tool. IEEE J. Transl. Eng. Health Med 2019;7:1-8. [doi: 10.1109/jtehm.2018.2878534]
18. Solomon GM, Marshall SG, Ramsey BW, Rowe SM. Breakthrough therapies: Cystic fibrosis (CF) potentiators and correctors. Pediatr Pulmonol 2015 Oct;50 Suppl 40:S3-S13 [FREE Full text] [doi: 10.1002/ppul.23240] [Medline: 26097168]
19. Schechter MS, Schmidt HJ, Williams R, Norton R, Taylor D, Molzhon A. Impact of a program ensuring consistent response to acute drops in lung function in children with cystic fibrosis. Journal of Cystic Fibrosis 2018 Nov;17(6):769-778. [doi: 10.1016/j.jcf.2018.06.003]
20. Caban JJ, Gotz D. Visual analytics in healthcare--opportunities and research challenges. J Am Med Inform Assoc 2015 Mar;22(2):260-262. [doi: 10.1093/jamia/ocv006] [Medline: 25814539]
21. Simpao A, Ahumada L, Larru Martinez B, Cardenas A, Metjian T, Sullivan K, et al. Design and Implementation of a Visual Analytics Electronic Antibiogram within an Electronic Health Record System at a Tertiary Pediatric Hospital. Appl Clin Inform 2018 Jan 17;09(01):037-045. [doi: 10.1055/s-0037-1615787]
22. Luz CF, Berends MS, Dik JH, Lokate M, Pulcini C, Glasner C, et al. Rapid Analysis of Diagnostic and Antimicrobial Patterns in R (RadaR): Interactive Open-Source Software App for Infection Management and Antimicrobial Stewardship. J Med Internet Res 2019 May 24;21(6):e12843. [doi: 10.2196/12843]
23. Kaiser T, Laireiter AR. DynAMo: A Modular Platform for Monitoring Process, Outcome, and Algorithm-Based Treatment Planning in Psychotherapy. JMIR Med Inform 2017 Jul 20;5(3):e20 [FREE Full text] [doi: 10.2196/medinform.6808] [Medline: 28729233]
24. Caron A, Chazard E, Muller J, Perichon R, Ferret L, Koutkias V, et al. IT-CARES: an interactive tool for case-crossover analyses of electronic medical records for patient safety. J Am Med Inform Assoc 2017 Mar 01;24(2):323-330 [FREE Full text] [doi: 10.1093/jamia/ocw132] [Medline: 27678461]
25. Feller DJ, Burgermaster M, Levine ME, Smaldone A, Davidson PG, Albers DJ, et al. A visual analytics approach for pattern-recognition in patient-generated data. J Am Med Inform Assoc 2018 Oct 01;25(10):1366-1374 [FREE Full text] [doi: 10.1093/jamia/ocy054] [Medline: 29905826]
26. Kaplan SH, Greenfield S, Gandek B, Rogers WH, Ware JE. Characteristics of physicians with participatory decision-making styles. Ann Intern Med 1996 Mar 01;124(5):497-504. [doi: 10.7326/0003-4819-124-5-199603010-00007] [Medline: 8602709]
27. Heisler M, Bouknight RR, Hayward RA, Smith DM, Kerr EA. The relative importance of physician communication, participatory decision making, and patient understanding in diabetes self-management. J Gen Intern Med 2002 Apr;17(4):243-252. [doi: 10.1046/j.1525-1497.2002.10905.x]
28. Cystic Fibrosis Foundation. CF Learning Network: Fall Community Conference and Implementation Phase. CF Learning Network: Fall Community Conference and Implementation Phase. 2019 Jan 01. URL: https://tinyurl.com/y42ggn4n [accessed 2020-03-27]

Abbreviations

CF: cystic fibrosis
CFFPR: Cystic Fibrosis Foundation Patient Registry
CFPOPD: Cystic Fibrosis Point of Personalized Detection

CFTR: cystic fibrosis transmembrane conductance regulator
EHR: electronic health record
FEV1: forced expiratory volume in 1 second
MRSA: methicillin-resistant Staphylococcus aureus
Pa: Pseudomonas aeruginosa
PEs: pulmonary exacerbations

Edited by G Eysenbach; submitted 14.08.20; peer-reviewed by M Cheng, S Sarbadhikari, D Newman; comments to author 19.09.20; revised version received 02.10.20; accepted 30.10.20; published 16.12.20.

Please cite as:
Wolfe C, Pestian T, Gecili E, Su W, Keogh RH, Pestian JP, Seid M, Diggle PJ, Ziady A, Clancy JP, Grossoehme DH, Szczesniak RD, Brokamp C
Cystic Fibrosis Point of Personalized Detection (CFPOPD): An Interactive Web Application
JMIR Med Inform 2020;8(12):e23530
URL: https://medinform.jmir.org/2020/12/e23530
doi:10.2196/23530
PMID:33325834

Original Paper

Missing-Data Handling Methods for Lifelogs-Based Wellness Index Estimation: Comparative Analysis With Panel Data

Ki-Hun Kim[1,2], PhD; Kwang-Jae Kim[3], PhD

[1]Faculty of Industrial Design Engineering, Delft University of Technology, Delft, Netherlands

[2]Department of Industrial Engineering, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea

[3]Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang, Republic of Korea

Corresponding Author:
Ki-Hun Kim, PhD
Faculty of Industrial Design Engineering
Delft University of Technology
Landbergstraat 15
Delft, 2628 CE
Netherlands
Phone: 31 625244785
Fax: 31 152787316
Email: K.Kim-1@tudelft.nl

Abstract

Background: A lifelogs-based wellness index (LWI) is a function for calculating wellness scores based on health behavior lifelogs (eg, daily walking steps and sleep times collected via a smartwatch). A wellness score intuitively shows the users of smart wellness services the overall condition of their health behaviors. LWI development includes estimation (ie, estimating coefficients in LWI with data). A panel data set comprising health behavior lifelogs allows LWI estimation to control for unobserved variables, thereby resulting in less bias. However, these data sets typically have missing data due to events that occur in daily life (eg, smart devices stop collecting data when batteries are depleted), which can introduce biases into LWI coefficients. Thus, the appropriate choice of method to handle missing data is important for reducing biases in LWI estimations with panel data. However, there is a lack of research in this area.

Objective: This study aims to identify a suitable missing-data handling method for LWI estimation with panel data.

Methods: Listwise deletion, mean imputation, expectation maximization–based multiple imputation, predictive-mean matching–based multiple imputation, k-nearest neighbors–based imputation, and low-rank approximation–based imputation were comparatively evaluated by simulating an existing case of LWI development. A panel data set comprising health behavior lifelogs of 41 college students over 4 weeks was transformed into a reference data set without any missing data. Then, 200 simulated data sets were generated by randomly introducing missing data at proportions from 1% to 80%. The missing-data handling methods were each applied to transform the simulated data sets into complete data sets, and coefficients in a linear LWI were estimated for each complete data set. For each proportion for each method, a bias measure was calculated by comparing the estimated coefficient values with values estimated from the reference data set.

Results: Methods performed differently depending on the proportion of missing data. For 1% to 30% proportions, low-rank approximation–based imputation, predictive-mean matching–based multiple imputation, and expectation maximization–based multiple imputation were superior. For 31% to 60% proportions, low-rank approximation–based imputation and predictive-mean matching–based multiple imputation performed best. For over 60% proportions, only low-rank approximation–based imputation performed acceptably.

Conclusions: Low-rank approximation–based imputation was the best of the 6 data-handling methods regardless of the proportion of missing data. This superiority is generalizable to other panel data sets comprising health behavior lifelogs given their verified low-rank nature, for which low-rank approximation–based imputation is known to perform effectively. This result will guide missing-data handling in reducing coefficient biases in new development cases of linear LWIs with panel data.

(JMIR Med Inform 2020;8(12):e20597) doi:10.2196/20597

KEYWORDS

lifelogs-based wellness index; missing-data handling; health behavior lifelogs; panel data; smart wellness service

Introduction

Background

Smart wellness services are designed to help individuals monitor their own wellness through smart devices, including smartphones and smartwatches [1]. Reports indicate that these services will see exponential growth alongside continued smart device penetration and the increasing size of the wellness market [2]. Their popularity is further evidenced by the high number of mobile health apps, with around 325,000 available in app stores in 2017 [3,4].

Smart wellness services can collect various health behavior lifelogs through the aid of smart devices [5]. For example, smartwatches, such as Fitbit, can record daily walking steps, total distances, and the number of sleeping hours [6], while smart patches, such as HealthPatch, can monitor heart rate, breathing rate, skin temperature, posture, number of walking steps, activity patterns, and sleep habits [7]. There are also devices for infants, such as Owlet smart socks, that send the child's vital signs to their parents via smartphones, including information on heartrate, oxygen level, skin temperature, sleep quality, and sleeping position [8].

Existing smart wellness services utilize health behavior lifelogs to provide users with detailed records about health behaviors [9]. Fitbit provides a smart wellness service that primarily shows users detailed activity records (eg, daily walking steps), exercise habits (eg, type, time, and duration), sleep information (eg, start and end times), and dietary facts (eg, daily calorie intake). By focusing on the details of each health behavior, existing smart wellness services have a limitation in supporting users to easily identify their aggregate condition from multiple health behaviors. Users must synthesize the information, making it difficult to monitor overall progress.

A lifelogs-based wellness index (LWI), a function that transforms health behavior lifelogs into wellness scores for smart wellness service users, resolves this limitation [10]. The wellness scores quantitatively represent how well the user meets relevant recommended health behaviors. Such information, including a user's current or past wellness scores, wellness score progress over time, and comparisons of their wellness scores [11], can be offered by smart wellness services. According to Platt et al [12], a wellness index is a critical feature of wellness apps for younger demographics. The utility of LWIs is thus expected to stimulate new LWI development .

An LWI can be developed through 3 key phases: definition, estimation, and assessment [10,11]. The definition phase refers to the selection of the LWI function type and a model for estimating the function that consists of behavior variables and a proxy variable as its independent variables and dependent variable, respectively. The behavior variables are potential constituents of an LWI, while the proxy variable is used in place of wellness scores, immeasurable during the development process. The estimation phase refers to the process of estimating the coefficients of the behavior variables in LWIs by collecting and preprocessing data, which are then fit with the estimation model. The assessment phase refers to the assessment of LWI generalizability and utility for users.

LWI estimation can lead to the reduction of coefficient biases through a panel data set of health behavior lifelogs. A panel data set follows a given sample of participants over time, thus providing multiple observations for each participant. Existing panel data analysis methods (eg, 1-way random effects regression) can only be applied to panel data sets. These methods can reduce biases in the coefficients by controlling for heterogeneity across participants, which is caused by unobserved variables [13].

A panel data set comprising health behavior lifelogs will likely contain large proportions of missing data. Such a data set is collected based on everyday user activities and is therefore exposed to various random events that result in missing data. For example, users may forget to wear smart devices or to record health behavior lifelogs, and the smart devices themselves will no longer record health behavior lifelogs when batteries are depleted. These random events often lead to large proportions of missing data. For example, missing data accounted for 18% of a panel data set in an LWI development case [10]. This rate was considered high considering that participants received reminders for the data collection.

Missing data can lead to 2 severe problems when attempting to estimate LWI coefficients. First, it can introduce biases to the coefficients [14,15]. This leads to low LWI generalizability for users. Second, most existing data analysis methods are only applicable to complete data sets (ie, data sets without missing data). Thus, incomplete data sets must be modified into complete ones [16]. A variety of missing data handling methods exist to address these problems, the choice of which becomes increasingly significant as the proportion of missing data increases [17]. However, few studies have identified which existing method is suitable for handling missing data in a panel data set that is composed of health behavior lifelogs.

This study identified a suitable method for LWI estimation with panel data based on an examination of 6 representative missing-data handling methods: listwise deletion, mean imputation, expectation maximization–based multiple imputation, predictive-mean matching–based multiple imputation, k-nearest neighbors–based imputation, and low-rank approximation–based imputation. These were selected from common missing-data handling methods from previous studies, specifically because they represented possible missing-data handling approaches in the context of LWI estimation.

The 6 abovementioned missing-data handling methods were comparatively evaluated for various missingness proportions of a panel data set by simulating an LWI development case originally presented by Kim et al [10]. The case estimated the coefficients in a linear LWI with a panel data set composed of health behavior lifelogs. Such cases are expected to become

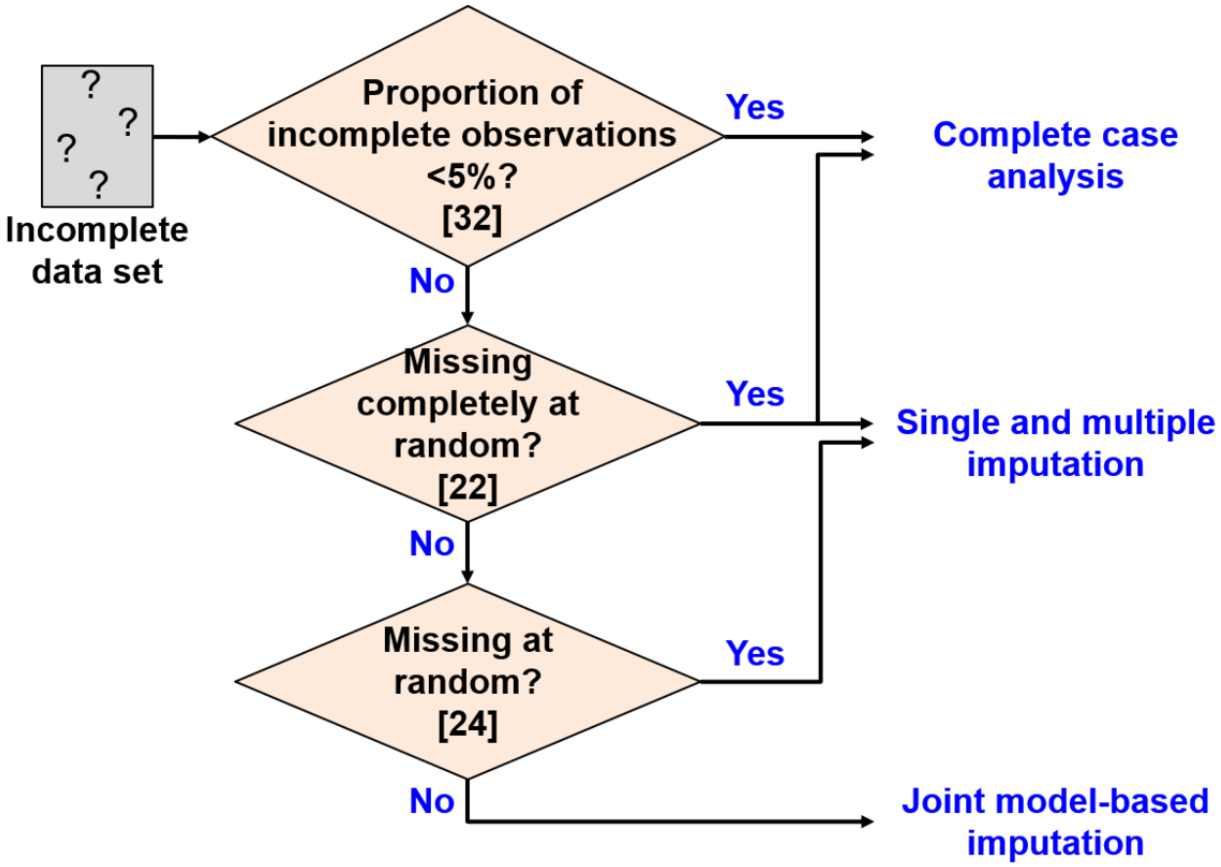
prevalent because linear functions help users understand how changes in each behavior variable influence their overall wellness scores [18]. This advantage of linear LWIs enables users to obtain 2 types of valuable insights. First, users can easily see which behavior variables substantially decrease or increase their wellness scores, thus motivating them to manage those variables. Second, users can create optimized plans for improving their wellness scores based on the relative effects of each behavior variable. Linear functions are also already prevalent in existing wellness-related indexes (eg, [10,19,20]).

Missing-Data Handling Methods

Missing-data handling can be divided into 4 approaches, including complete case analysis, single imputation, multiple imputation, and joint model-based imputation (Figure 1). Complete case analysis excludes observations with missing values when analyzing data [21]. Single imputation produces only one complete data set by imputing missing values [22]. Multiple imputation creates multiple imputed data sets, applies a statistical analysis model to each one, and ultimately combines all analysis results to create an overall result [23]. Joint model-based imputation utilizes different distributions to model individuals with and without incomplete observations or directly models the relationship between the probability of a variable being missing and its missing value [24].

When selecting these 4 approaches, previous studies have used the missingness proportions and missingness mechanisms of data sets as major criteria for ensuring adequate selection for the data sets [25,26]. The missingness proportion is the ratio of the amount of missing values to the amount of missing and nonmissing values in the data set. The missingness mechanism can be divided into 3 types [14], including missing completely at random, missing at random, and missing not at random. First, missing completely at random is not related to any nonmissing or missing values in the data set. Second, missing at random entails that the missingness is independent of the missing values and is also conditional on nonmissing values. Third, the mechanism is missing not at random when the missingness depends on the missing values. As shown above, Figure 1 outlines the current recommendations for selecting adequate approaches based on both the missingness proportion and missingness mechanism.

Figure 1. Existing recommendations for missing data handling.



A panel data set of health behavior lifelogs is likely to contain 5% or more of incomplete observations with a missingness mechanism similar to missing completely at random. This property is attributed to a variety of random daily events that result in missing data. For example, the LWI development case presented by Kim et al [10] showed an 18% proportion of incomplete observations even though participants received interventions reminding them about the need to collect data. Participants also reported that random daily events resulted in missing or abnormal data, specifically including issues such as forgetting to wear a smartwatch or not entering data via the smartphone app, depleted smartwatch batteries, and data transmission errors. Based on the flowchart shown in Figure 1, 3 of the missing-data handling approaches may be implemented

for this property of a panel data set composed of health behavior lifelogs, including the complete case analysis, single imputation, and multiple imputation.

The 6 missing-data handling methods presented in Table 1 were selected to represent the complete case analysis, single imputation, and multiple imputation [21,27-31]. These methods are known to yield similar results given low missingness proportions (eg, less than 5% incomplete observations) [17,32]. The choice of missing-data handling method is known to become increasingly significant as the missingness proportion increases [17,32].

However, few previous studies have recommended which of the 6 missing-data handling methods are suitable for reducing coefficient biases according to the missingness proportion of a panel data set composed of health behavior lifelogs. This study filled that gap in the literature by comparatively evaluating the LWI coefficient biases of the 6 missing-data handling methods according to the missingness proportion of exactly such a panel data set.

Table 1. Representative missing-data handling methods applicable for LWI estimation.

Approach and method	Description
Complete case analysis	
Listwise deletion [21]	Excludes all observations with missing values to conduct analysis
Single imputation	
Mean imputation [21]	Imputes each missing value of a variable with the mean of observed values of the variable
k-nearest neighbor–based imputation [30]	Imputes each missing value of a variable based on the observed values of the k-nearest neighbors
Low-rank approximation–based imputation [29]	Predicts missing values as a linear combination of a small set of singular vectors
Multiple imputation	
Expectation maximization–based multiple imputation [28]	Draws imputed values from the multivariate normal distribution of the data set estimated by expectation–maximization; multiple imputed data sets are estimated by repeating the imputation and separately analyzed; analysis results are pooled into the final result
Predictive-mean matching–based multiple imputation [31]	Substitutes a missing value with a value randomly from complete observations, with regression-predicted values that are closest to the regression-predicted value for the missing value from the simulated regression model; multiple imputed data sets are estimated by repeating the imputation and separately analyzed; analysis results are pooled into the final result

Methods

Development Case: LWI for College Students

We previously developed an LWI for college students [10]. As a component of Onecare, a smart wellness service that supports individual-level health behavior monitoring for Korean college students based on their health behavior lifelogs, the index was developed to calculate daily wellness scores from lifelogs, thus intuitively showing users whether they were meeting recommended daily health behaviors. Daily wellness scores ranged from 0 to 100, indicating the worst and best conditions, respectively. The index was defined as a linear function consisting of 7 behavior variables (see Table 2), representing the critical health behaviors that Korean college students needed or wanted to manage. All such behaviors were identified based on expert interviews, target-user group discussions, and a literature review. As the daily wellness score was immeasurable during the development process, its proxy variable was also defined to estimate the index. More specifically, the proxy variable was the perceived score described in Table 2. Previous studies have regarded these types of perceived scores as valid measures for representing health. For example, patient-reported outcome measures are increasingly used in medical studies to represent psychometric self-evaluations of patient health [33,34].

Table 2. Variable descriptions.

Category and variable	Description (value meaning)
Behavior variable	
Breakfast (or Lunch or Dinner)	Student's self-rating of the day's breakfast (or lunch or dinner) based on nutrition (0: skip, 33: low, 66: medium, 100: high)
Exercise	Whether the student exercises or works out for more than 30 minutes during the day (0: no exercising, 100: exercising)
Step achievement	Percentage indicating a ratio that the total number of walking steps in the day reached 10,000
Sleep duration achievement	Percentage that the student's sleep duration reached 7 hours between 6 PM of the previous day and 6 PM of the current day
Golden time achievement	Percentage that the student slept during the golden time, which is 10 PM of the previous day to 2 AM of the current day
Proxy variable	
Perceived score	Score that the student determines by evaluating overall condition of their critical health behaviors over the day

To establish an intuitive scoring system, all behavior variables and the proxy variable were set to range from 0 (worst) to 100 (best) [35]. Each variable was defined to minimize user participation in the data collection process. From this perspective, data on the 3 behavior variables (ie, golden time achievement, sleep duration achievement, and step achievement) were automatically collected by smartwatches worn by students. Students also could easily record data on the remaining 5 variables through a smartphone app.

A 1-way random effects regression model was used to estimate the index coefficients:

where i, t, and k denote the ith student, day t, and kth behavior variable, respectively; y_{it} is the perceived score of the ith student on day t; β_0 and β_k are unknown coefficients; $x_{k,it}$ is the value of the kth behavior variable observed for the ith student on day t; μ_i the unobserved student-specific random effect of the ith student, is independent and identically distributed, $N(0, \sigma_\mu^2)$, and is independent of $x_{k,it}$; μ_i controls for the effects of student-specific heterogeneity on y_{it} and u_{it}, the error term, is independent and identically distributed, $N(0, \sigma_u^2)$.

This regression model was selected for 2 reasons. First, the index is a linear function. Second, the regression model was set to control for the unobserved student-specific random effects on the perceived score. Unobserved (or unmeasured) student-specific heterogeneity could exist in the regression model and thus influence the perceived score. For example, students may have different levels of interest in wellness, but these are unobserved in the regression model. However, those who are more interested in wellness may have higher standards for health behaviors, thus resulting in lower perceived scores. As the failure to control for such unobserved student-specific effects may produce misleading results [36], this was addressed by adding the effects to the regression model as μ_i.

The data set used to estimate the regression model was compiled by collecting data on the daily life activities of 41 students including 21 undergraduate (15 males and 6 females) and 20 graduate students (15 males and 5 females), all of whom were attending a university in Korea. Their age statistics were as follows: average of 24.7, maximum of 30, minimum of 19, and a standard deviation of 2.8. A total of 1148 observations were thus collected over a 28-day period (November 3-30, 2015). An observation consisted of 1 student's 1-day data for the 8 variables in the regression model.

Data preprocessing excluded the 264 observations including missing or abnormal values. Notably, students reported that these observations went through data collection problems (eg, forgetting to wear smartwatches, neglecting to enter data through the smartphone app, or depleting their smartwatch batteries). In this regard, they did not accurately reflect actual daily health behaviors of students. By excluding these observations, a panel data set comprised 884 complete observations from 41 students.

The LWI coefficients were estimated by fitting Eq (1) to the data set. Based on the estimated coefficients, the LWI was defined as a linear function consisting of the 7 following behavior variables: 0.151 × Breakfast + 0.163 × Lunch + 0.135 × Dinner + 0.135 × Exercise + 0.095 × Step achievement + 0.219 × Sleep duration achievement + 0.102 × Golden time achievement.

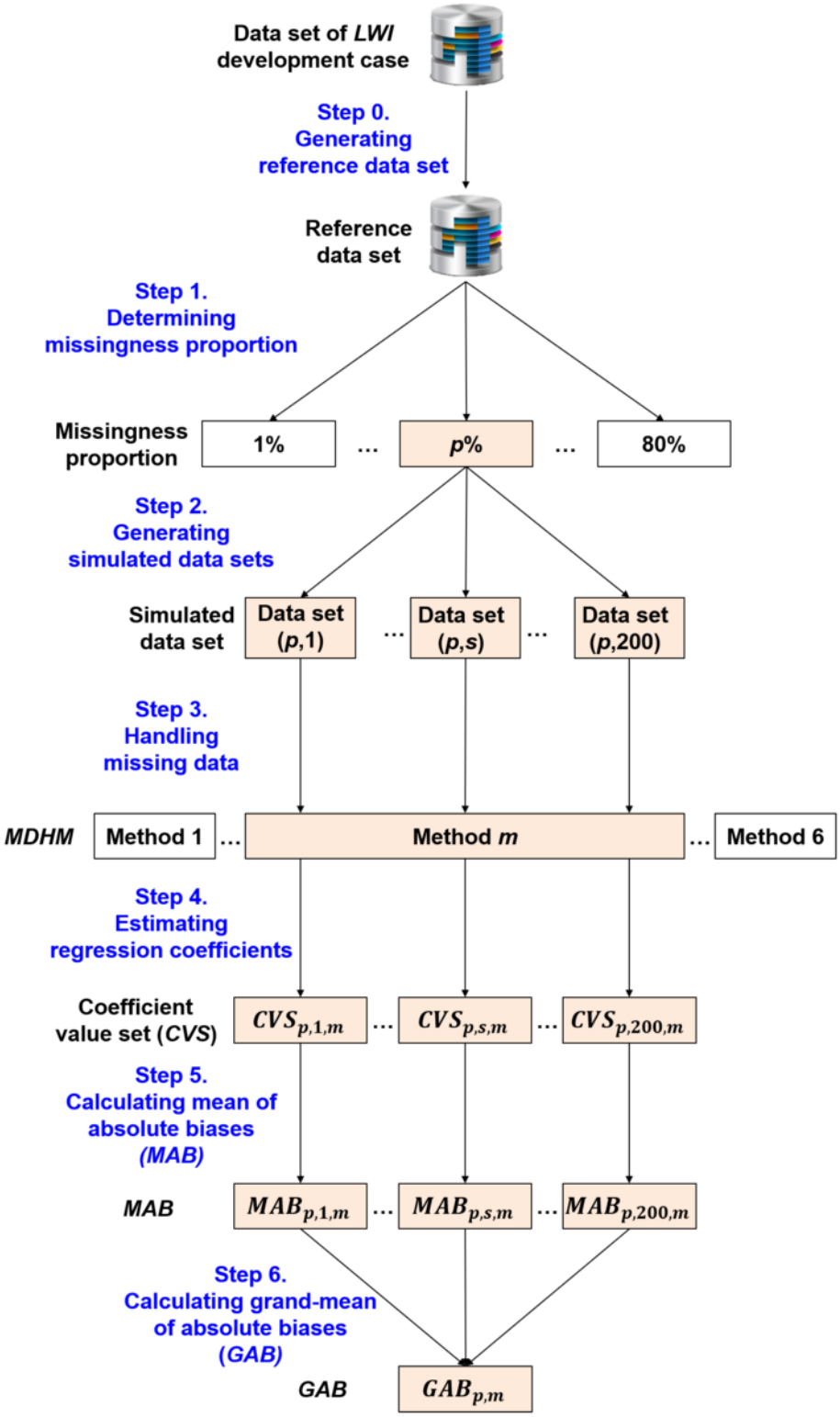
This study simulated the aforementioned LWI development case to evaluate biases regarding the regression coefficients that each of the 6 missing-data handling methods led to, as follows: the data set of the LWI development case was transformed into a reference data set that did not include any missing data; incomplete data sets were simulated by introducing missing data to the reference data set at various missingness proportions; the missing-data handling method changed all simulated data sets into complete data sets by handling their missing data; regression coefficients were estimated by fitting Eq (1) to the complete data sets; a bias measure of the missing-data handling method was calculated by comparing the estimated coefficient values with coefficient reference values. The coefficient reference values were estimated by fitting Eq (1) to the reference data set.

Overview

In this study, we conducted a simulation to calculate a bias measure for incremental missingness proportions for each of the 6 methods. The bias measure was referred to as the grand-mean of absolute biases (GAB). For each missingness proportion, GAB was used to compare the coefficient biases, thus determining which missing-data handling methods was superior.

Simulation steps are shown in Figure 2. In step 0, a reference data set was generated by transforming the data set from the development case. Steps 1 through 6 were then repeated for each missingness proportion, with each repetition calculating GAB for the 6 missing-data handling methods.

Figure 2. Research process.



Step 0: Generating the Reference Data Set

Step 0 was performed to generate a reference data set from the data set used in [10]. The reference data set included 884 observations of 41 students for 7 behavior variables and a perceived score variable. The descriptive statistics are provided in Table 3. Ranges of the variables were transformed from [x_{min}, x_{max}] to [z_{min}=0, z_{max}=1] using minimum-maximum normalization [37]:

This normalization is generally recommended as preprocessing for data-mining algorithms, including missing-data handling methods [38].

Table 3. Descriptive statistics of the data set for developing the LWI for college students and regression results for the reference data set.

Variable	Descriptive statistics		Regression results	
	Mean (SD)	Range	Estimate (SE)	*P* value
Perceived score	63.4 (15.9)	0-100	N/A[a]	N/A
Breakfast	24.2 (36.2)	0-100	0.097 (0.014)	<.001
Lunch	63.5 (32.3)	0-100	0.105 (0.013)	<.001
Dinner	75.5 (27.5)	0-100	0.088 (0.015)	<.001
Exercise	5.3 (22.4)	0-100	0.087 (0.019)	<.001
Step achievement	74.6 (28.6)	0-100	0.061 (0.015)	<.001
Sleep duration achievement	86.0 (19.3)	6.7-100	0.131 (0.021)	<.001
Golden time achievement	14.2 (25.1)	0-100	0.066 (0.018)	<.001
(Intercept)	N/A	N/A	0.305 (0.029)	<.001

[a]N/A: not applicable.

The reference data set also included 40 dummy variables and a time variable. Here, the dummy variables coded the 41 students, while the value of time variable was determined based on the dates the data were collected, that is, between the first and last days of the data collection period (November 3-30, 2015):

The resulting reference data set was 884×49 in dimension, as it contained all 884 observations mentioned above. Each observation included values for the 40 dummy variables, time variable, 7 behavior variables, and perceived score variable for a particular student on a given day. All variables ranged from 0 to 1.

Step 1: Determining the Missingness Proportion

In Step 1, the missingness proportion was selected to evaluate the 6 missing-data handling methods. The missingness proportion increased from 1% to 80% by 1%. An increment of 1% was sufficiently small to observe how the performance of each method changed according to the missingness proportion. Previous studies [39-41] have used larger increments, for example, Hasan et al [39] used 4 levels (10%, 20%, 30%, and 40%), Marshall et al [40] used 5 levels (5%, 10%, 25%, 50%, and 75%), and Song et al [41] used 4 levels (10%, 15%, 20%, and 30%) of missingness proportion for simulations to evaluate method performance.

We used a range up to 80% because one method continued to show outstanding performance for proportion above 60% and a missingness proportion of 80% was too high to estimate coefficients with low biases. If a data set had such a high missingness proportion in practice, then it may be preferable to collect another data set instead of using data from the initial data set.

Step 2: Generating the Simulated Data Sets

As shown in Figure 2, Step 2 generated 200 simulated data sets by randomly deleting the variable values from the reference data set according to missingness proportion *p*%. The random deletion implemented missing completely at random into the simulated data sets to reflect the missingness mechanism of a panel data set composed of health behavior lifelogs.

For proportion *p*%, there were many ways that missing data could be distributed across variables within the data set. Such a wide and varied distribution could affect missing-data handling method performance. However, there were too many possible missing data distributions to simulate all of them. Thus, this study randomly generated 200 simulated data sets for the missingness proportion, and then calculated the average of regression coefficient biases that each missing-data handling method produced across the 200 data sets. The average of each missing-data handling method was its performance measure (ie, GAB) for the missingness proportion. Similarly, Young and Johnson [42] had also calculated GABs of different missing-data handling methods across 200 simulated panel data sets in order to compare performance, although their work focused on multiple imputation and panel data sets related to family research.

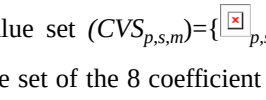
Step 3: Handling Missing Data

In Step 3, each of the 6 missing-data handling methods were applied to each of the 200 simulated data sets using R software (version 3.6.0). Listwise deletion and mean imputation were

implemented by several lines of R code to automatically delete incomplete observations and substitute a missing value for a variable with the mean of its observed values, respectively. k-nearest neighbor–based imputation used the knnImputation function in the DMwR package [30]. The number of nearest neighbors was the odd value close to the squared root of complete observations in each simulated data set [43]. The package softImpute [29] was utilized as a low-rank approximation–based imputation. Its maximum rank and lambda were determined based on "warm starts [29]." Expectation maximization–based multiple imputation and predictive-mean matching–based multiple imputation used Amelia II [28] and MICE [31] packages, respectively. The number of multiple imputations was set to 5, based on published recommendations [44].

As a result of this step, each of the listwise deletion, mean imputation, k-nearest neighbor–based imputation, and low-rank approximation–based imputation methods resulted in a complete data set. For expectation maximization–based and predictive-mean matching–based multiple imputations, there were 5 complete data sets.

Step 4: Estimating the Regression Coefficients

Eq (1) was fitted to each complete data set resulting from Step 3 using the plm package [45]. As a result, 8 coefficients (ie, β_k) were estimated for each complete data set. Each listwise deletion, mean imputation, k-nearest neighbor–based imputation, and low-rank approximation–based imputation contained a set of the 8 coefficient values for a simulated data set because each one resulted in a compete data set for the simulated data set in Step 3. Each expectation maximization–based and predictive-mean matching–based multiple imputation contained 5 sets of the 8 coefficient values for a simulated data set, which were pooled into a single set each, following rules established by Rubin [14]. For each method, the set of 8 coefficient values was defined as coefficient value set $(CVS_{p,s,m})=\{$  $\}$, where $CVS_{p,s,m}$ is the set of the 8 coefficient values that originated from the application of mth missing-data handling method to sth simulated data set of missing proportion p%; $_{p,s,m,k}$ is kth coefficient value in $CVS_{p,s,m}$; $p \in$ {1%, 2%,..., 80%}; $s \in$ {1, 2,…, 200}; and $m \in$ {listwise deletion,..., predictive-mean matching–based multiple imputation}.

Step 5: Calculating the Mean of Absolute Biases

Step 5 was performed to calculate a bias measure for each coefficient value set. Because a coefficient could have a certain amount of bias, each coefficient value set contained a total of 8 coefficient biases. The mean of absolute biases (MAB) was defined as a bias measure to calculate the average amount of the 8 coefficient biases for a given coefficient value set:

where  $_{p,s,m,k} \in CVS_{p,s,m}$; $\hat{a}_k$ is the reference value of  $_k$; $\hat{a}_k$ was estimated by fitting Eq (1) to the reference data set, as all simulated data sets were generated by deleting the missingness proportion p% of the reference data set. The estimate column in Table 3 provides the estimated values of $\hat{a}_k$. For missingness proportion p%, this step resulted in the 200 MABs of each missing-data handling method.

Step 6: Calculating the GAB

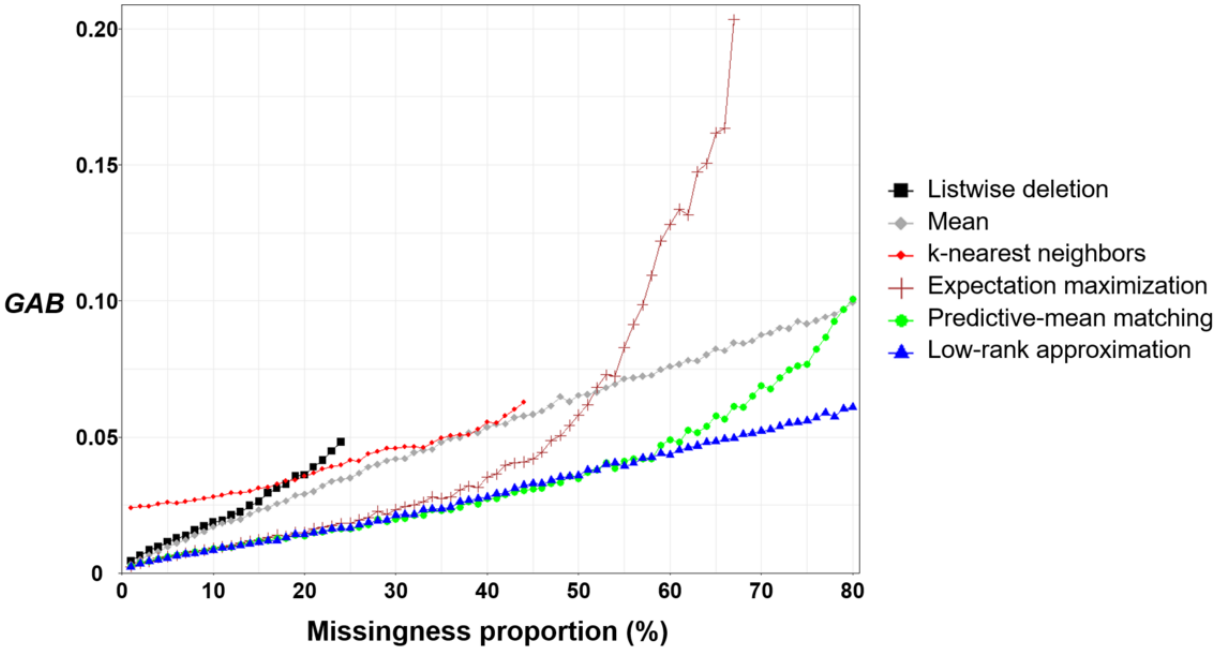
We combined the 200 MABs for each method to create a bias measure that represented the average of its coefficient biases over the 200 simulated data sets of missingness proportion p%. By following Young and Johnson [42], the bias measure was defined as the GAB:

A low GAB indicated that the missing-data handling method led to small coefficient biases across the 200 simulated data sets of the missingness proportion. The GAB was used as the criterion for evaluating method performance.

Results

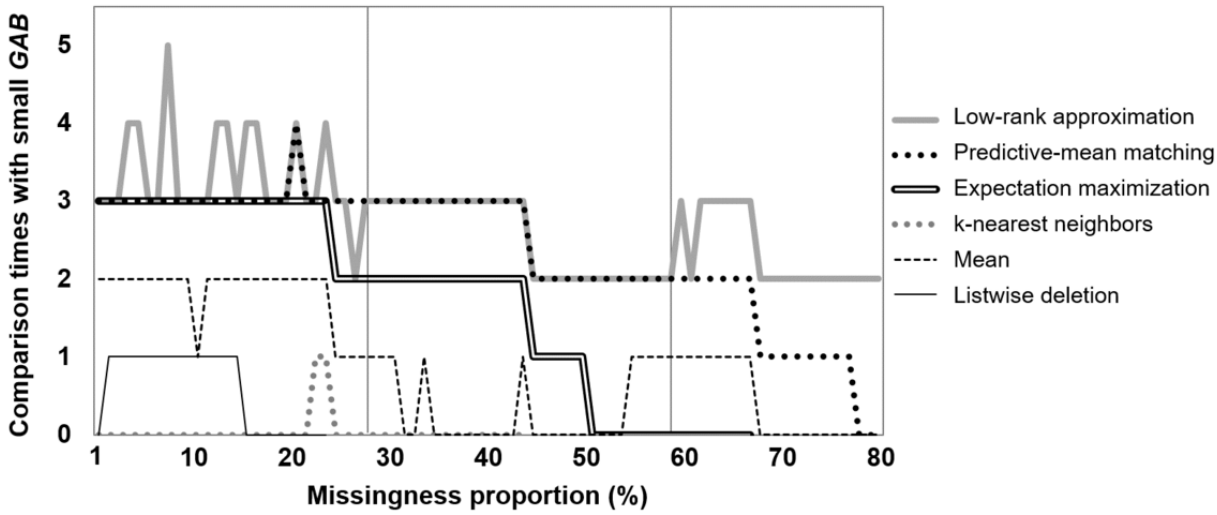
Figure 3 shows GABs for each missingness proportion. The listwise deletion, k-nearest neighbor–based imputation, and expectation maximization–based multiple imputation did not have GABs over missingness proportions of 24%, 44%, and 67%, respectively. Listwise deletion left too small number of complete observations to estimate the regression coefficients over missingness proportions of 24%. Both the k-nearest neighbor–based imputation and expectation maximization–based multiple imputation also failed to impute missing values over missingness proportions of 44% and 67%, respectively. The simulated data sets for these missingness proportions contained smaller numbers of complete observations than the minimum required for them to impute missing values.

Figure 3. GAB results.



Pairwise multiple comparison tests were conducted to statistically compare relative superiority among the 6 missing-data handling methods for each missingness proportion. The tests were conducted using Dunnett modified Tukey-Kramer pairwise multiple comparison at the .05 significance level [46]. Results provided the number of pairwise comparisons in which each missing-data handling method had statistically small GAB compared with all other missing-data handling methods for each missingness proportion. For interpretation purposes, a superior missing-data handling method will show the maximum number of pairwise comparisons with statistically small GAB (Figure 4). For example, the low-rank approximation–based imputation, predictive-mean matching–based multiple imputation, and expectation maximization–based multiple imputation were shown to be superior at a 1% missingness proportion (Figure 4).

Figure 4. Number of pairwise comparisons with statistically small GAB differences.



Different missing-data handling methods were shown to be superior depending on the missingness proportion. As shown in Figure 4, this included the low-rank approximation–based imputation, predictive-mean matching–based multiple imputation, and expectation maximization–based multiple imputation for the 1% to 30% missingness proportions, while the low-rank approximation–based imputation and predictive-mean matching–based multiple imputation were

superior for the 31% to 60% proportion, and only the low-rank approximation–based imputation was superior for proportions over 60%. These results are also shown in Table 4, which shows the sum of the pairwise comparison times with statistically small GAB for each missing-data handling method and missingness proportion. Listwise deletion, mean imputation, k-nearest neighbor–based imputation, expectation maximization–based multiple imputation, predictive-mean matching–based imputation, and low-rank approximation–based imputation achieved 15, 53, 2, 84, 91, and 99 as sums for the pairwise comparison times with statistically small GAB for 1% to 30% missingness proportions, respectively. The low-rank approximation–based imputation, predictive-mean matching–based multiple imputation, and expectation maximization–based multiple imputation were shown to be superior for these missingness proportions, with the low-rank approximation–based imputation revealing the maximum number (the predictive-mean matching–based and expectation maximization–based multiple imputations were also close to the maximum). The second and third rows of Table 4 show that the low-rank approximation–based imputation and predictive-mean matching–based multiple imputation were superior for the 30% to 60% missingness proportions, while only the low-rank approximation–based imputation was superior for over 60%.

Table 4. Sum of pairwise comparison times with statistically small GAB for each missing-data handling method and missingness proportion range.

Missingness proportion range	Listwise deletion	Mean imputation	k-nearest neighbor	Expectation–maximization	Predictive-mean matching	Low-rank approximation
1%-30%	15	53	2	84[a]	91[a]	99[a]
31%-60%	0	9	0	34	74[a]	75[a]
61%-80%	0	7	0	0	24	46[a]

[a]These methods had the best performance for the missingness proportion range.

Discussion

Principal Findings

The low-rank approximation–based imputation showed superior performance for 1% to 80% missingness proportions and has previously shown excellent performance with low-rank data sets [47]. In this context, low rank indicates that a data set can be approximated by a small subset of its singular vectors. Early studies [48,49] established strong theoretical guarantees about the perfect performance of low-rank approximation–based imputation for low-rank data sets without noise, with extensive research later supporting its superiority for low-rank data sets with noise [50-52]. These studies [48-52] suggest that the low-rank nature of the simulated data sets may be the primary reason that low-rank approximation–based imputation was shown to be superior in this study. In this regard, the low-rank property of the simulated data sets was investigated based on the chosen ranks for the low-rank approximation–based imputation to impute them. The rank of 13 was the maximum among the chosen ranks to impute all simulated data sets, while the maximum rank was much lower than the dimensions of the simulated data sets (ie, 884 × 49). It is therefore reasonable to assume that the low-rank nature of the simulated data sets is the primary reason that low-rank approximation–based imputation was shown to be superior.

Low-rank approximation–based imputation is also expected to perform well with other panel data sets comprising health behavior lifelogs, as previous studies [53,54] have verified that such data sets are generally low-rank. For instance, Eagle and Pentland [53] found that panel data sets comprising human behaviors were low-rank. They specifically proposed eigenbehaviors as principal components for panel data sets on human behaviors. The weighted sums of only 6 eigenbehaviors achieved more than 90% accuracy in reconstruction of a data set on the daily behaviors of 100 individuals for 400,000 hours. Furthermore, Saint Onge and Kreuger [54] found 7 distinct health lifestyle typologies for US adults in terms of 8 health behaviors, including sleep, physical activity, and alcohol intake. This result implied that panel data sets comprising health behaviors can be approximated by several typologies and are thus of a low-rank nature.

Both the expectation maximization–based and predictive-mean matching–based multiple imputations showed larger biases than the low-rank approximation–based imputation as the missingness proportion increased. Larger proportions increased the loss of information with missing values, which then increases uncertainty. Multiple imputation reflects such uncertainty in the standard errors of the estimates [14], with greater uncertainty resulting in larger standard errors for the estimates and larger coefficient biases [55].

In summary, the low-rank approximation–based imputation was the superior missing-data handling method for handling missing data when estimating a linear LWI with a panel data set comprising health behavior lifelogs, regardless of the missingness proportion.

Future Research

Three future research issues can improve and expand on this research. The first involves validating generalizability of the current research to nonlinear LWIs (eg, functions with polynomial or interaction variables and logistic functions). New LWI development cases can aim to develop nonlinear LWIs that this study did not cover. Thus, additional research is needed to establish the validity of our findings in regard to nonlinear LWIs.

The second issue involves the need to identify which health behavior-related covariates (eg, age, gender, and BMI) can enhance the performance of missing-data handling for LWI

estimation. While previous studies have already suggested several such covariates [56-58], additional covariates can enhance missing-data handling method performance. However, this study did not investigate these elements. Furthermore, few studies have identified covariates that can improve missing-data handling for panel data sets comprising health behavior lifelogs.

The third issue concerns the need to develop guidelines for predicting the size of bias in LWI coefficients for a certain missingness proportion of a given panel data set. In Figure 3, all missing-data handling methods showed increased coefficient biases as the missingness proportion increases. This suggests that missing-data handling methods can lead to large biases in LWI coefficients when missingness proportions are excessively large. Thus, a panel data set with a remarkably large missingness proportion requires careful attention to prevent excessively biased LWI coefficients. However, few previous studies have provided guidelines for predicting such biases according to the given missingness proportion. As shown in Figure 3, the low-rank approximation–based imputation exhibited linear growth in GAB as the missingness proportion increased. The slope of linear growth can be estimated through an experiment in which the change in GAB is calculated according to the unit change in the missingness proportion. The slope enables the prediction of GAB at a given missingness proportion. Such a guideline will help investigators decide whether the missingness proportion is acceptable for preventing highly biased coefficients of LWI. This requires additional research aimed at identifying relationships between biases and missingness proportions. Efforts are also needed to validate the generalizability of any guidelines.

Conclusion

A panel data set comprising health behavior lifelogs will likely contain a large amount of missing data due to various events. These missing data can result in LWI coefficient biases. While there are various methods for handling missing data, few previous studies have set out to determine which are the most effective for reducing LWI coefficient biases. This study comparatively evaluated 6 representative missing-data handling methods by simulating an existing LWI development case. Results suggested that low-rank approximation–based imputation was superior for reducing biases when estimating a linear LWI with a panel data set composed of health behavior lifelogs. This finding is expected to contribute to the reduction of coefficient biases in new development cases where linear LWIs are estimated with panel data.

Acknowledgments

This work was supported by the National Research Foundation of Korea grant funded by the Korean government (Ministry of Science and ICT; no. 2020R1C1C1014312).

Conflicts of Interest

None declared.

References

1. Market Research Future. Smart Wellness Market Research Report - Global Forecast 2023. Maharashtra, India: Market Research Future; 2018.
2. Grand View Research. mHealth App Market by Type (Fitness, Lifestyle Management, Nutrition & Diet, Women's Health, Healthcare Providers, Disease Management) and Segment Forecasts, 2014 – 2025. San Francisco, CA, USA: Grand View Research; 2017.
3. mHealth App Developer Economics 2016. Research2guidance. Berlin, Germany; 2016. URL: http://research2guidance.com/product/mhealth-app-developer-economics-2016/ [accessed 2020-08-06] [WebCite Cache ID 6lY0vJ78i]
4. 325,000 mobile health apps available in 2017 – android now the leading mHealth platform. Research2guidance. Berlin, Germany; 2017. URL: https://research2guidance.com/325000-mobile-health-apps-available-in-2017/ [accessed 2020-08-06] [WebCite Cache ID 71ZIAzZe7]
5. Luxton DD, June JD, Sano A, Bickmore T. Intelligent mobile, wearable, and ambient technologies for behavioral health care. In: Luxton DD, editor. Artificial Intelligence in Behavioral and Mental Healthcare. New York, NY, USA: Academic Press; 2015:137-162.
6. H-Jennings F, Clément M, Brown M, Leong B, Shen L, Dong C. Promote students' healthy behavior through sensor and game: a randomized controlled trial. Med Sci Educ 2016 May 3;26(3):349-355 [FREE Full text] [doi: 10.1007/s40670-016-0253-8]
7. Rodgers MM, Pai VM, Conroy RS. Recent advances in wearable sensors for health monitoring. IEEE Sensors J 2015;15(6):3119-3126. [doi: 10.1109/jsen.2014.2357257]
8. Ajami S, Teimouri F. Features and application of wearable biosensors in medical care. J Res Med Sci 2015;20(12):1208-1215 [FREE Full text] [doi: 10.4103/1735-1995.172991] [Medline: 26958058]
9. Lee J, Kim D, Ryoo HY, Shin BS. Sustainable wearables: wearable technology for enhancing the quality of human life. Sustainability 2016 May 11;8(5):466 [FREE Full text] [doi: 10.3390/su8050466]
10. Kim K, Kim K, Lim C, Heo J. Development of a lifelogs-based daily wellness score to advance a smart wellness service. Serv Sci 2018;10(4):408-422 [FREE Full text] [doi: 10.1287/serv.2018.0216]

11. Nardo M, Saisana M, Saltelli A, Tarantola S, Hoffman A, Giovannini E. Handbook On Constructing Composite Indicators: Methodology and User Guide. Paris, France: OECD Publishing; 2008.
12. Platt A, Outlay C, Sarkar P, Karnes S. Evaluating user needs in wellness apps. Int J Hum Comput Int 2016;32(2):119-131 [FREE Full text] [doi: 10.1080/10447318.2015.1099803]
13. Hsiao C. Panel data analysis—advantages and challenges. TEST 2007;16(1):1-22. [doi: 10.1007/s11749-007-0046-x]
14. Rubin DB. Multiple Imputation for Nonresponse In Surveys. New York, NY, USA: Wiley; 1987.
15. Schafer JL. Analysis of Incomplete Multivariate Data. London, UK: Chapman & Hall/CRC; 1997.
16. Dong Y, Peng CYJ. Principled missing data methods for researchers. Springerplus 2013;2(1):222 [FREE Full text] [doi: 10.1186/2193-1801-2-222] [Medline: 23853744]
17. Croninger RG, Douglas KM. Missing data and institutional research. In: Umbach PD, editor. Survey Research. Emerging Issues. New Directions for Institutional Research #127. San Fransisco, CA, USA: Jossey-Bass; 2005:33-50.
18. Belton V, Stewart T. Multiple Criteria Decision Analysis. Dordrecht, The Netherlands: Kluwer Academic Publishers; 2002.
19. Gallup. Gallup-Healthways Well-being Index: Methodology Report for Indexes. Washington, DC: Gallup; 2009.
20. Jung YS, Chae HG, Kim YW, Cho WD, Park RW, Han TH. Method for producing wellbeing life care index model in ubiquitous environment patent 1015555410000. Korean Intellectual Property Office. 2015 Sep 18. URL: http://engpat.kipris.or.kr/engpat/searchLogina.do?next=MainSearch#page1 [accessed 2020-11-30]
21. Little RJA, Rubin DB. Statistical Analysis with Missing Data. New York, NY, USA: John Wiley & Sons; 1987.
22. Nakagawa S. Missing data: mechanisms, methods, and messages. In: Fox GA, Negrete-Yankelevich S, Sosa VJ, editors. Ecological Statistics: Contemporary Theory and Application. Oxford, UK: Oxford University Press; 2015:81-105.
23. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009 Jun 29;338:b2393 [FREE Full text] [doi: 10.1136/bmj.b2393] [Medline: 19564179]
24. Galimard JE, Chevret S, Curis E, Resche-Rigon M. Heckman imputation models for binary or continuous MNAR outcomes and MAR predictors. BMC Med Res Methodol 2018 Aug 31;18(1):90 [FREE Full text] [doi: 10.1186/s12874-018-0547-1] [Medline: 30170561]
25. Adèr HJ, Mellenbergh GJ, Hand DJ. Advising on Research Methods: A Consultant's Companion. Huizen, The Netherlands: Johannes van Kessel Publishing; 2008.
26. Lodder P. To Impute or Not Impute: That's the Question. In: Mellenbergh JG, Adèr HJ. editors. Advising on Research Methods: Selected Topics. Huizen, The Netherlands: Johannes van Kessel Publishing; 2013.
27. Bertsimas D, Pawlowski C, Zhuo YD. From predictive methods to missing data imputation: an optimization approach. J Mach Learn Res 2017;18(1):7133-7171 [FREE Full text]
28. Honaker J, King G, Blackwell M. Amelia II: a program for missing data. J Stat Softw 2011;45(7):1-47. [doi: 10.18637/jss.v045.i07]
29. Mazumder R, Hastie T, Tibshirani R. Spectral regularization algorithms for learning large incomplete matrices. J Mach Learn Res 2010 Mar 01;11:2287-2322 [FREE Full text] [Medline: 21552465]
30. Torgo L. Data Mining Using R: Learning with Case Studies. Boca Raton, FL, USA: CRC Press; 2010.
31. van Buuren S, Groothuis-Oudshoorn K. MICE: multivariate imputation by chained equations in R. J Stat Soft 2011;45(3):1-67. [doi: 10.18637/jss.v045.i03]
32. Dettori JR, Norvell DC, Chapman JR. The sin of missing data: is all forgiven by way of imputation? Global Spine J 2018 Dec;8(8):892-894 [FREE Full text] [doi: 10.1177/2192568218811922] [Medline: 30560043]
33. Anthoine E, Moret L, Regnault A, Sébille V, Hardouin JB. Sample size used to validate a scale: a review of publications on newly-developed patient reported outcomes measures. Health Qual Life Outcomes 2014;12(126):1-10 [FREE Full text] [doi: 10.1186/s12955-014-0176-2] [Medline: 25492701]
34. Garrard L, Price LR, Bott MJ, Gajewski BJ. A novel method for expediting the development of patient-reported outcome measures and an evaluation of its performance via simulation. BMC Med Res Methodol 2015;15:77 [FREE Full text] [doi: 10.1186/s12874-015-0071-5] [Medline: 26419748]
35. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 2009;4(3):114-123.
36. Hospido L. Modelling heterogeneity and dynamics in the volatility of individual wages. J Appl Econ 2012;27(3):386-414 [FREE Full text] [doi: 10.1002/jae.1204]
37. Han J, Kamber M, Pei J. Data Mining Concepts and Techniques. San Francisco, CA, USA: Elsvier; 2006.
38. Al Shalabi L, Shaaban Z. Normalization as a preprocessing engine for data mining and the approach of preference matrix. 2006 Presented at: DepCos-RELCOMEX '06; 24-28 May 2006; Szklarska Poreba, Poland p. 207-214. [doi: 10.1109/depcos-relcomex.2006.38]
39. Hasan H, Ahmad S, Osman BM, Sapri S, Othman N. A comparison of model-based imputation methods for handling missing predictor values in a linear regression model: a simulation study. A comparison of model-based imputation methods for handling missing predictor values in a linear regression model: American Institute of Physics; 2017 Presented at: 24th National Symposium on Mathematical Sciences; 27-29 September 2016; Kuala Terengganu, Malaysia p. 060003-1-060003-8. [doi: 10.1063/1.4995930]

40. Marshall A, Altman DG, Royston P, Holder RL. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. BMC Med Res Methodol 2010 Jan 19;10:7 [FREE Full text] [doi: 10.1186/1471-2288-10-7] [Medline: 20085642]
41. Song Q, Shepperd M, Cartwright M. A short note on safest default missingness mechanism assumptions. Empir Softw Eng 2005;10(2):235-243. [doi: 10.1007/s10664-004-6193-8]
42. Young R, Johnson DR. Handling missing values in longitudinal panel data with multiple imputation. J Marriage Fam 2015 Mar;77(1):277-294 [FREE Full text] [doi: 10.1111/jomf.12144] [Medline: 26113748]
43. Jonsson P, Wohlin C. An evaluation of k-nearest neighbour imputation using Likert data. : IEEE Computer Society; 2004 Presented at: 10th International Symposium on Software Metrics; 11-17 September 2004; Chicago, IL, USA p. 108-118. [doi: 10.1109/metric.2004.1357895]
44. Schafer JL, Olsen MK. Multiple imputation for multivariate missing data problems: a data analyst's perspective. Multivariate Behav Res 1998;33:545-571 [FREE Full text] [doi: 10.1207/s15327906mbr3304_5]
45. Croissant Y, Millo G. Panel data econometrics in R: the plm package. J Stat Softw 2008;27(2):1-43. [doi: 10.18637/jss.v027.i02]
46. Dunnett CW. Pairwise multiple comparisons in the unequal variance case. J Am Stat Assoc 1980;75(372):796-800. [doi: 10.1080/01621459.1980.10477552]
47. Mao X, Chen SX, Wong RKW. Matrix completion with covariate information. J Am Stat Assoc 2019;114(525):198-210 [FREE Full text] [doi: 10.1080/01621459.2017.1389740]
48. Candès EJ, Recht B. Exact matrix completion via convex optimization. Found Comput Math 2009;9(6):717-772. [doi: 10.1007/s10208-009-9045-5]
49. Recht B. A simpler approach to matrix completion. J Mach Learn Res 2011;12:3413-3430 [FREE Full text]
50. Candes EJ, Plan Y. Matrix completion with noise. Proc IEEE 2010 Jun;98(6):925-936. [doi: 10.1109/jproc.2009.2035722]
51. Koltchinskii V, Lounici K, Tsybakov AB. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. Ann Stat 2011;39(5):2302-2329. [doi: 10.1214/11-aos894]
52. Rohde A, Tsybakov AB. Estimation of high-dimensional low-rank matrices. Ann Stat 2011;39(2):887-930. [doi: 10.1214/10-aos860]
53. Eagle N, Pentland AS. Eigenbehaviors: identifying structure in routine. Behav Ecol Sociobiol 2009;63(7):1057-1066. [doi: 10.1007/s00265-009-0739-0]
54. Saint Onge JM, Krueger PM. Health lifestyle behaviors among U.S. adults. SSM Popul Health 2017 Dec;3:89-98 [FREE Full text] [doi: 10.1016/j.ssmph.2016.12.009] [Medline: 28785602]
55. Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. Int J Epidemiol 2019;48(4):1294-1304 [FREE Full text] [doi: 10.1093/ije/dyz032] [Medline: 30879056]
56. Greene GW, Schembre SM, White AA, Hoerr SL, Lohse B, Shoff S, et al. Identifying clusters of college students at elevated health risk based on eating and exercise behaviors and psychosocial determinants of body weight. J Am Diet Assoc 2011;111(3):394-400. [doi: 10.1016/j.jada.2010.11.011] [Medline: 21338738]
57. Olson JS, Hummer RA, Harris KM. Gender and health behavior clustering among U.S. young adults. Biodemography Soc Biol 2017;63(1):3-20 [FREE Full text] [doi: 10.1080/19485565.2016.1262238] [Medline: 28287308]
58. Ruiz-Palomino E, Giménez-García C, Ballester-Arnal R, Gil-Llario MD. Health promotion in young people: identifying the predisposing factors of self-care health habits. J Health Psychol 2020;25(10-11):1410-1424. [doi: 10.1177/1359105318758858] [Medline: 29468900]

Abbreviations

CVS: coefficient value set
GAB: grand-mean of absolute biases
LWI: lifelogs-based wellness index
MAB: mean of absolute biases

Edited by C Lovis; submitted 22.05.20; peer-reviewed by A Benis, B Loo Gee, C Reis; comments to author 19.08.20; revised version received 10.10.20; accepted 18.10.20; published 17.12.20.

Please cite as:
Kim KH, Kim KJ
Missing-Data Handling Methods for Lifelogs-Based Wellness Index Estimation: Comparative Analysis With Panel Data
JMIR Med Inform 2020;8(12):e20597
URL: http://medinform.jmir.org/2020/12/e20597/
doi:10.2196/20597
PMID:33331831

Original Paper

Detecting Miscoded Diabetes Diagnosis Codes in Electronic Health Records for Quality Improvement: Temporal Deep Learning Approach

Sina Rashidian[1], MSc; Kayley Abell-Hart[2], BSc; Janos Hajagos[2], PhD; Richard Moffitt[2], PhD; Veena Lingam[2], MD; Victor Garcia[2], MD; Chao-Wei Tsai[2], MD; Fusheng Wang[1,2], PhD; Xinyu Dong[1], MSc; Siao Sun[3], MSc; Jianyuan Deng[2], MPhil; Rajarsi Gupta[2], MD, PhD; Joshua Miller[4], MD; Joel Saltz[2], MD, PhD; Mary Saltz[2], MD

[1]Department of Computer Science, Stony Brook University, Stony Brook, NY, United States

[2]Department of Biomedical Informatics, Renaissance School of Medicine at Stony Brook, Stony Brook, NY, United States

[3]Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY, United States

[4]Department of Medicine, Renaissance School of Medicine at Stony Brook, Stony Brook, NY, United States

Corresponding Author:
Sina Rashidian, MSc
Department of Computer Science
Stony Brook University
2212 Computer Science
Stony Brook, NY, 11794
United States
Phone: 1 631 632 8470
Email: srashidian@cs.stonybrook.edu

Abstract

Background: Diabetes affects more than 30 million patients across the United States. With such a large disease burden, even a small error in classification can be significant. Currently billing codes, assigned at the time of a medical encounter, are the "gold standard" reflecting the actual diseases present in an individual, and thus in aggregate reflect disease prevalence in the population. These codes are generated by highly trained coders and by health care providers but are not always accurate.

Objective: This work provides a scalable deep learning methodology to more accurately classify individuals with diabetes across multiple health care systems.

Methods: We leveraged a long short-term memory-dense neural network (LSTM-DNN) model to identify patients with or without diabetes using data from 5 acute care facilities with 187,187 patients and 275,407 encounters, incorporating data elements including laboratory test results, diagnostic/procedure codes, medications, demographic data, and admission information. Furthermore, a blinded physician panel reviewed discordant cases, providing an estimate of the total impact on the population.

Results: When predicting the documented diagnosis of diabetes, our model achieved an 84% F1 score, 96% area under the curve–receiver operating characteristic curve, and 91% average precision on a heterogeneous data set from 5 distinct health facilities. However, in 81% of cases where the model disagreed with the documented phenotype, a blinded physician panel agreed with the model. Taken together, this suggests that 4.3% of our studied population have either missing or improper diabetes diagnosis.

Conclusions: This study demonstrates that deep learning methods can improve clinical phenotyping even when patient data are noisy, sparse, and heterogeneous.

(JMIR Med Inform 2020;8(12):e22649) doi:10.2196/22649

KEYWORDS

electronic health records; diabetes; deep learning

Introduction

The widespread adoption of an electronic health record (EHR) has generated large amounts of data, providing an opportunity for clinical phenotyping to identify patients with characteristics of interest [1,2]. Analyzing these rich EHR data has many potential uses such as predicting mortality, defining cohorts, evaluating health care policy, and driving health care finance that affect patient care, revenue, and performance evaluation. The ability to use large amounts of clinical data to discover or validate information is of particular interest for research studies as well as clinical practice [3]. Over the years, disease phenotyping methods from EHR data have evolved from traditional manually developed rule-based analysis for concept curation such as eMERGE and PheKB [4-6] to statistical and traditional machine learning techniques [7-9], and more recently, deep learning techniques which offer better performance while reducing the need for data preprocessing and feature engineering [10-12]. However, EHR data are often incomplete, inaccurate, fragmented, and heterogeneously structured, reflecting the challenges of real-world information gathering, extraction, and interpretation [1,4,13].

Being able to accurately predict diseases in a population could lead to targeted clinical interventions [14], while applying predictive models retrospectively may identify patients with incorrect or missing diagnoses, documentation, or billing codes. We chose diabetes mellitus for such phenotyping applications because it is a highly prevalent disease with heterogeneous presentations and objective diagnostic criteria. In the United States, more than 34 million people have diabetes, and 1 out of 4 people are undiagnosed. Diabetes is associated with many serious medical comorbidities such as heart disease and stroke, as well as high costs of medical care [15]. Previous efforts assessing errors in diagnosis, classification, and disease coding in patients with diabetes using clinical trial data and primary care data have shown that significant errors from misdiagnosis, misclassification, and miscoded patient data are associated with worse therapeutic outcomes [16-21].

In this study, we aim to characterize clinical phenotype for diabetes using data available at the time of discharge by using a generalizable sequential-based deep learning method. We employ all laboratory results, medications, demographic data, and other admission data such as days from prior admission or duration of current visit for each patient. We also include diagnostic codes and procedure codes from all encounters except the most recent one, which is the target to predict. The goal of this work is to train a model that can identify diseases—diabetes in this study—for each patient based on all available information. This model has the potential to merge into hospital real-time monitoring systems for flagging patients, potentially improving patient care and EHR documentation quality, among countless other downstream benefits.

In recent years, there are many interesting studies applying deep learning methods on EHR data. Using dense neural networks (DNNs) for finding patients at high risk of mortality [22], discovering characteristic patterns of physiology [23], representing patient data for machine learning purposes [14], improving coding accuracy in EHR data [24,25], taking advantage of recurrent neural networks (RNNs) for predicting future diagnosis codes and clinical events [26-30], forecasting kidney transplant success [31], early detection of heart failure [32], using bidirectional RNNs for medical event detection [33], and combining convolutional neural networks and RNNs for improving patient representation [34] are just a few of these inspiring projects. There are extensive survey papers exploring and categorizing recent projects based on methods and their goal [35,36]. However, in most of them limited EHR data elements are used, patients have extensive background information, and the goal is to predict what is recorded in a future visit for a patient. The real-world disease classification problem in a health system is different and requires a more general and scalable model that can make robust predictions using all data elements.

Our study offers the following key contributions: (1) A minimally curated, real-world data set for model training is employed, where about 76% of patients had only 1 encounter, reflecting the incomplete and fragmented nature of EHR data. (2) Data from 5 different health care facilities in the United States are combined to show the generalizability of the model, avoiding overfitting on a single facility, and demonstrating the capability of neural networks to learn from data with diverse and complex structures. (3) Precise measurements are provided to show improvements and performance of this model. (4) A thorough validation with a panel of clinicians is conducted to adjudicate the clinical phenotype from longitudinal data in cases where the model disagreed with the documented disease coding. (5) The total impact on the population for patients is calculated with both improper and missed diagnosis codes in their EHR data.

Methods

Data Set Description

We obtained data from the CERNER Health Facts database, a large multi-institutional deidentified database derived from EHR data and administrative systems. The database has 599 facilities. For this study, we extracted inpatient encounter data from the 5 acute care facilities with the most inpatient discharges from January 1, 2016, to December 31, 2017. The extracted encounters all have ICD-10 (International Classification of Diseases, 10th edition) diagnosis codes and at least one laboratory test. Table 1 summarizes general information including statistics on the reported cases of diabetes in each facility and the mean number of medications and unique laboratory tests. Population demographic information is summarized in Multimedia Appendix 1.

Table 1. General and diabetes-related inpatient statistics in facilities studied.

Facility ID	131	143	384	898	1157
Number of encounters	62,318	60,175	45,390[a]	55,444	52,080
Number of patients	41,854	38,657	31,387[a]	38,953	36,336
Mean number of ICD[b] codes	13.74	19.07	3.61[a]	13.77	10.22
Percentage of encounters with diabetes	34.55	27.82	9.93[a]	23.34	25.91
Mean number of medications	21.56	12.58	16.43	1.71[a]	8.79
Percentage with metformin	3.06	0.76	1.53	0.08[a]	0.28
Mean number of unique laboratories	56.72	49.94	26.89[a]	48.73	61.7
Percentage with hemoglobin A1c (HbA1c)	28.91	13.20	0.00[a]	24.16	19.47

[a]ICD: International Classification of Diseases.

[b]The lowest value in each row.

EHRs from different facilities usually have various formats, structures, and may not be directly interoperable. For this reason, demographic information, laboratory results, diagnosis codes, procedure codes, and medications were mapped to the Observational Health Data Sciences and Informatics (OHDSI) Common Data Model (version 5.3; vocabulary release on October 2, 2018), a standard data model for observational health studies [37-39]. Clinical notes are not available in the database and were not included in this study.

Laboratory Tests

There are 2 major challenges for representing laboratory values. First, laboratory tests may be performed multiple times in a single encounter. Second, there are a large number of test types, which form a huge sparse matrix with many missing values. We proposed 2 approaches to represent laboratory tests: (1) We used statistical summaries including median, max, min, total count, and the values of the first and last instance of a laboratory test for each single encounter. A laboratory test is ordered by a physician if there are concerns that it may not be normal. Therefore, when it is unavailable the value is either expected to be normal or its result is reflected in other available features clearly. For these laboratory tests we used median imputation for filling missing values. It is worth mentioning that we explored more complicated imputation methods as well, including MICE [40], Soft-Impute [41], and SVD-Impute [42]. However, these methods did not provide distinct improvement and took much more computation power. (2) We counted the number of laboratory values that were classified as “high,” “low,” “within the range,” or “normal,” “abnormal,” and “unspecified” according to standards provided by each facility. In a case that a laboratory value is not recorded, these values are exactly 0, thus imputation is not needed. However, ranges for some features are undefined in the EHR system that makes it necessary to have numerical values as well.

Diagnosis and Procedure Codes

Because the model is designed to use all information available at the time of discharge, codes from past encounters are included. However, the codes for the current encounter are the target to be predicted and not included in the input feature matrix. Codes are represented as binary values for each ICD code in the data set.

Medications

Medications were mapped from National Drug Codes to RxNorm’s Concept Unique Identifiers using mappings associated with the OHDSI-controlled vocabularies. Total counts of drug exposure and per inpatient visit were added to the feature matrix.

Demographic/Personal Information

We also included age, weight, height, race, ethnicity, and gender from the data set. For categorical features (race, ethnicity, and gender), we added them to the feature matrix through one-hot encoding.

Derived Features

We further derived calculated features, such as the number of days from the latest previous encounter, days hospitalized, and the facility IDs represented with a one-hot encoding scheme.

Target

The ICD-10-CM codes that defined clinical diabetes were derived from the Clinical Classification Software (CCS) [43] categories 49, 50, and 186. We excluded conditions that do not clearly fit the clinical definition of diabetes as a chronic disease, such as “unspecified hyperglycemia,” “prediabetes,” and “gestational diabetes.” All ICD codes under the mentioned CCS codes were included except conditions specified in Multimedia Appendix 2.

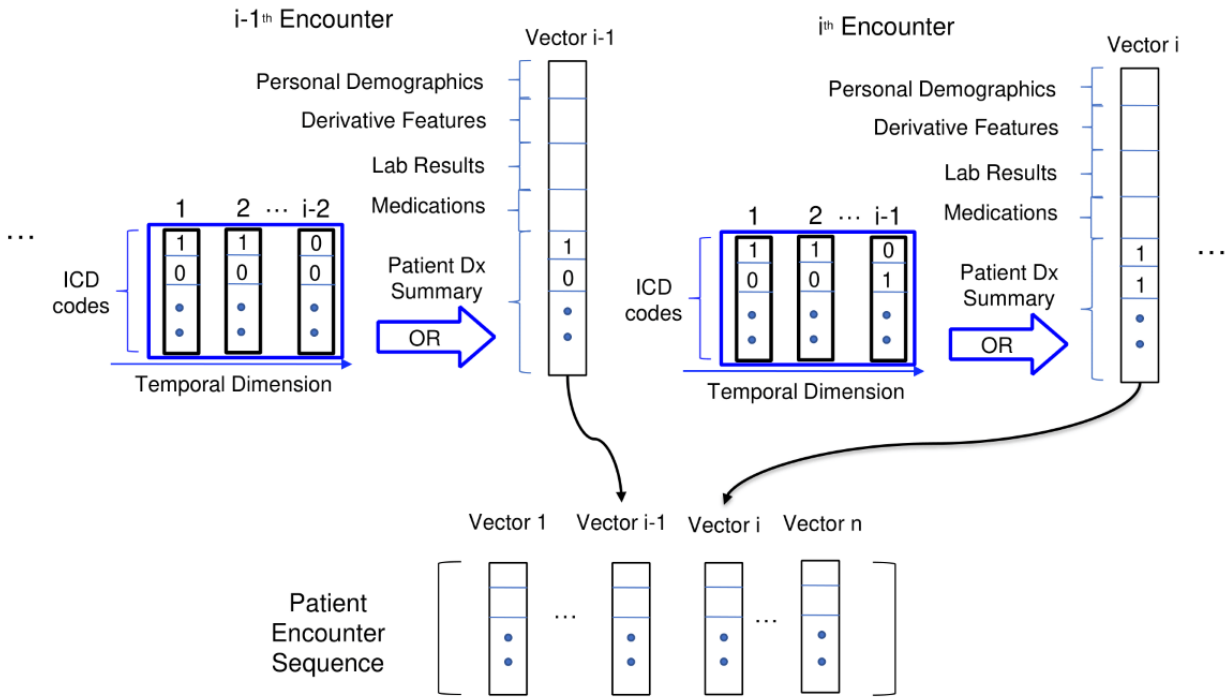
In order to reduce the sparsity of the feature matrix and remove features that are not available or relevant to the target disease, we only kept features with a nonzero value and appearing in at least 5% of positive cases in the training set.

Data Vectorization

As previously mentioned, diagnosis and procedure codes from the final encounter are the prediction goal and are not included in the input to the model. We combined the target diagnosis codes using CCS categorization to create a binary value for the presence of disease. For each encounter i, we created a vector

v_i by concatenating laboratories, medications, demographics from the *i*th encounter, and the accumulated ICD code presence value from prior encounters: "0" for no presence and "1" for at least one instance as shown in Figure 1. The idea behind this "or" operation is to represent the history data as physicians would review them, that is, focusing on the presence or absence of diseases in the patient history. Thus, in mimicking our stated goal, these vectors hold the information that would be available at the time of discharge when the codes must be determined.

Figure 1. Feature matrix construction from patient encounters. All information from the ith encounter, except ICD codes, was combined with ICD codes from prior encounters to build a slice in the sequence. Dx: diagnosis code; ICD: International Classification of Diseases.



Machine Learning/Deep Learning–Based Predictive Models

We employed both nonsequential and sequential models in this study. In nonsequential models, the order of input features does not matter and does not distinguish features based on their temporal occurrence. On the contrary, sequential models care about which features happened when and they are designed to capture temporal information.

Nonsequential Models

We took 2 traditional machine learning approaches, random forest and logistic regression, as baselines for comparison. Furthermore, we took advantage of DNNs which are powerful classifiers and have been widely used in previous studies [22-25]. The main advantage of DNNs over other machine learning methods is the capability to learn patterns more effectively from large data sets with numerous features without the need for feature selection.

Sequential Models

Because of the inherently sequential nature of a patient's medical history, we expect that sequential models should outperform those that do not consider the order of inputs. RNNs are among the most powerful tools for prediction and classification when there is a sequence of data leading to the result. Standard or vanilla RNNs face vanishing and exploding gradients in back-propagation during the training phase as the longer the sequence of inputs grows, the longer and more unstable the chain of gradients becomes to calculate. Because of these problems, we leveraged long short-term memory (LSTM) [44] and gated recurrent unit (GRU) [45] which use "forget" and "update" elements to selectively turn off portions of the model, effectively reducing the parameter space during each training step. Furthermore, we added additional dense layers after the output of recurrent layers [46,47]. We call these models LSTM-DNN and GRU-DNN, respectively.

Model Training

As is the case in almost any phenotyping study, the data set is imbalanced, with only 21.59% of cases positive for diabetes. In this subsection, we briefly go through techniques and parameters used to increase prediction power and avoid overfitting. These parameters also make it possible to replicate experiments. Data set is normalized (mean = 0, variance = 1) before training to improve performance and stability. The data set (combination data of 5 acute care facilities that were mentioned earlier) was split using stratified random sampling to 80% for the training set and 20% for the test set. The training and test sets were the same for training and evaluation of all models.

Traditional Machine Learning Methods

For the logistic regression model, we used L2 regularization (1.0) and in the random forest model we limited the tree maximum depth to 30. The class weights for both models were adjusted inversely proportional to class frequencies to give more weight to the minor class (positive cases).

Neural Networks

For the DNN model both L2 regularization (0.0002) and dropout (with rate 0.45) [48] were used. We applied weight balancing with log proportion as the prevalence ratio (2.22) to calculate loss in each epoch. We employed mini batches (2048) which are more computationally efficient, use less memory, and are generally more robust as they avoid local minima in optimization steps [49]. After hyper-tuning using 12.5% of training data for cross validation, the best model was trained with mean squared error loss, Adam optimizer [50], Xavier uniform initializer [51], tanh activation functions in hidden layers, and a sigmoid activation function in the output layer. The dense network consists of 4 hidden layers (512, 512, 512, 512) and the recurrent networks have 2 recurrent layers (512, 512) (LSTM/GRU) and 2 dense layers (512, 512). All have a single neuron output. Adding additional embedding layers did not improve models' performances.

As the search space is enormous, we had 2 steps for finding the best parameters. First, we fixed all parameters except one and hyper-tuned that specific parameter. After reaching a short list of candidates for each variable, we used grid search on all of them to find the best combination. The network configuration was reached by extensive hyperparameter search over the following parameters: activation functions (tanh, relu, selu), loss functions (mean squared error, mean absolute error, binary cross entropy), optimizers (Adam, sgd), batch size (512, 1024, 2048), L2 regularization (0.001, 0.01, 0.10, 0.05, 1, 2, 10), dropout rate (from 0 to 0.80 every 0.05), number of layers (1 to 7), and various number of neurons in each layer (different combinations of powers of 2 as expected to be faster while using GPU nodes).

Review Panel Validation Method

Identifying inaccuracy in coded disease states was a major motivation for the study, and we hypothesized that a well-trained model would be accurate even when some diagnosis codes in the training set were incorrectly coded. Because it is impossible to evaluate this goal using existing diagnosis codes which themselves can be flawed, we asked 3 board-certified practicing physicians to review cases where the model contradicted the documented diagnosis. In this experiment, experts were provided with the same information as the model, including all demographic information, laboratory results, and medications as well as event timelines for inpatient encounters. Furthermore, the experiment was performed in a blinded manner—experts did not have any knowledge of the diagnosis from either the model prediction or EHR documentation. We believe this experiment can shed light into the usefulness of such a model for flagging cases in hospital systems.

Results

Experimental Setup

For training and testing the deep learning models, we used Keras framework [52] backed by Tensorflow [53] and the scikit-learn library [54]. The training was performed on a NVIDIA Tesla V100 GPU with 640 Tensor Cores.

Performance of Phenotyping Diabetes According to EHR Labels

We compared our sequential-based model with other models based on a variety of metrics. As the data set is imbalanced (21.59% positive cases), accuracy cannot be a distinguishing metric among models. The area under the receiver operating characteristic curve (AUROC) also can be misleading in these data sets. The F1 score (harmonic mean of precision and recall) and area under the precision–recall curve (AUPRC) are more suitable metrics for this purpose [22,55,56]. In this project it is important to capture the majority of patients, therefore a model with high recall is desired. The precision for 0.80 recall is also measured and reported in Table 2. As shown in Figure 2, the LSTM-DNN model outperforms other models in both the AUROC and AUPRC curves. We excluded GRU-DNN in Figure 2 as it is close to the LSTM-DNN model.

Table 2. Methods performance comparison.

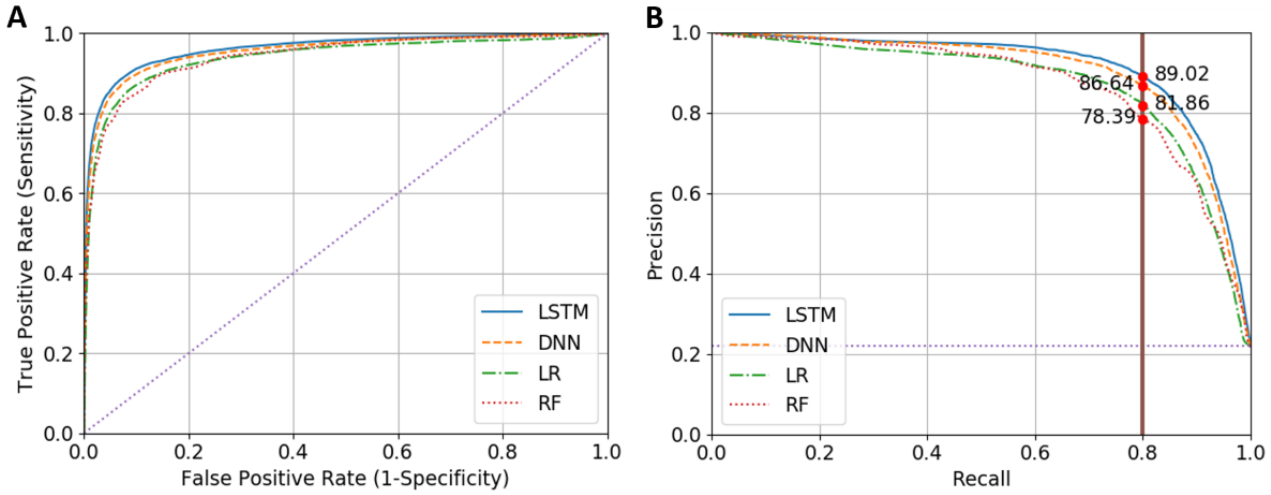
Models	Accuracy	Precision @0.8 recall	F1 score	AUPRC[a]	AUROC[b]
LSTM-DNN	93.04[c]	89.02[c]	84.30[c]	91.18[c]	96.15[c]
GRU-DNN	92.80	88.04	83.92	90.65	95.77
DNN	92.49	86.64	83.17	90.10	95.49
Logistic regression	90.77	81.86	80.03	86.47	93.96
Random forest	90.95	78.39	76.78	86.86	94.17

[a]AUPRC: area under the precision–recall curve.

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

[c]Numbers for the best method.

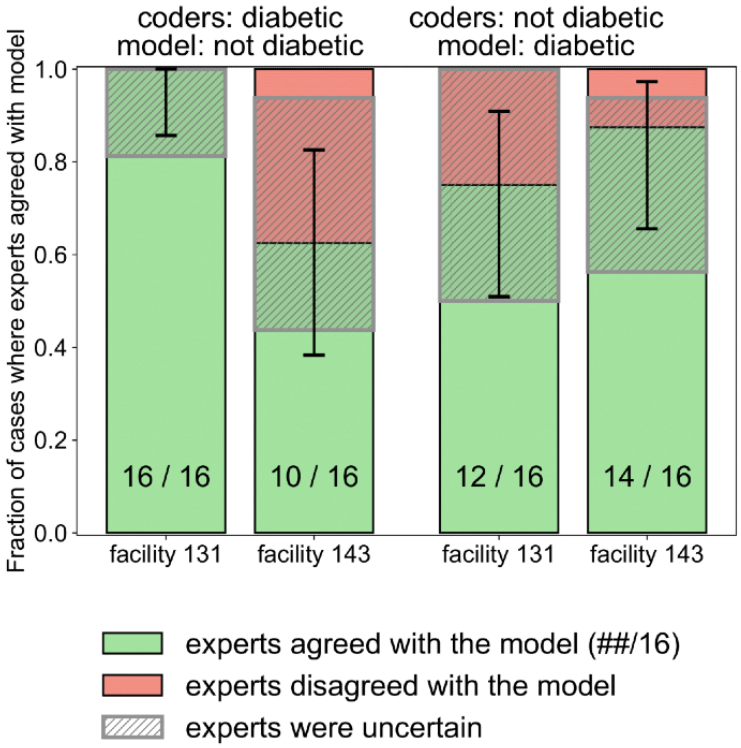
Figure 2. ROC and PR curves for all models. (A) ROC curve. Diagonal dotted purple line is the performance of random model. (B) PR curve. The vertical solid line shows precision of different models for achieving 0.8 recall. Straight dotted purple line is the performance of random model. DNN: dense neural network; LR: logistic regression; LSTM: long short-term memory; PR: precision-recall; RF: random forest; ROC: receiver operating-characteristic curve.



Review Panel Validation Results

For analyzing discordant cases where the model disagreed with what was recorded in the EHR, we performed a blinded review with a group of domain experts including at least three board-certified practicing physicians for each case review. For facilities 131 and 143, we used 32 sampled cases per facility where the model-predicted diagnosis was discordant with the EHR and the model had a high confidence (sigmoid output >0.83 or <0.17). We asked the review panel to answer 2 questions: (1) does the patient have diabetes; and (2) what is their confidence level? (high or low). In 52 out of the 64 cases, the panel's conclusions agreed with the results from the model's prediction. In 37 out of the 39 cases with which the panel had high confidence, the model's prediction (output of the LSTM-DNN model) was consistent with the panel's conclusion. Generally, the panel would have low confidence when there was insufficient evidence from the data to support a conclusion. The evaluation results are shown in Figure 3.

Figure 3. Expert review of cases where the model prediction disagreed with coded diagnosis. The error bars were 5% confidence intervals calculated from the beta binomial distribution.



Through expert validation, we can provide a conservative estimate of how frequently a case flagged by the model for review would result in a correction at each facility. We calculated the range of the total population that would be

potentially impacted for each facility with lower and upper bounds. The lower bound considers only the model's high confidence interval—probability of more than 0.83 or less than 0.17 for positive and negative labeling, respectively, on sigmoid output—and the upper bound is for all predictions made by the model. Each value bound is multiplied by the probability of the model being correct, as derived from the expert validation (Figure 3). This final value is the percentage of the impacted population. In facility 131, we estimated that 1.25%-3.03% of the total population were missing a diabetes-related diagnosis code, and 1.65%-2.98% were improperly labeled as having diabetes. These numbers varied for facility 143, where there were 1.61%-3.73% missing a diabetes code and 1.12%-1.89% improperly labeled. Taking the mean of the intervals across facilities, we estimate that the error rate is 4.3% across these facilities. This suggests a considerable impact of this misclassification that can impact patients, hospitals, health systems, and payers.

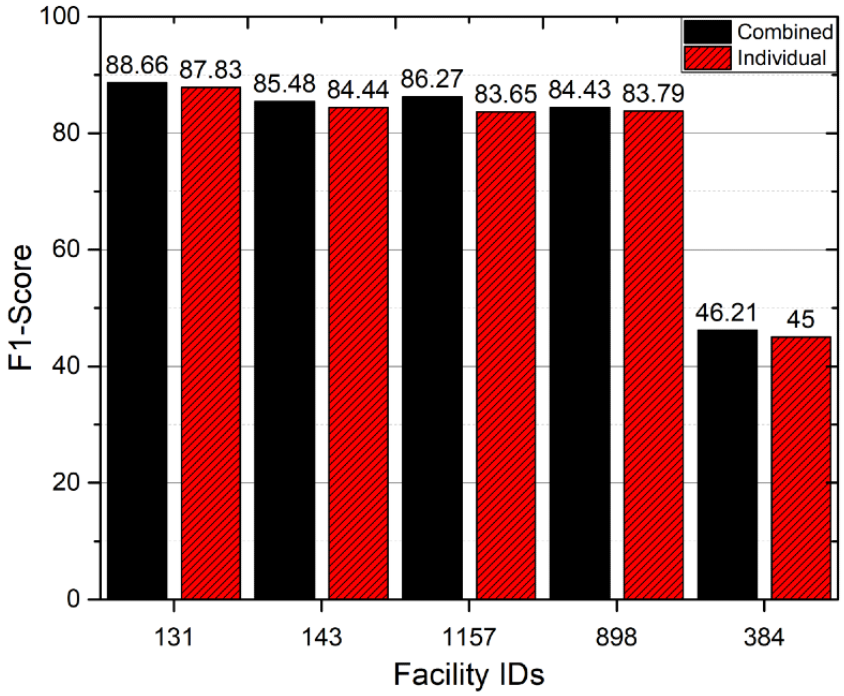
These results demonstrate that when the model prediction contradicts the coders, the model is most often correct even for patients with several past encounters. From 32 cases with background information in 24 cases, experts agreed with the model. This suggests that a deep learning model trained from EHR data, which are often noisy, is capable of phenotyping and flagging cases for further review.

Multiple Facilities Versus Single-Facility Models

In our study, we found that different facilities used different coding schemes for laboratory tests and medications. As a result, the diversity of features is higher than we had anticipated. For instance, blood glucose measurement, a standard test in diabetes, has a variety of names and Logical Observation Identifiers Names and Codes (LOINC) across facilities. Facilities reported "Glucose lab," "Glucose [Mass/volume] in Blood," "Glucose [Mass/volume] in Body fluid," "Glucose [Mass/volume] in Blood by Test strip manual," "Glucose; blood, reagent strip," and "Glucose finger stick." Each name has a different LOINC, making automated consolidation difficult. This problem exists in other data elements such as medications, where brand names, generic names, and various similar formulations are recorded. For this reason, a model trained on a single facility will not perform as well on another facility. Our goal was to develop a generalizable model that could perform well on all facilities independent of features available. Because features might vary widely, we proposed to collate all information from all facilities, and created 1 data set containing all features rather than manual or automatic merging of them (the data set we used for previous experiments). We were curious to see how does a model trained on this "combined" data set would differ from a model trained on just a single facility? From one perspective, with more data the model should perform better. However, as coding patterns and features vary significantly between facilities, this combination can end up misleading the model.

We trained a model for each facility using the exact same steps we did previously using our best architecture (LSTM-DNN). As shown in Figure 4, the results from the combined model are very similar to those from the single facility–based models. In another experiment, we repeated the training on the combined data set without including facility IDs, and the results were almost the same. This suggests that the model trained on the combined data has the capability to learn all different patterns and can benefit from this approach.

Figure 4. Comparison of F1 scores on single facility-based models and multifacility combined model.



Facility 384 showed very low performance, and we suspect that this is due to poor data quality and feature availability. We found that facility 384 reported fewer laboratory tests than other facilities (Table 1). It also lacked some laboratory tests essential to diabetes diagnosis, such as hemoglobin A1c. The facility also reported far fewer diagnoses per patient, including much lower

prevalence of diabetes, even though it recorded metformin (a typical drug used for diabetes treatment) as much as other facilities. Thus, we believe that the low performance was due to the low availability of vital training features and the poor quality of recorded diagnosis codes. Interestingly, the model appeared to be resilient to other data problems, such as the paucity of medication data in facility 898.

Limitations of Rule-Based Models

The traditional approach for phenotyping is based on a predefined set of rules and steps to determine whether a patient has a specific disease. To compare with such rule-based methods, we followed the steps in the eMERGE project [46]. Because of the lack of required data elements such as family history of diabetes and counts of dates that the patient had face-to-face outpatient clinic encounters, the performance of this algorithm was not ideal on our data set. For 75.28% of the patients, the results from the method were undecided and no final decision could be made. Another major limitation of such rule-based methods is the need for constant updates for new ICD codes, laboratory codes, and medications. Even after mapping and updating codes to current ICD-10, the method would often fail and detect only obvious cases and discard uncertain cases. As a result, it was not possible to make a reasonable comparison between models' performances and the eMERGE criteria.

Discussion

Principal Findings

Our study demonstrates the successful identification of patient phenotypes using a deep learning model trained on heterogenous, minimally curated data. The model identifies a noticeable subset of potential coding errors in instances when patients are either improperly labeled as having or not having diabetes and is able to avoid errors arising from missing clinician documentation or sporadic coder errors. Given that the data were mapped to the OHDSI data model, the model is independent of facility-specific data representations and could be adopted by different health care systems based on normalization using OHDSI.

For much of the work on phenotyping, there is a presumption that the documented EHR diagnosis codes represent ground truth. However, human error can result in improper classification of a patient's comorbidities and true illness severity. The motivation for this work was to detect and reclassify individuals in whom the wrong diagnosis was assigned at the time of discharge from the hospital, a fact that drives the development of such phenotyping algorithms. Our efforts can be used to flag discordant records for human review, leading to more accurate patient and population characterization. This strategy can be used to guide coders at the time of discharge to re-evaluate charts detected by the algorithm, with more directed attention to the potential missed diagnosis.

To validate the simulation of operational deployment of such a model, we used a double-blinded physician review panel to review the discordant cases where the model prediction was in contrary to the documented diagnosis. From this review, we not only captured the panel's diagnosis but also the confidence level of their decision. During the review, the experts felt that some cases were too complex or needed more data for a model to classify correctly. Despite this, our panel and algorithm agreed on the final diagnosis among 81.25% of cases when the algorithm was confident in its prediction. In a real health system, this would equate to an anticipated 4 corrections to the coding for every 5 cases flagged by the model for further review. This is estimated to impact about 2.4% of a facility's entire population missing a diabetes code that should be present, and about 1.9% of the population who were given the code of diabetes when it should not have been present. This suggests that our methodology is highly promising for improving clinical decision support to flag possibly missing or improper ICD classifications.

Limitations

This work could benefit from expert validation at larger scale, which would result in a more accurate estimation of the effect on the population. As patients' background information was very limited in this study, we did not expect significant difference using other methods such as attention-based models; however, they can be beneficial where more background data are available. Moreover, we are collaborating with the diabetes care group of our network hospitals to incorporate our prediction model into a pilot study.

Conclusions

As research continues to advance the capabilities of predictive algorithms to medicine, we demonstrate a successful application of deep learning methodology bridging the gap at the intersection of computer science and clinical medicine.

We can classify a disease state in patients using a generalizable model that is deployable in institutions adopting the OHDSI standard. Our sequential deep learning–based model outperformed both traditional machine learning and nonsequential DNN as shown earlier. Results proved that the deep learning model can capture patterns for phenotyping from a high-dimension feature space without hand-crafted feature engineering. The findings also provide insights into how to build a framework/workflow using real-world EHR data for enhancing operations in real-world health care organizations, especially in applications to clinical intervention, documentation and billing, as well as quality improvement. The success of such disease prediction models can also benefit academic and translational research, as a faster and more refined disease phenotyping process allows researchers to better refine their study cohorts and minimize bias or confounding variables. Most importantly, one cannot understate the potential impact to patient care and clinical outcomes afforded by this approach to diagnostic validation and case ascertainment.

Acknowledgments

This work was partially supported by T32GM127253 for the Scholars in BioMedical Sciences Training Program (K.A-H).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Population demographic statistics.
[DOCX File , 15 KB - medinform_v8i12e22649_app1.docx]

Multimedia Appendix 2
Excluded ICD-10 codes from CCS 49, 50, 186.
[DOCX File , 15 KB - medinform_v8i12e22649_app2.docx]

References

1. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013 Jan 01;20(1):117-121 [FREE Full text] [doi: 10.1136/amiajnl-2012-001145] [Medline: 22955496]
2. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inform Assoc 2013 Dec 01;20(e2):e206-e211 [FREE Full text] [doi: 10.1136/amiajnl-2013-002428] [Medline: 24302669]
3. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. Annu Rev Biomed Data Sci 2018 Jul 20;1(1):53-68 [FREE Full text] [doi: 10.1146/annurev-biodatasci-080917-013315] [Medline: 31218278]
4. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014 Mar 01;21(2):221-230 [FREE Full text] [doi: 10.1136/amiajnl-2013-001935] [Medline: 24201027]
5. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med 2013 Oct;15(10):761-771 [FREE Full text] [doi: 10.1038/gim.2013.72] [Medline: 23743551]
6. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc 2016 Dec;23(6):1046-1052 [FREE Full text] [doi: 10.1093/jamia/ocv202] [Medline: 27026615]
7. Carroll RJ, Eyler AE, Denny JC. Naïve Electronic Health Record phenotype identification for Rheumatoid arthritis. AMIA Annu Symp Proc 2011;2011:189-196 [FREE Full text] [Medline: 22195070]
8. Li D, Simon G, Chute CG, Pathak J. Using association rule mining for phenotype extraction from electronic health records. AMIA Jt Summits Transl Sci Proc 2013;2013:142-146 [FREE Full text] [Medline: 24303254]
9. Chen Y, Ghosh J, Bejan CA, Gunter CA, Gupta S, Kho A, et al. Building bridges across electronic health record systems through inferred phenotypic topics. J Biomed Inform 2015 Jun;55:82-93 [FREE Full text] [doi: 10.1016/j.jbi.2015.03.011] [Medline: 25841328]
10. Ayala Solares JR, Diletta Raimondi FE, Zhu Y, Rahimian F, Canoy D, Tran J, et al. Deep learning for electronic health records: A comparative review of multiple deep neural architectures. J Biomed Inform 2020 Jan;101:103337. [doi: 10.1016/j.jbi.2019.103337] [Medline: 31916973]
11. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. $\mathtt {Deepr}$: A Convolutional Net for Medical Records. IEEE J Biomed Health Inform 2017 Jan;21(1):22-30. [doi: 10.1109/JBHI.2016.2633963] [Medline: 27913366]
12. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J Biomed Health Inform 2018 Sep;22(5):1589-1604 [FREE Full text] [doi: 10.1109/JBHI.2017.2767063] [Medline: 29989977]
13. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 01;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
14. 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]
15. Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2020. Atlanta, GA: Centers for Disease Control and Prevention; 2020. URL: https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf [accessed 2020-11-28]

16. de LS, Khunti K, Belsey J, Hattersley A, van VJ, Gallagher H, et al. A method of identifying and correcting miscoding, misclassification and misdiagnosis in diabetes: a pilot and validation study of routinely collected data. Diabet Med 2010 Feb;27(2):203-209. [doi: 10.1111/j.1464-5491.2009.02917.x] [Medline: 20546265]
17. Farmer A, Fox R. Diagnosis, classification, and treatment of diabetes. BMJ 2011 Jun 09;342(jun09 4):d3319. [doi: 10.1136/bmj.d3319] [Medline: 21659368]
18. Sadek AR, van Vlymen J, Khunti K, de Lusignan S. Automated identification of miscoded and misclassified cases of diabetes from computer records. Diabet Med 2012 Mar;29(3):410-414. [doi: 10.1111/j.1464-5491.2011.03457.x] [Medline: 21916978]
19. de Lusignan S, Sadek N, Mulnier H, Tahir A, Russell-Jones D, Khunti K. Miscoding, misclassification and misdiagnosis of diabetes in primary care. Diabet Med 2012 Feb;29(2):181-189. [doi: 10.1111/j.1464-5491.2011.03419.x] [Medline: 21883428]
20. Seidu S, Davies MJ, Mostafa S, de Lusignan S, Khunti K. Prevalence and characteristics in coding, classification and diagnosis of diabetes in primary care. Postgrad Med J 2014 Jan 13;90(1059):13-17. [doi: 10.1136/postgradmedj-2013-132068] [Medline: 24225940]
21. Mata-Cases M, Mauricio D, Real J, Bolíbar B, Franch-Nadal J. Is diabetes mellitus correctly registered and classified in primary care? A population-based study in Catalonia, Spain. Endocrinología y Nutrición (English Edition) 2016 Nov;63(9):440-448. [doi: 10.1016/j.endoen.2016.10.005]
22. Avati A, Jung K, Harman S, Downing L, Ng A, Shah NH. Improving palliative care with deep learning. BMC Med Inform Decis Mak 2018 Dec 12;18(Suppl 4):122 [FREE Full text] [doi: 10.1186/s12911-018-0677-8] [Medline: 30537977]
23. Che Z, Kale D, Li W, Bahadori MT, Liu Y. Deep computational phenotyping. In: 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY: Association for Computing Machinery (ACM); 2015 Presented at: KDD '15: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2015; Sydney, Australia p. 1-10 URL: http://www-scf.usc.edu/~dkale/papers/che_kale-kdd2015-deep_phenotyping.pdf [doi: 10.1145/2783258.2783365]
24. Rashidian S, Hajagos J, Moffitt R, Wang F, Noel KM, Gupta RR, et al. Deep Learning on Electronic Health Records to Improve Disease Coding Accuracy. AMIA Jt Summits Transl Sci Proc 2019;2019:620-629 [FREE Full text] [Medline: 31259017]
25. Rashidian S, Hajagos J, Moffitt R, Wang F, Dong X, Abell-Hart K, et al. Disease phenotyping using deep learning: A diabetes case study. 2020. URL: https://arxiv.org/abs/1811.11818 [accessed 2020-11-28]
26. Choi E, Schuetz A, Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. Proceedings of Machine Learning for Healthcare 2016. 2016. URL: http://proceedings.mlr.press/v56/Choi16.pdf [accessed 2020-11-28]
27. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Liu PJ, et al. Scalable and accurate deep learning for electronic health records. arXiv. 2018. URL: https://arxiv.org/abs/1801.07860 [accessed 2020-11-28]
28. Pham T, Tran T, Phung D, Venkatesh S. Predicting healthcare trajectories from medical records: A deep learning approach. J Biomed Inform 2017 May;69:218-229 [FREE Full text] [doi: 10.1016/j.jbi.2017.04.001] [Medline: 28410981]
29. Pham T, Tran T, Phung D, Venkatesh S. DeepCare: A Deep Dynamic Memory Model for Predictive Medicine. In: Bailey J, Khan L, Washio T, Dobbie G, Huang J, Wang R, editors. Advances in Knowledge Discovery and Data Mining. PAKDD 2016. Lecture Notes in Computer Science, vol 9652. Cham, Switzerland: Springer; Apr 12, 2016.
30. Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent neural networks. 2016. URL: http://zacklipton.com/media/papers/learning-to-diagnose-Lipton-Kale2015.pdf [accessed 2020-11-28]
31. Esteban C, Staeck O, Yang Y, Tresp V. Predicting Clinical Events by Combining Static and Dynamic Information Using Recurrent Neural Networks. arXiv. 2016 Feb 08. URL: arXiv:1602.02685 [accessed 2020-11-28]
32. 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]
33. Jagannatha AN, Yu H. Bidirectional RNN for Medical Event Detection in Electronic Health Records. Proc Conf 2016 Jun;2016:473-482 [FREE Full text] [doi: 10.18653/v1/n16-1056] [Medline: 27885364]
34. Ma T, Xiao C, Wang F. Health-ATM: A deep architecture for multifaceted patient health record representation and risk prediction. Philadelphia, PA: SIAM Publications URL: https://epubs.siam.org/doi/pdf/10.1137/1.9781611975321.30 [accessed 2020-11-28]
35. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2018 Nov 27;19(6):1236-1246 [FREE Full text] [doi: 10.1093/bib/bbx044] [Medline: 28481991]
36. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc 2018 Oct 01;25(10):1419-1428 [FREE Full text] [doi: 10.1093/jamia/ocy068] [Medline: 29893864]
37. Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, et al. Characterizing treatment pathways at scale using the OHDSI network. Proc Natl Acad Sci U S A 2016 Dec 05;113(27):7329-7336 [FREE Full text] [doi: 10.1073/pnas.1510502113] [Medline: 27274072]
38. Park RW. Sharing Clinical Big Data While Protecting Confidentiality and Security: Observational Health Data Sciences and Informatics. Healthc Inform Res 2017 Jan;23(1):1-3 [FREE Full text] [doi: 10.4258/hir.2017.23.1.1] [Medline: 28261525]

39. Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. Proc Natl Acad Sci U S A 2018 Mar 13;115(11):2571-2577 [FREE Full text] [doi: 10.1073/pnas.1708282114] [Medline: 29531023]
40. Buuren SV, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. J. Stat. Soft 2011;45(3). [doi: 10.18637/jss.v045.i03]
41. Mazumder R, Hastie T, Tibshirani R. Spectral regularization algorithms for learning large incomplete matrices. J Mach Learn Res 2010;11:2287-2322 [FREE Full text]
42. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics 2001 Jun 01;17(6):520-525. [doi: 10.1093/bioinformatics/17.6.520] [Medline: 11395428]
43. Healthcare Cost and Utilization Project (HCUP) Agency for Healthcare Research and Quality R, MD. Beta Clinical Classifications Software (CCS) for ICD-10-CM/PCS Healthcare Cost and Utilization Project (HCUP). 2019. URL: https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp [accessed 2020-11-28]
44. 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]
45. Chung J, Gulcehre C, Cho KH, Bengio Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv. 2014 Dec 11. URL: https://arxiv.org/abs/1412.3555 [accessed 2020-11-28]
46. Zhao R, Wang D, Yan R, Mao K, Shen F, Wang J. Machine Health Monitoring Using Local Feature-Based Gated Recurrent Unit Networks. IEEE Trans. Ind. Electron 2018 Feb;65(2):1539-1548. [doi: 10.1109/tie.2017.2733438]
47. Ren L, Cheng X, Wang X, Cui J, Zhang L. Multi-scale Dense Gate Recurrent Unit Networks for bearing remaining useful life prediction. Future Generation Computer Systems 2019 May;94:601-609. [doi: 10.1016/j.future.2018.12.009]
48. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res 2014 Jan;15(1):1929-1958 [FREE Full text]
49. Ruder S. An overview of gradient descent optimization algorithms. arXiv. 2016 Sep 15. URL: http://arxiv.org/abs/1609.04747 [accessed 2020-11-28]
50. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv. 2014 Dec 22. URL: arXiv:1412.6980v9 [accessed 2020-11-28]
51. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. 2010 Presented at: Thirteenth International Conference on Artificial Intelligence and Statistics; 13-15 May 2010; Sardinia, Italy p. 249-256 URL: http://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf
52. Keras. URL: https://keras.io/ [accessed 2020-11-28]
53. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Google Brain. TensorFlow: A system for large-scale machine learning. In: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16). 2016 Nov Presented at: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16); November 2–4, 2016; Savannah, GA, URL: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf
54. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12:2825-2830 [FREE Full text]
55. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning (ICML '06). New York, NY: ACM; 2006 Jun Presented at: 23rd International Conference on Machine Learning (ICML '06); June 25-29, 2006; Pittsburgh, PA p. 233-240. [doi: 10.1145/1143844.1143874]
56. Boyd K, Santos Costa V, Davis J, Page CD. Unachievable Region in Precision-Recall Space and Its Effect on Empirical Evaluation. Proc Int Conf Mach Learn 2012 Dec 01;2012:349 [FREE Full text] [Medline: 24350304]

Abbreviations

AUPRC: area under the precision–recall curve
AUROC: area under receiver operating-characteristic curve
CCS: Clinical Classification Software
DNN: dense neural network
EHR: electronic health record
GRU: gated recurrent unit
LOINC: Logical Observation Identifiers Names and Codes
LSTM: long short-term memory
OHDSI: Observational Health Data Sciences and Informatics
RNN: recurrent neural network

Edited by C Lovis; submitted 19.07.20; peer-reviewed by Y Ye, G Lim; comments to author 13.09.20; revised version received 24.09.20; accepted 27.09.20; published 17.12.20.

Please cite as:
Rashidian S, Abell-Hart K, Hajagos J, Moffitt R, Lingam V, Garcia V, Tsai CW, Wang F, Dong X, Sun S, Deng J, Gupta R, Miller J, Saltz J, Saltz M
Detecting Miscoded Diabetes Diagnosis Codes in Electronic Health Records for Quality Improvement: Temporal Deep Learning Approach
JMIR Med Inform 2020;8(12):e22649
URL: http://medinform.jmir.org/2020/12/e22649/
doi:10.2196/22649
PMID:33331828

Original Paper

User Experience of a Chatbot Questionnaire Versus a Regular Computer Questionnaire: Prospective Comparative Study

Mariska E te Pas[1], MD; Werner G M M Rutten[2], PhD; R Arthur Bouwman[1,3], MD, PhD; Marc P Buise[1], MD, PhD

[1]Anesthesiology Department, Catharina Hospital, Eindhoven, Netherlands

[2]Game Solutions Lab, Eindhoven, Netherlands

[3]Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, Netherlands

Corresponding Author:
Mariska E te Pas, MD
Anesthesiology Department
Catharina Hospital
Michelangelolaan 2
Eindhoven, 5623 EJ
Netherlands
Phone: 31 627624857
Email: mariska.t.pas@catharinaziekenhuis.nl

Abstract

Background: Respondent engagement of questionnaires in health care is fundamental to ensure adequate response rates for the evaluation of services and quality of care. Conventional survey designs are often perceived as dull and unengaging, resulting in negative respondent behavior. It is necessary to make completing a questionnaire attractive and motivating.

Objective: The aim of this study is to compare the user experience of a chatbot questionnaire, which mimics intelligent conversation, with a regular computer questionnaire.

Methods: The research took place at the preoperative outpatient clinic. Patients completed both the standard computer questionnaire and the new chatbot questionnaire. Afterward, patients gave their feedback on both questionnaires by the User Experience Questionnaire, which consists of 26 terms to score.

Results: The mean age of the 40 included patients (25 [63%] women) was 49 (SD 18-79) years; 46.73% (486/1040) of all terms were scored positive for the chatbot. Patients preferred the computer for 7.98% (83/1040) of the terms and for 47.88% (498/1040) of the terms there were no differences. Completion (mean time) of the computer questionnaire took 9.00 minutes by men (SD 2.72) and 7.72 minutes by women (SD 2.60; *P*=.148). For the chatbot, completion by men took 8.33 minutes (SD 2.99) and by women 7.36 minutes (SD 2.61; *P*=.287).

Conclusions: Patients preferred the chatbot questionnaire over the computer questionnaire. Time to completion of both questionnaires did not differ, though the chatbot questionnaire on a tablet felt more rapid compared to the computer questionnaire. This is an important finding because it could lead to higher response rates and to qualitatively better responses in future questionnaires.

(JMIR Med Inform 2020;8(12):e21982) doi:10.2196/21982

KEYWORDS

chatbot; user experience; questionnaires; response rates; value-based health care

Introduction

Questionnaires are routinely used in health care to obtain information from patients. Patients complete these questionnaires before and after a treatment, an intervention, or a hospital admission. Questionnaires are an important tool which provides patients the opportunity to voice their experience in a safe fashion. In turn, health care providers gather information that cannot be picked up in a physical examination. Through the use of patient-reported outcome measures (PROMs), the patient's own perception is recorded, quantified, and compared to normative data in a large variety of domains such as quality of life, daily functioning, symptoms, and other aspects of their health and well-being [1,2]. To enable the usage of data delivered by the PROMs for the evaluation of services, quality

of care, and also outcome for value-based health care correctly, respondent engagement is fundamental [3].

Subsequently, adequate response rates are needed for generalization of results. This implies that maximum response rates from questionnaires are desirable in order to ensure robust data. However, recent literature suggests that response rates of these PROMs are decreasing [4,5].

From previous studies, it is clear that factors which increase response rates include short questionnaires, incentives, personalization of questionnaires as well as repeat mailing strategies or telephone reminders [6-9]. Additionally, it seems that the design of the survey has an effect on response rates. Conventional survey designs are often perceived as dull and unengaging, resulting in negative respondent behavior such as speeding, random responding, premature termination, and lack of attention. An alternative to conventional survey designs is chatbots with implemented elements of gamification, which is defined as the application of game-design elements and game principles in nongame contexts [10].

A chatbot is a software application that can mimic intelligent conversation [11]. The assumption is that by bringing more fun and elements of gamification in a questionnaire, response rates will subsequently rise.

In a study comparing a web survey with a chatbot survey the conclusion was that the chatbot survey resulted in higher-quality data [12]. Patients may also feel that chatbots are safer interaction partners than human physicians and are willing to disclose more medical information and report more symptoms to chatbots [13,14].

In mental health, chatbots are already emerging as useful tools to provide psychological support to young adults undergoing cancer treatment [15]. However, literature investigating the effectiveness and acceptability of chatbot surveys in health care is limited. Because a chatbot is suitable to meet the aforementioned criteria to improve response rates of questionnaires, this prospective preliminary study will focus on the usage of a chatbot [13,16]. The aim of this study is to measure the user experience of a chatbot-based questionnaire at the preoperative outpatient clinic of the Anesthesiology Department (Catharina Hospital) in comparison with a regular computer questionnaire.

Methods

Recruitment

All patients scheduled for an operation who visit the outpatient clinic of the Anesthesiology Department (Catharina Hospital) complete a questionnaire about their health status. Afterward there is a preoperative intake consultation with a nurse or a doctor regarding the surgery, anesthesia, and risks related to their health status. The Medical Ethics Committee and the appropriate Institutional Review Board approved this study and the requirement for written informed consent was waived by the Institutional Review Board.

We performed a preliminary prospective cohort study and included 40 patients who visited the outpatient clinic between September 1, 2019, and October 31, 2019. Because of the lack of previous research on this topic and this is a preliminary study, we discussed the sample size (N=40) with the statistician of our hospital and this was determined to be clinically sufficient. Almost all patients could participate in the study. The exclusion criteria included patients under the age of 18, unable to speak Dutch, and those who were illiterate.

Patients were asked to participate in the study and were provided with information about the study if willing to participate. After permission for participation was obtained from the patient, the researcher administered the questionnaires. As mentioned above, informed consent was not required as patients were anonymous and no medical data were analyzed.

The Two Questionnaires

The computer questionnaire is the standard method at the Anesthesiology Outpatient Department (Figure 1). We developed a chatbot questionnaire (Figure 2) with identical questions to the computer version. This ensured that the questionnaires were of the same length, avoiding bias due to increased or decreased appreciation per question. The patients completed both the standard and chatbot questionnaires, as the standard computer questionnaire was required as part of the preoperative system in the hospital. Patients started alternately with either the chatbot or the computer questionnaire, in order to prevent bias in length of time and user experience. During the completion of both questionnaires, time required to complete was documented.

Figure 1. Computer questionnaire.

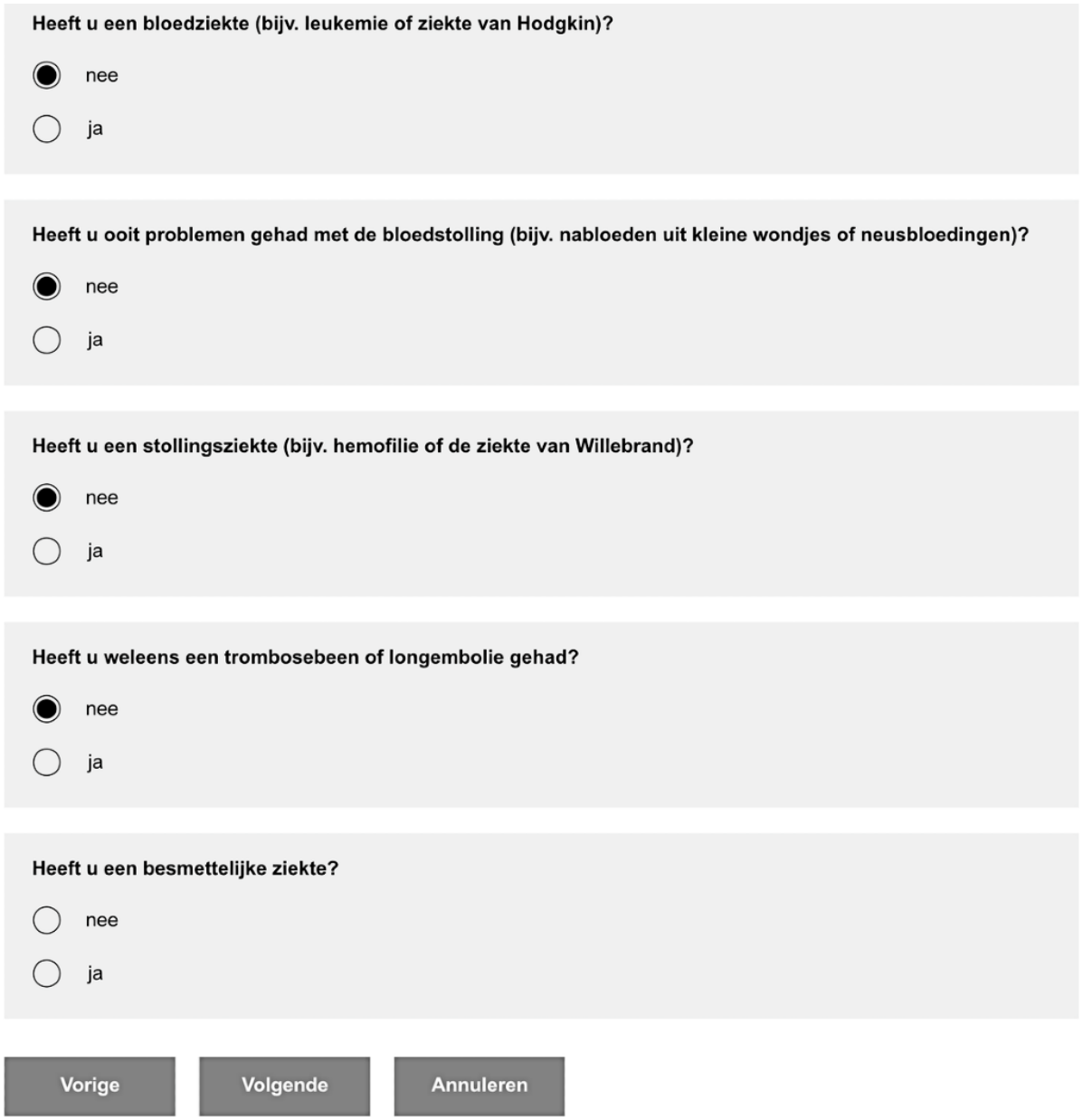
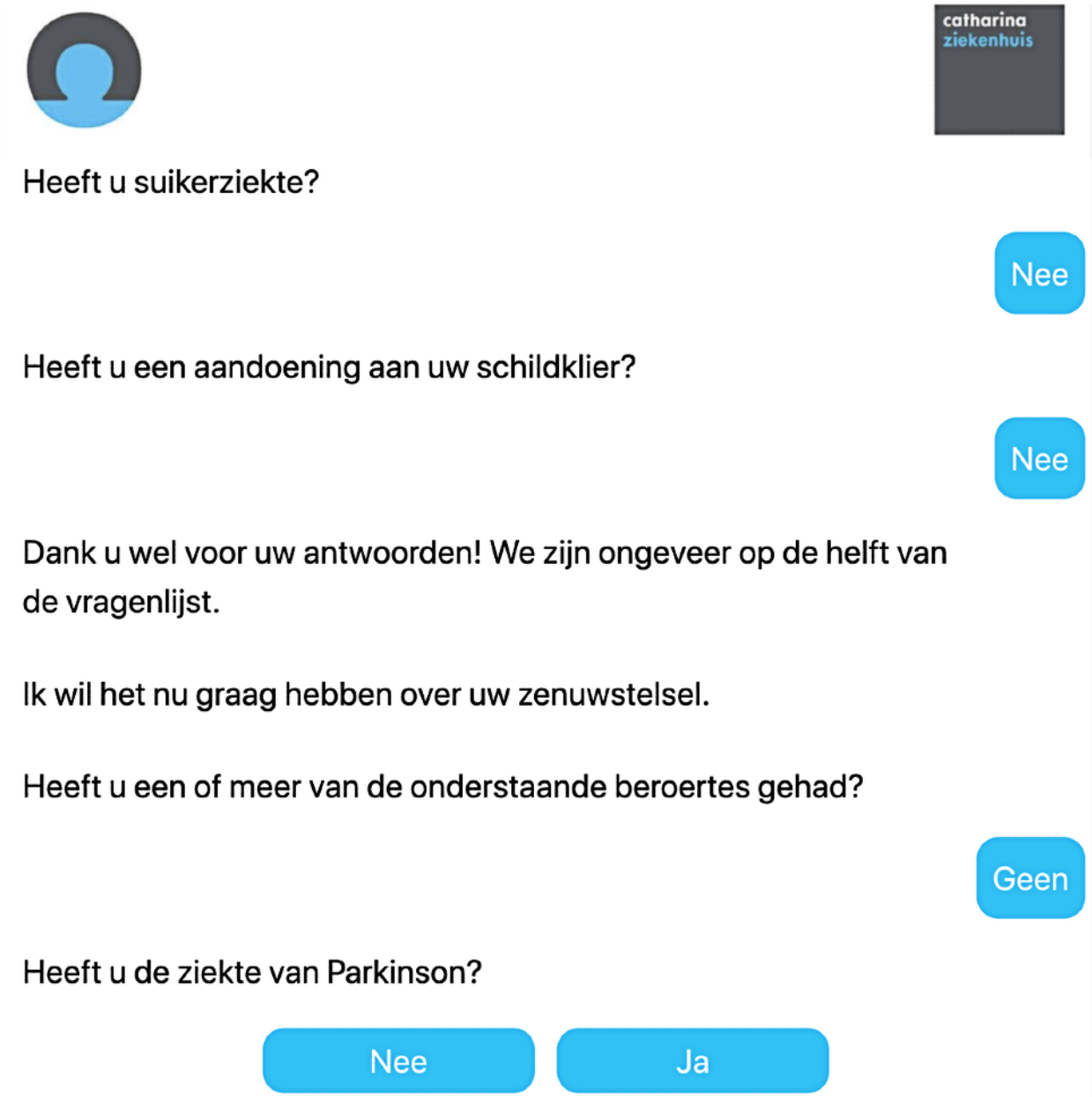


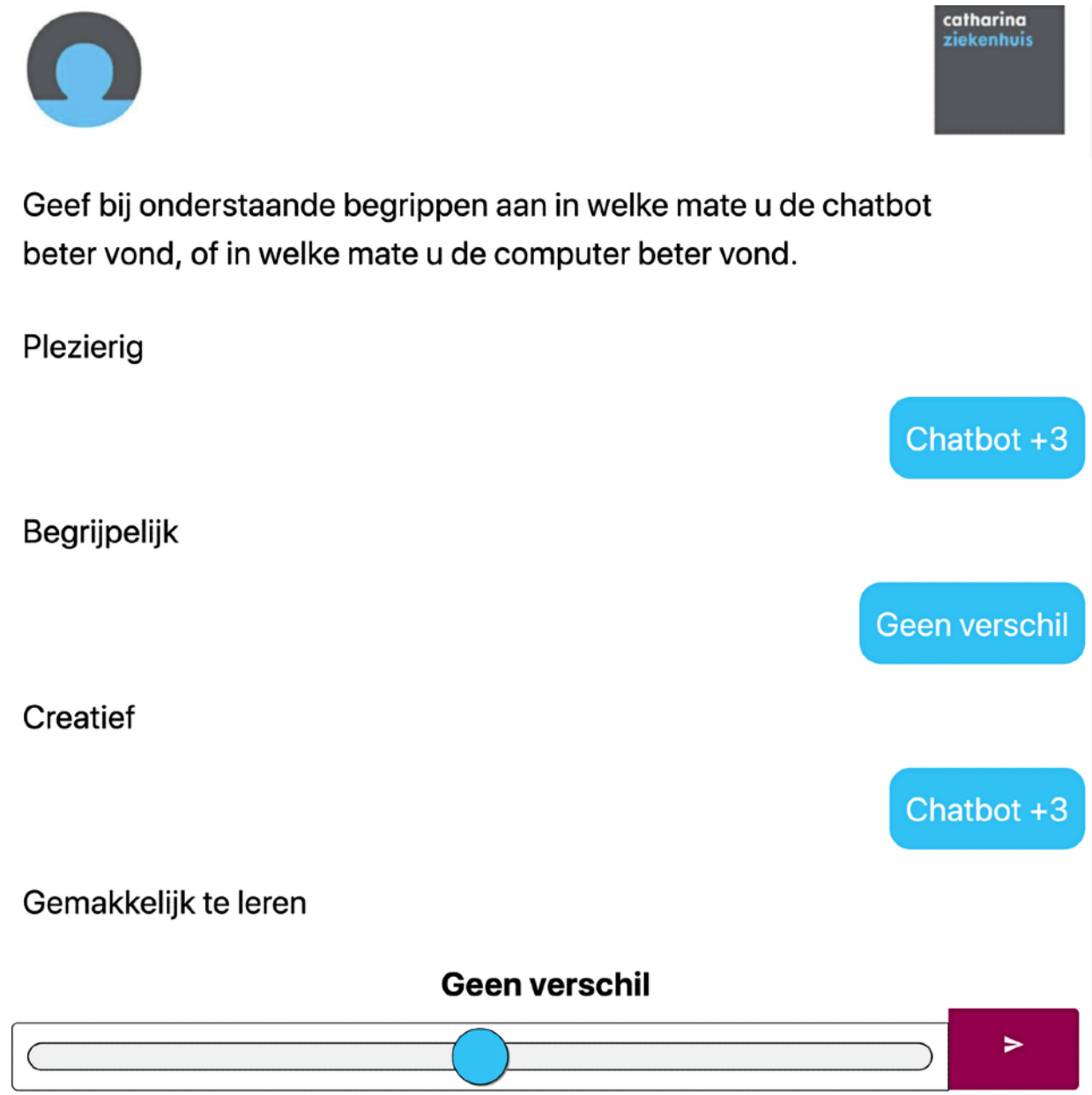
Figure 2. Chatbot questionnaire.



The User Experience Questionnaire

After completion of both questionnaires, patients provided feedback about the user experience. Patients were asked to rate their experience by providing scores for both questionnaires with the User Experience Questionnaire (UEQ; Figure 3). The reliability and validity of the UEQ scales were investigated in 11 usability tests which showed a sufficiently high reliability of the scales measured by Cronbach α [17-19]. Twenty-six terms were shown on a tablet and for each term patients gave their opinion by dragging the button to the “chatbot side” or to the “computer side.” They could choose to give 1, 2, 3, or 4 points to either the computer or the chatbot in relation to a specific term. If, according to the patient, there was no difference between the computer and the chatbot, he or she let the button in the middle of the bar.

Figure 3. User Experience Questionnaire.



The UEQ tested the following terms: pleasant, understandable, creative, easy to learn, valuable, annoying, interesting, predictable, rapid, original, obstructing, good, complex, repellent, new, unpleasant, familiar, motivating, as expected, efficient, clear, practical, messy, attractive, kind, and innovative.

As much as 20 of the 26 items were positive terms, such as "pleasant." The other 6 are negative terms, such as "annoying."

Outcome Measures

The primary outcome measure of this research is the user experience score and the difference in score between the standard computer questionnaire and the chatbot questionnaire. Secondary outcome was duration to complete a questionnaire.

Statistical Analysis

Data analysis primarily consisted of descriptive statistics and outcomes were mainly described in percentages or proportions. The unpaired *t* test was used to quantify significant differences between men and women and for time differences, because the data were normally distributed. A *P* value of .05 or less was chosen for statistical significance. Data were analyzed with SPSS statistics version 25 (IBM). Microsoft Excel version 16.1 was used for graphics.

This manuscript adheres to the applicable TREND guidelines [20].

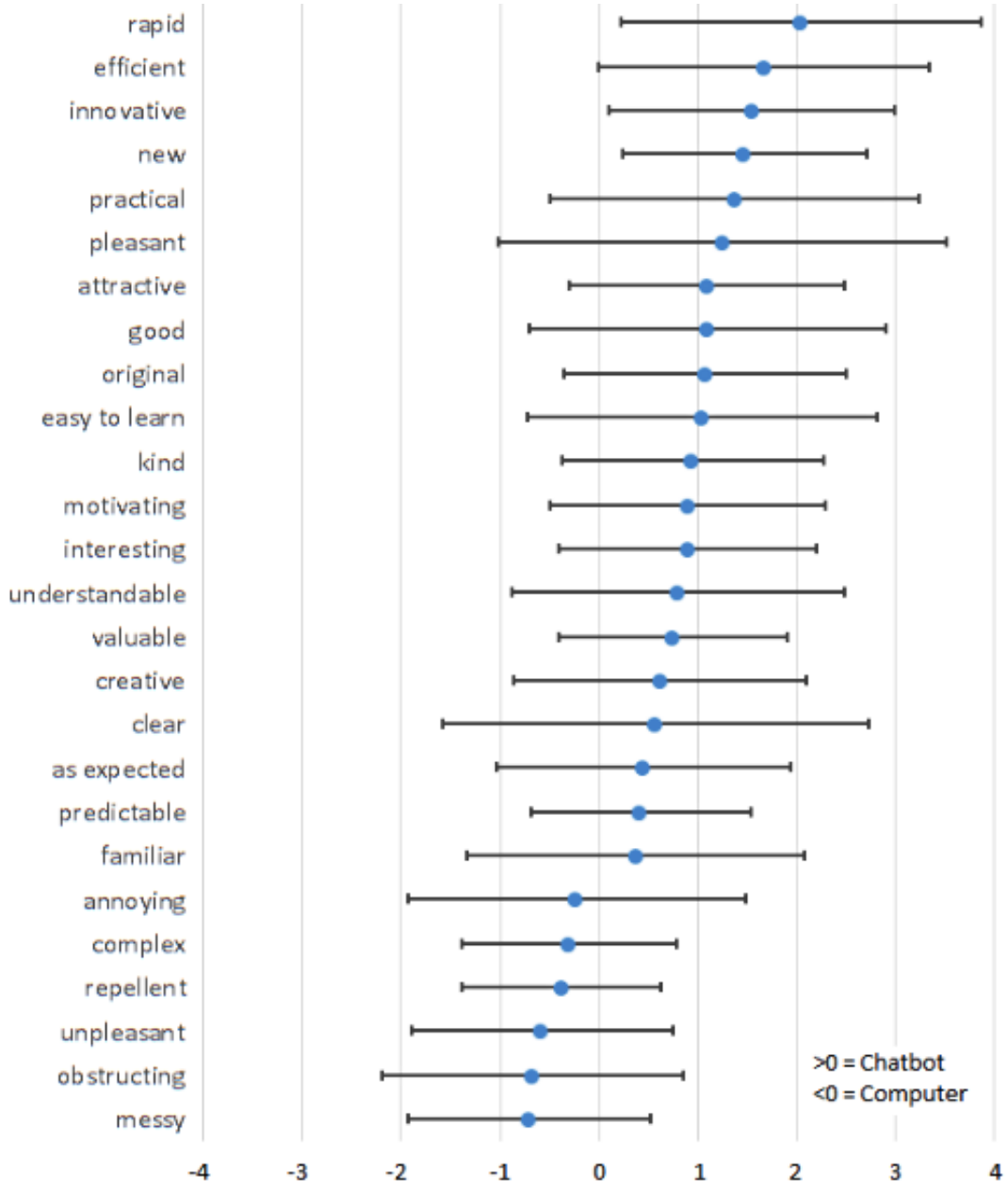
Results

The mean age of the 40 patients included, of whom 25 (63%) were women, was 49 (SD 18-79) years.

The average score per term was calculated and shown in Figure 4. The UEQ scores showed that patients favored the chatbot over the standard questionnaire. According to the graph, the patients prefer the chatbot for 20 of the 26 terms (77%), all of which are positive terms. The average values for the other 6 terms, which are the negative terms (23%), are shown to have

a negative value. This indicates that on average the patients associated the standard questionnaire with negative terms.

Figure 4. Average User Experience Questionnaire (UEQ) scores per term and standard deviation. A score above 0 illustrates that the term fits best with the chatbot. A score below 0 illustrates that the term fits best with the computer.



In total, 1040 terms were scored. As much as 46.73% (n=486) of the user experience terms were scored positive for the chatbot, 47.88% (n=498) of the terms had preference neither for chatbot nor computer, and for 7.98% (n=83) of the terms patients preferred the computer.

Average time to completion of the computer questionnaire was 8.20 (SD 2.69) minutes; for the chatbot questionnaire this was 7.72 (SD 2.76) minutes. The questionnaire completed initially took on average more time to complete, as the data in Table 1 indicate.

Time to completion differed between men and women, but did not reach statistical significance. Every patient completed the second questionnaire statistically significantly faster than the initial one (chatbot P=.044, computer P=.012), irrespective of which questionnaire was completed initially (Table 1).

Table 1. Time to completion (minutes).

Criteria	Computer questionnaire completion time (minutes), mean (SD)	Chatbot questionnaire completion time (minutes), mean (SD)
Average time to completion of computer- and chatbot-based questionnaire (n=40)		
All patients	8.20 (2.6)	7.72 (2.7)
Average time to completion for men (n=15) versus women (n=25)		
Men	9.00 (2.7)	8.33 (2.9)
Women	7.72 (2.6)	7.36 (2.6)
P value	.148	.287
Average time to completion depending on computer first (n=20) or chatbot first (n=20)		
Computer first	9.25 (2.4)	6.85 (2.1)
Chatbot first	7.15 (2.6)	8.60 (3.0)
P value	.012	.044

Discussion

Principal Findings

In this prospective observational study, we evaluated the user experience of a chatbot questionnaire and compared it to a standard computer questionnaire in an anesthesiology outpatient setting. Our results demonstrate that patients favored the chatbot questionnaire over the standard computer questionnaire according to the UEQ, which is in line with the previous research by Jain et al [21], who showed that users preferred chatbots as these provide a "human-like" natural language conversation.

Another intriguing result, as seen in Figure 4, is that the highest score to the chatbot was given for "rapid." However, the time to completion of the questionnaires did not differ between the computer questionnaire and the chatbot questionnaire. This indicates that a questionnaire answered on a tablet may give the perception of being faster than a standard model answered on a computer. In addition, by using more capabilities of a chatbot it is possible to shorten the questionnaire, possibly leading to higher response rates, as mentioned by Nakash et al [6].

The second questionnaire took significantly less time to complete than the initial one, as the contents are identical between the 2 questionnaires. This is not an unexpected observation. Although time to completion of the initial questionnaire was significantly different compared to that of the second questionnaire, bias in the results was minimized by alternating the order of questionnaires.

Comparison With Prior Work

Explanations for low response rates can be disinterest, lack of time, or inability to comprehend the questions. Furthermore, patient characteristics such as age, social economic status, relationship status, and those with preoperative comorbidities appear to have a negative influence on response rates, with the majority being nonmodifiable factors [22]. However, Ho et al [23] demonstrated that the method employed to invite and inform patients of the PROM collection, and the environment in which it is undertaken, significantly alters the response rate in the completion of PROMs. This means that, as expected in this study, there is a chance that response rates will rise by using a chatbot instead of a standard questionnaire.

Gamification

As described in the study by Edwards et al [7], response rates will rise when incentives are used. Currently, questionnaires are often lacking elements motivating the patient to complete them. The introduction of nudging techniques, such as gamification, can help. Nudging is the subtle stimulation of someone to do something in a way that is gentle rather than forceful or direct, based on insights from behavioral psychology [24,25]. In a recent study by Warnock et al [26], where the strong positive impact of gamification on survey completion was demonstrated, respondents spent 20% more time on gamified questions than on questions without a gamified aspect, suggesting they gave thoughtful responses [26]. Gamification has been proposed to make online surveys more pleasant to complete and, consequently, to improve the quality of survey results [27,28].

Limitations

There are some limitations to this research. First, as mentioned in the "Introduction" section, a chatbot can mimic intelligent conversation and is a form of gamification. In our study we had identical questionnaires and therefore did not explore how the chatbot could mimic intelligent conversation. However, this research demonstrates that only minor changes in the questionnaire's design lead to improved user experience. Second, because both the tablet and the chatbot were different from the standard computer questionnaire, it is possible that the user experience was influenced by the use of a tablet rather than by the characteristics of a chatbot solely. Third, although the UEQ shows us that the patients appreciated the chatbot more than the computer, we did not use qualitative methods to understand what factors drove users to identify the chatbot as a more positive experience. Fourth, although we recommend the use of a chatbot in the health care setting to improve

questionnaire response rate as seen in previous literature, we did not formally investigate this outcome.

Future Research

Because patients preferred the chatbot questionnaire over the computer questionnaire, we expect that a chatbot questionnaire can result in higher response rates. This research is performed as a first step in the development of a tool by which we can achieve adequate response rates in questionnaires such as the PROMs. Further research is needed, however, to investigate whether response rates of a questionnaire will rise due to alteration of the design. In future research it will be interesting to investigate which elements of gamification are needed to have beneficial effects such as higher response rates and higher quality of the answers as well.

Conclusions

Patients preferred the chatbot questionnaire over the conservative computer questionnaire. Time to completion of both questionnaires did not differ, though the chatbot questionnaire on a tablet felt more rapid compared to the computer questionnaire. Possibly, a gamified chatbot questionnaire could lead to higher response rates and to qualitatively better responses. The latter is important when outcomes are used for the evaluation of services, quality of care, and also outcome for value-based health care.

Authors' Contributions

All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by MP and WR. The first draft of the manuscript was written by MP and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Australian Commission on Safety and Quality in Health Care. URL: https://www.safetyandquality.gov.au/our-work/indicators-measurement-and-reporting/patient-reported-outcome-measures [accessed 2020-11-06]
2. Baumhauer JF, Bozic KJ. Value-based Healthcare: Patient-reported Outcomes in Clinical Decision Making. Clin Orthop Relat Res 2016 Jun;474(6):1375-1378. [doi: 10.1007/s11999-016-4813-4] [Medline: 27052020]
3. Gibbons E, Black N, Fallowfield L, Newhouse R, Fitzpatrick R. Essay 4: Patient-reported outcome measures and the evaluation of services. In: Raine R, Fitzpatrick R, Barratt H, Bevan G, Black N, Boaden R, et al, editors. Challenges, Solutions and Future Directions in the Evaluation of Service Innovations in Health Care and Public Health. Southampton, UK: NIHR Journals Library; May 2016.
4. Hazell ML, Morris JA, Linehan MF, Frank PI, Frank TL. Factors influencing the response to postal questionnaire surveys about respiratory symptoms. Prim Care Respir J 2009 Sep;18(3):165-170 [FREE Full text] [doi: 10.3132/pcrj.2009.00001] [Medline: 19104738]
5. Peters M, Crocker H, Jenkinson C, Doll H, Fitzpatrick R. The routine collection of patient-reported outcome measures (PROMs) for long-term conditions in primary care: a cohort survey. BMJ Open 2014 Feb 21;4(2):e003968 [FREE Full text] [doi: 10.1136/bmjopen-2013-003968] [Medline: 24561495]
6. Nakash RA, Hutton JL, Jørstad-Stein EC, Gates S, Lamb SE. Maximising response to postal questionnaires--a systematic review of randomised trials in health research. BMC Med Res Methodol 2006 Feb 23;6:5 [FREE Full text] [doi: 10.1186/1471-2288-6-5] [Medline: 16504090]
7. Edwards P, Roberts I, Clarke M, DiGuiseppi C, Pratap S, Wentz R, et al. Methods to increase response rates to postal questionnaires. Cochrane Database Syst Rev 2007 Apr 18(2):MR000008. [doi: 10.1002/14651858.MR000008.pub3] [Medline: 17443629]
8. Toepoel V, Lugtig P. Modularization in an Era of Mobile Web. Social Science Computer Review 2018 Jul:089443931878488. [doi: 10.1177/0894439318784882]
9. Sahlqvist S, Song Y, Bull F, Adams E, Preston J, Ogilvie D, iConnect Consortium. Effect of questionnaire length, personalisation and reminder type on response rate to a complex postal survey: randomised controlled trial. BMC Med Res Methodol 2011 May 06;11:62 [FREE Full text] [doi: 10.1186/1471-2288-11-62] [Medline: 21548947]
10. Robson K, Plangger K, Kietzmann JH, McCarthy I, Pitt L. Is it all a game? Understanding the principles of gamification. Business Horizons 2015 Jul;58(4):411-420. [doi: 10.1016/j.bushor.2015.03.006]
11. A. S, John D. Survey on Chatbot Design Techniques in Speech Conversation Systems. ijacsa 2015;6(7). [doi: 10.14569/ijacsa.2015.060712]
12. Kim S, Lee J, Gweon G. Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality. In: CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. New York, NY: ACM Press; Sep 04, 2019:1-12.

13. Palanica A, Flaschner P, Thommandram A, Li M, Fossat Y. Physicians' Perceptions of Chatbots in Health Care: Cross-Sectional Web-Based Survey. J Med Internet Res 2019 Apr 05;21(4):e12887. [doi: 10.2196/12887] [Medline: 30950796]
14. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. Digit Health 2019;5:2055207619871808 [FREE Full text] [doi: 10.1177/2055207619871808] [Medline: 31467682]
15. Greer S, Ramo D, Chang Y, Fu M, Moskowitz J, Haritatos J. Use of the Chatbot. JMIR Mhealth Uhealth 2019 Oct 31;7(10):e15018 [FREE Full text] [doi: 10.2196/15018] [Medline: 31674920]
16. Tudor Car L, Dhinagaran DA, Kyaw BM, Kowatsch T, Joty S, Theng Y, et al. Conversational Agents in Health Care: Scoping Review and Conceptual Analysis. J Med Internet Res 2020 Aug 07;22(8):e17158 [FREE Full text] [doi: 10.2196/17158] [Medline: 32763886]
17. Schrepp M, Hinderks A, Thomaschewski J. Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios. 2014 Jun Presented at: International Conference of Design, User Experience, and Usability; 2014; Heraklion, Crete, Greece p. 383-392. [doi: 10.1007/978-3-319-07668-3_37]
18. Laugwitz B, Held T, Schrepp M. Construction and Evaluation of a User Experience Questionnaire. In: Holzinger A, editor. USAB 2008: HCI and Usability for Education and Work. Berlin, Germany: Springer; 2008:63-76.
19. Baumhauer JF, Bozic KJ. Value-based Healthcare: Patient-reported Outcomes in Clinical Decision Making. Clin Orthop Relat Res 2016 Jun;474(6):1375-1378. [doi: 10.1007/s11999-016-4813-4] [Medline: 27052020]
20. Des Jarlais CC, Lyles C, Crepaz N, TREND Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. Am J Public Health 2004 Mar;94(3):361-366. [doi: 10.2105/ajph.94.3.361] [Medline: 14998794]
21. Jain M, Kumar P, Kota R, Patel SN. Evaluating and Informing the Design of Chatbots. In: DIS '18: Proceedings of the 2018 Designing Interactive Systems Conference. New York, NY: ACM; 2018 Presented at: Designing Interactive Systems (DIS) Conference; June 11-13, 2018; Hong Kong p. 895-906. [doi: 10.1145/3196709.3196735]
22. Schamber EM, Takemoto SK, Chenok KE, Bozic KJ. Barriers to completion of Patient Reported Outcome Measures. J Arthroplasty 2013 Oct;28(9):1449-1453. [doi: 10.1016/j.arth.2013.06.025] [Medline: 23890831]
23. Ho A, Purdie C, Tirosh O, Tran P. Improving the response rate of patient-reported outcome measures in an Australian tertiary metropolitan hospital. Patient Relat Outcome Meas 2019;10:217-226 [FREE Full text] [doi: 10.2147/PROM.S162476] [Medline: 31372076]
24. Nagtegaal R. [A nudge in the right direction? Recognition and use of nudging in the medical profession]. Ned Tijdschr Geneeskd 2020 Aug 20;164. [Medline: 32940980]
25. Cambridge Dictionary. URL: https://dictionary.cambridge.org/dictionary/english/nudging [accessed 2020-06-30]
26. Warnock S, Gantz JS. Gaming for respondents: a test of the impact of gamification on completion rates. Int J Market Res 2017;59(1):117. [doi: 10.2501/ijmr-2017-005]
27. Harms J, Biegler S, Wimmer C, Kappel K, Grechenig T. Gamification of Online Surveys: Design Process, Case Study, and Evaluation. In: Human-Computer Interaction – INTERACT 2015. Lecture Notes in Computer Science. Cham, Switzerland: Springer; 2015:219-236.
28. Guin TD, Baker R, Mechling J, Ruyle E. Myths and realities of respondent engagement in online surveys. Int J Mark Res 2012 Sep;54(5):613-633. [doi: 10.2501/ijmr-54-5-613-633]

Abbreviations

PROM: patient-reported outcome measure
UEQ: User Experience Questionnaire

Edited by C Lovis; submitted 30.06.20; peer-reviewed by R Watson, A Mahnke, J Shenson, T Freeman; comments to author 06.09.20; revised version received 12.10.20; accepted 03.11.20; published 07.12.20.

Please cite as:
te Pas ME, Rutten WGMM, Bouwman RA, Buise MP
User Experience of a Chatbot Questionnaire Versus a Regular Computer Questionnaire: Prospective Comparative Study
JMIR Med Inform 2020;8(12):e21982
URL: http://medinform.jmir.org/2020/12/e21982/
doi:10.2196/21982
PMID:33284125

Original Paper

Effects of Erythropoietin Payment Policy on Cardiovascular Outcomes of Peritoneal Dialysis Patients: Observational Study

Ying-Hui Hou[1*], PhD; Feng-Jung Yang[2,3*], MD, PhD; I-Chun Lai[4*], MD; Shih-Pi Lin[5], PhD; Thomas TH Wan[6], PhD; Ray-E Chang[5], PhD

[1]Department of Health Industry Management, School of Healthcare Management, Kainan University, Taoyuan, Taiwan
[2]Renal Division, Department of Internal Medicine, National Taiwan University Hospital Yun Lin Branch, Douliu, Taiwan
[3]School of Medicine, College of Medicine, National Taiwan University, Taipei, Taiwan
[4]Center for Drug Evaluation, Taipei, Taiwan
[5]Institute of Health Policy and Management, College of Public Health, National Taiwan University, Taipei, Taiwan
[6]Public Affairs PhD Program, College of Health and Public Affairs, University of Central Florida, Orlando, FL, United States
[*]these authors contributed equally

Corresponding Author:
Ray-E Chang, PhD
Institute of Health Policy and Management
College of Public Health
National Taiwan University
17 Xu-Zhou Road, Room 639
Taipei, 100
Taiwan
Phone: 886 2 3366 8069
Email: rchang@ntu.edu.tw

Abstract

Background: The change in the reimbursement policy of erythropoietin administration to patients receiving peritoneal dialysis by the Taiwan National Health Insurance (NHI) system provided a natural experimental venue to examine whether cardiovascular risk differs when maintaining the hematocrit (Hct) level below or above 30%.

Objective: The aim of this study was to analyze the impact of loosening the erythropoietin payment criteria for peritoneal dialysis patients on their cardiovascular outcomes.

Methods: Two cohorts of incident peritoneal dialysis patients were identified according to the time before and after relaxation of the NHI's erythropoietin payment criteria, designated cohort 1 (n=1759) and cohort 2 (n=2981), respectively. The cohorts were matched according to propensity scores (1754 patients in each cohort) and then followed up for cardiovascular events, which were analyzed with Cox regressions.

Results: For the composite cardiovascular endpoint, patients in cohort 2 had a significantly lower risk than those in cohort 1. However, subgroup analysis showed that this risk reduction was observed only in patients with diabetes.

Conclusions: After loosening erythropoietin payment criteria, reduced cardiovascular risks were observed, particularly for patients with diabetes. These results indicate that it is crucial to maintain an Hct level above 30% to reduce the cardiovascular risk in patients with diabetes undergoing peritoneal dialysis.

(JMIR Med Inform 2020;8(12):e18716) doi:10.2196/18716

KEYWORDS

erythropoietin; cardiovascular disease; peritoneal dialysis; diabetes mellitus

Introduction

Erythropoietin is a major regulatory hormone of erythrocyte production that is produced from the kidney, and its levels are decreased in patients with chronic kidney disease (CKD). A reduction in erythropoietin further decreases erythrocyte survival and leads to a chronic inflammatory status that contribute to anemia. Administration of exogenous erythropoietin for CKD

patients, especially those receiving dialysis, is the standard treatment for anemia.

Early studies showed that the use of erythropoietin tended to increase the hematocrit (Hct) target to the normal level (ie, 40.5% for men and 36% for women). However, more recent large, randomized outcome trials [1-3] showed that elevating the Hct level above 36% compared to maintaining Hct in the range of 30%-36% was associated with a higher risk of cardiovascular events for patients with CKD. These findings led to establishing the limitation of the Hct upper bound; however, the optimal Hct target remains debatable. The recommendations from the National Kidney Foundation-Kidney Disease Outcomes and Quality Initiative [4] and Taiwan's nephrology professionals [5] suggest maintaining the level of Hct between 33% and 36%.

The public statement of the European Medical Agency in 2007 concluded that the target Hct range should be 30%-36% [6]. The 2011 safety announcement of the US Food and Drug Administration recommended reducing or interrupting erythropoietin administration if the Hct level approaches or exceeds 33% for patients undergoing dialysis [7]. The recommendation from the Kidney Disease Improving Global Outcome in 2012 Clinical Practice Guideline was to maintain Hct below 34.5% [8]. Accordingly, an Hct range of 30%-36% might be considered the minimal bandwidth to accommodate all of these recommendations.

To reduce the cost of providing end-stage renal disease (ESRD) treatments while maintaining, or preferably improving, patient care, the US Center for Medicare and Medicaid (CMS) implemented the ESRD Prospective Payment System, known as the "expanded ESRD bundle," on January 1, 2011 [9]. Moreover, in response to a quality incentive program (QIP) required by US congress, two quality measures of anemia management were established to identify poor performance: patients with a hemoglobin (Hb) level less than 10 g/dL and those with an Hb level greater than 12 g/dL [9]. These Hb levels are equivalent to an Hct level less than 30% and above 36%, respectively, since 1 g/dL of Hb is equal to 3% Hct. However, the CMS retired the measure of an Hb level less than 10 g/dL in its later QIP requirements [10,11]; that is, dialysis facilities would receive no penalties for patients with Hb levels lower than 10g/dL, who might be spotted more often in the future. The elimination of penalties for the lower bound of Hb levels has indeed removed the financial incentives to provide costly erythropoietin treatment, while raising some concerns about patient care [12]. Nevertheless, it remains unclear whether patients with an Hb level lower than 10 g/dL or an Hct level lower than 30% have a higher risk of adverse events, which is a logical inquiry that warrants further investigation.

Limited studies have reported cardiovascular events or mortality associated with Hct levels lower than 30%. Studies comparing dialysis patients with an Hct level maintained below 30% to those with Hct levels maintained in the range of 30%-36% showed no significant difference in adverse outcomes [13-15]. However, more recent studies [1-3] comparing the risk of pushing Hct levels above 36% with those maintained between 30%-36% included a larger sample size of more than 1200 patients with a follow-up period of more than 14 months, in contrast to the early studies with a relatively small sample size of 152 patients or less and a short follow-up period of 6-9 months. Moreover, the design of these studies was not specifically focused on assessing this question. Recently, the change in the reimbursement policy of erythropoietin administration to patients undergoing peritoneal dialysis by the Taiwan National Health Insurance (NHI) system provides a natural experimental venue for directly examining this clinical research issue.

The incidence and prevalence rates of ESRD in Taiwan have been ranked at the top internationally since 2001 [16], placing an immense burden of caring and funding for ESRD patients on the Taiwan NHI system. The low renal transplant rate, at less than 1% annually [17], results in nearly all of Taiwan's ESRD patients relying on dialysis treatments to prolong their lives, with more than 93.5% of ESRD patients receiving hemodialysis treatments in 2004 [18]. To increase peritoneal dialysis utilization, Taiwan's NHI has introduced a series of encouragement policies since 2005, including loosening the reimbursement criteria. Before November 1, 2006, the treatment of erythropoietin to a patient undergoing peritoneal dialysis could only be reimbursed by the NHI if the patient's Hct level was ≤30% and they were receiving a maximal monthly erythropoietin dosage of 20,000 U epoetin alfa/beta or 100 μg darbepoetin alfa. After November 1, 2006, the Hct level at which erythropoietin administration could be reimbursed was relaxed to ≤36% with the same maximal monthly erythropoietin dosage requirements. Subsequent to this relaxation of erythropoietin administration criteria, the Hct levels for both prevalent and incident peritoneal dialysis patients increased from 28%-29% to 30%-31% [19-21].

The main purpose of this study was to analyze the impact of loosening the erythropoietin administration criteria for patients undergoing peritoneal dialysis in Taiwan with a focus on exploring the risk of cardiovascular events when maintaining Hct at 30%-31% as compared to 28%-29%.

Methods

Ethics Statement

Data were obtained from the National Health Insurance Research Database [22], which are accessible to researchers after ethical and scientific review processes. Prior to applying for this access, this study was approved by the ethical review board of National Taiwan University Hospital (NTUH-REC No. 201406018W). There are 27 institutional review boards capable of issuing approvals, and all are supervised and regulated by the Taiwan Ministry of Health and Welfare. To protect individuals' confidentiality, all datasets in the Data Science Centre are pseudonymized. Personal ID, birth date, and names are encrypted, and this deidentification process was approved by an independent third party. We performed data analysis in the branches of the Data Science Centre. The analyzed results were also examined by the Data Science Centre before exporting. The Institutional Review Board verified the anonymity of data analysis performed in this study. All research procedures followed the directives of the Declaration of Helsinki.

Study Design

This was an observational study designed to compare the cardiovascular events of two cohorts of newly treated (incident) patients undergoing peritoneal dialysis before and after relaxation of the NHI's erythropoietin payment criteria. Cohort 1 included dialysis patients who started to receive maintenance peritoneal dialysis treatments during a specified period of 28 months before relaxation of the NHI's erythropoietin payment criteria. To ensure an adequate observation period, this cohort was followed up for an additional 14 months after the month in which the last patient was enrolled in the study. Cohort 2 included incident dialysis patients who started to receive maintenance peritoneal dialysis treatments within a 28-month time interval after relaxation of the NHI's erythropoietin payment criteria. Additional 14-month follow-up observations were also made after the month in which the last patient of this cohort was enrolled in the study. We set a 6-month time lag between the initiation of relaxing the erythropoietin payment criteria and the time that the first patient was enrolled in cohort 2 to accommodate possible adaptations of the physician prescribing practices to the new policy.

Because of potential imbalances in the distributions of many measured and unmeasured baseline covariates between the two cohorts, propensity score (PS) analysis, which was developed by Rosenbaum et al [23], was used in this study. Thus, the influence of any potential enrollment biases between these two cohorts was attenuated through a PS-matching approach and identification of patients with comparable characteristics in the two cohorts. This study defined PS as the probability of a patient having experienced a cardiovascular event. Patients in cohorts 1 and 2 were matched with PS scores estimated by age, sex, and the comorbidity index with the Greedy nearest neighbor algorithm [24]. The comorbidity index was developed by Liu et al [25] specifically for the US Medicare dialysis population and was subsequently validated for Taiwanese dialysis patients [26].

After matching with the PS, patients were followed up until experiencing either one of the following three events: (1) the occurrence of cardiovascular endpoints, (2) change to hemodialysis, or (3) the data cutoff point (October 31, 2006 for cohort 1 and October 31, 2010 for cohort 2), whichever occurred earlier. Survival analysis models were then employed to investigate the differences in the risk of cardiovascular events between the two cohorts of incident peritoneal dialysis patients. Baseline demographics and comorbid conditions were used as covariates in the statistical analyses. Monthly erythropoietin doses administered to patients of cohort 1 and cohort 2 during the follow-up period were compared to examine a difference between the two cohorts of incident peritoneal dialysis patients. In calculation of erythropoietin dosage, epoetin alfa and epoetin beta were considered to be equivalent, whereas darbepoetin alfa was converted to epoetin alfa based on the equivalence of 1 μg of darbepoetin alfa to 200 U of epoetin alfa [27].

Cardiovascular risk could be affected by treatments with concomitant medications related to cardiovascular comorbidities. Therefore, patients taking medications related to cardiovascular comorbidities during the follow-up period in the two cohorts were also examined. The concomitant medications related to cardiovascular comorbidities were identified by corresponding Anatomical Therapeutic Chemical classification codes, including acetylsalicylic acid (B01AC06) or clopidogrel (B01AC04), angiotensin-converting enzyme inhibitors (C09A) or angiotensin receptor blockers (C09C), beta blockers (C07), calcium channel blockers (C08), and statins (C10AA). A patient who received such medication for any of the 3 months during the follow-up period would be considered to be under treatment of concomitant medications related to cardiovascular comorbidities.

Finally, in addition to administering erythropoietin, because the patient's Hct level could also be affected by the use of iron and red cell transfusion, the differences in iron and red cell transfusion were compared between patients in the two cohorts.

Patient Selection

Incident peritoneal dialysis patients were identified from the claim data of entire beneficiaries covered by the NHI system from 2003 to 2010. Collection and analysis of the NHI claimed data were approved by the National Taiwan University Hospital Human Research Ethics Committee. The analyses were performed on deidentified data extracted from the NHI research database compiled by Taiwan National Health Research Institutes. A patient receiving over 90-day consecutive dialysis treatments and with peritoneal dialysis performed on day 90 and thereafter was considered to be an incident peritoneal dialysis patient in this study. Cohort 1 included patients who received dialysis as of the 90th day between May 1, 2003 and August 31, 2005, and cohort 2 included patients who received dialysis as of the 90th day between May 1, 2007 and August 31, 2009. Young patients (under 20 years) were excluded because comorbidities differed between pediatric and adult patients. There were 1759 patients in cohort 1 and 2981 patients in cohort 2. After PS-based matching, each cohort contained 1754 patients.

Statistical Analyses

The primary outcome measure was a composite cardiovascular endpoint, defined as myocardial infarction, heart failure hospitalization, stroke, or death. Myocardial infarction was defined by International Classification of Diseases, Ninth Revision (ICD-9) codes 410 and 411 in the hospital discharge diagnosis. Heart failure hospitalization was defined by ICD-9 hospital discharge diagnosis codes 398.91, 422, 425, 428, 402.x1, 404.x1, 404.x3, and V42.1. Stroke was defined by ICD-9 hospital discharge diagnosis codes 433, 434, 436, 437.0, and 437.1. For the primary outcome measure, all patients in both cohorts were followed up until the occurrence of myocardial infarction, heart failure hospitalization, stroke, or death, whichever occurred earlier. Secondary outcomes were the individual components of the composite primary outcome: myocardial infarction, heart failure hospitalization, stroke, and death. Each patient was followed up until the occurrence of each cardiovascular event. Data on patients who did not have an event were censored at the data cut-off point or date of transition to hemodialysis, whichever occurred earlier.

The selection and analyses of primary and secondary endpoints of cardiovascular risk in this study were the same as those

adopted in previous large-scale studies [1-3]. In addition to cardiovascular events, death was also considered an important clinical endpoint in the evaluation of cardiovascular risk because reducing mortality is an ultimate goal of reducing cardiovascular risk. Using a composite primary endpoint with each component evaluated as the secondary endpoint analysis is commonly adopted by many clinicians [2,3], such as in pivotal studies of new drug applications. This allows for a thorough evaluation of the contribution of each component of the composite primary endpoint and avoids any biases introduced by a dominating component.

The Cox proportional hazards model was employed to estimate the cardiovascular risk between the two cohorts. Estimated hazard ratios (HRs) for cohort 2 relative to cohort 1 and 95% CIs were calculated. To obtain more insightful results, patients were further stratified by diabetes status; Cox regression analyses for patients with and without diabetes were performed separately. All analyses were performed using SAS software, version 9.1.

Results

Patient Selection

Table 1 shows the baseline demographics and comorbid conditions of the equal number (n=1754) of incident peritoneal dialysis patients in the two cohorts. No statistically significant differences were observed, suggesting that patients in the two cohorts appeared to be similar in terms of age, gender, and comorbid conditions at baseline. There were also no significant differences in the usage of any concomitant medication related to cardiovascular comorbidities between the two cohorts.

Table 1. Baseline demographics and concomitant medications during the follow-up period in cohort 1 and cohort 2 after matching with the propensity score.

Characteristic	Matched[a] cohort 1 (n=1754)	Matched cohort 2 (n=1754)	*P* value[b]
Female, n (%)	994 (56.67)	991 (56.50)	.84
Age (years), mean (SD)	52.96 (15.36)	52.87 (15.02)	.33
Age group (years), n (%)			
20-39	326 (18.59)	327 (18.64)	
40-49	390 (22.23)	384 (21.89)	
50-59	431 (24.57)	444 (25.31)	
60-69	320 (18.24)	324 (18.47)	
≥70	287 (16.36)	275 (15.68)	
Comorbidity index, mean (SD)	2.52 (1.72)	2.52 (1.79)	.80
Comorbidity index, n (%)			
0	401 (22.86)	401 (22.86)	
1	268 (15.28)	269 (15.34)	
2	324 (18.47)	323 (18.42)	
3	245 (13.97)	243 (13.85)	
4	180 (10.26)	182 (10.38)	
5	148 (8.44)	148 (8.44)	
6	94 (5.36)	94 (5.36)	
7	49 (2.79)	50 (2.85)	
8	24 (1.37)	23 (1.31)	
9	10 (0.57)	10 (0.57)	
≥10	11 (0.63)	11 (0.63)	
Baseline comorbidity, n (%)			
Atherosclerotic heart disease	327 (18.64)	320 (18.24)	.49
Congestive heart failure	192 (10.95)	192 (10.95)	>.99
Cerebrovascular accident/transient ischemic attack	273 (15.56)	268 (15.28)	.67
Peripheral vascular disease	250 (14.25)	253 (14.42)	.76
Other cardiac disease	220 (12.54)	223 (12.71)	.75
Chronic obstructive pulmonary disease	106 (6.04)	110 (6.27)	.59
Gastrointestinal bleeding	212 (12.09)	207 (11.80)	.65
Liver disease	200 (11.40)	204 (11.63)	.66
Dysthymia	60 (3.42)	56 (3.19)	.48
Cancer	149 (8.49)	151 (8.61)	.80
Diabetes	581 (33.12)	584 (33.30)	.82
Hypertension	1297 (73.95)	1305 (74.40)	.70
Atrial fibrillation	19 (1.08)	15 (0.86)	.33
Coronary artery bypass graft	134 (7.64)	128 (7.30)	.59
Myocardial infarction	22 (1.25)	21 (1.20)	.89
Concomitant medications, n (%)			
Acetylsalicylic acid or clopidogrel	1369 (78.05)	1355 (77.3)	.39
ACEIs[c] or ARBs[d]	637 (36.32)	631 (35.97)	.38

Characteristic	Matched[a] cohort 1 (n=1754)	Matched cohort 2 (n=1754)	*P* value[b]
Beta blockers	589 (33.58)	586 (33.41)	.48
CCB[e]	683 (38.94)	693 (39.51)	.37
Statins	509 (29.02)	504 (28.73)	.32
Oral iron usage, n (%)	72 (4.10)	69 (3.93)	.63
Intravenous iron usage, n (%)	794 (45.27)	772 (42.01)	.61
Red cell transfusions, n (%)	194 (11.06)	170 (9.69)	.09
Red cell transfusion units per patient per month, mean (SD)	0.059 (0.216)	0.044 (0.172)	.03
Oral iron dose per patient per month (mg), mean (SD)	25.06 (129.66)	23.39 (125.0)	.23
Intravenous iron dose per patient per month (mg), mean (SD)	106.54 (92.29)	98.91 (89.38)	.19
Erythropoietin[f] usage per patient per month (U), median (IQR)	10,588[c] (7750-13,280)[d]	12,379[c] (8580-14,570)	<.001

[a]Matching with propensity score was based on age, sex, and comorbidity index using the Greedy method.

[b]Means (SD) were compared with the *t* test, n (%) values were compared with the proportion *z* test, and medians (IQR) were compared with the Wilcoxon rank-sum test.

[c]ACEIs: angiotensin converting enzyme inhibitors.

[d]ARBs: angiotensin receptor blockers.

[e]CCB: calcium channel blocker.

[f]Including epoetin alfa, epoetin beta, and darbepoetin alfa; epoetin alfa and beta were considered equivalent, and 100 μg darbepoetin was considered equivalent to 20,000 U erythropoietin according to the reimbursement criteria of the Taiwan National Health Institute.

Erythropoietin Dosage

The median monthly erythropoietin dosage was significantly higher in cohort 2 than in cohort 1 (12,739 U vs 10,588 U, *P*<.001). The usage of iron supplements (both oral and intravenous) and red cell transfusions were comparable in the two cohorts (Table 1).

Endpoint Evaluation

For the composite cardiovascular endpoint, the risk in cohort 2 was significantly lower after adjusting for age, sex, comorbidity index, diabetes mellitus, hypertension, history of coronary artery bypass graft, and congestive heart failure (Table 2). For each cardiovascular endpoint, the risk reduction in cohort 2 did not reach statistical significance.

Table 2. Comparison of primary and secondary endpoints between the cohorts.

Endpoint	Matched cohort 1 (n=1754), n (%)	Matched cohort 2 (n=1754), n (%)	Hazard ratio[a] (95% CI)	*P* value
Primary endpoint: cardiovascular composite events	299 (17.05)	261 (14.88)	0.82 (0.69-0.98)	.04
Secondary endpoints				
Myocardial infarction	40 (2.28)	36 (2.05)	0.81 (0.48-1.19)	.20
Stroke	58 (3.31)	45 (2.57)	0.72 (0.50-1.12)	.15
Heart failure hospitalization	173 (9.86)	162 (9.24)	0.76 (0.65-1.09)	.17
Death	91 (5.19)	89 (5.07)	0.92 (0.68-1.24)	.59

[a]Adjusted for age, sex, comorbidity index, diabetes, hypertension, history of coronary artery bypass graft, and congestive heart failure.

In the subgroup analysis (Table 3), for patients that did not have diabetes, no significant difference in either the composite cardiovascular endpoint or any individual cardiovascular endpoint was observed between the two cohorts. However, for patients with diabetes, the risk of the composite cardiovascular endpoint was significantly lower in cohort 2. In addition, the risks of stroke and heart failure hospitalization were significantly lower in cohort 2 than those of cohort 1.

Table 3. Subgroup analysis according to diabetes status in comparing the endpoints between matched cohort 1 and cohort 2.[a]

Endpoint	Patients with diabetes[b]		Patients without diabetes[c]	
	Hazard ratio[d] (95% CI)	*P* value	Hazard ratio[d] (95% CI)	*P* value
Primary endpoint: Cardiovascular composite	0.74 (0.60-0.93)	.006	0.97 (0.74-1.27)	.82
Secondary endpoints				
Myocardial infarction	0.67 (0.36-1.15)	.19	0.86 (0.33-2.25)	.76
Stroke	0.61 (0.39-0.98)	.04	1.02 (0.51-2.04)	.93
Heart failure hospitalization	0.72 (0.54-0.99)	.04	1.06 (0.74-1.51)	.76
Death	1.07 (0.73-1.58)	.73	0.79 (0.49-1.26)	.27

[a]Patients in cohorts 1 and 2 were matched with the propensity score by age, sex, and comorbidity index using the Greedy method.

[b]Cohort 1, n=581; cohort 2, n=584.

[c]Cohort 1, n=1173; cohort 2, n=1170.

[d]Adjusted by age, sex, comorbidity index, hypertension, history of coronary artery bypass graft, and congestive heart failure.

Discussion

Summary

No statistically significant difference was observed for baseline comorbidities and concomitant medications in the follow-up period between the matched cohort 1 and cohort 2 (Table 1). This suggests that both cohorts had similar cardiovascular risk factors. After loosening erythropoietin payment criteria, the erythropoietin dosage increased and the cardiovascular risk decreased; however, the reduction in cardiovascular risk was observed only in patients with diabetes. In addition, among patients with diabetes, significant risk reduction was found not only for the composite cardiovascular endpoint but also for the individual secondary endpoints, including stroke and heart failure hospitalization. Since similar percentages of patients in matched cohort 1 and cohort 2 received oral and intravenous iron, and the oral and intravenous iron dosage was comparable between these two cohorts, it is reasonable to assume that the higher Hct level in matched cohort 2 might have resulted from the higher erythropoietin dosage. Similarly, the reduction in cardiovascular risk in matched cohort 2 may be related to the higher erythropoietin dosage and maintenance of an adequate Hct range.

Comparison With Prior Work

Although previous findings that pushing Hct to more than 36% compared to 30%-36% tends to increase cardiovascular risk [1-3,7] have been widely accepted and recommended, there is a lack of sufficient evidence to demonstrate a difference in cardiovascular risk by maintaining Hct levels below 30% relative to 30%-36%. A few studies with small sample sizes and short follow-up periods showed no significant difference in cardiovascular risk or mortality for patients maintaining Hct below 30% compared to those maintaining Hct at 30%-36% [13-15]. Thus, these limitations have prevented investigators from detecting the potential difference in cardiovascular risk. By contrast, our national study showed that a lower cardiovascular risk is associated with increasing Hct from 28%-29% to 30%-31% for incident peritoneal dialysis patients in Taiwan. The number of subjects in our study was 3508 and the median follow-up duration was 23 months, which are comparable to those of more recent large-scale studies [1-3] with a sample size between 1265 and 4038 and median follow-up duration between 14 and 29 months.

Principal Findings

Although the Hct data reported in the NHI beneficiaries claim database did not directly link to observations of patients' Hct levels of this study, we used the data from the whole NHI population (census) and government documents publishing Hct statistics for dialysis patients supported by the NHI [19-21]. Moreover, from the governmental published data, the Hct levels of both prevalent and incident peritoneal dialysis patients were very similar (28.9% to 30.4% vs 29.1% to 30.4% from 2005 to 2008) and the Hct of both peritoneal dialysis patients with and without diabetes mellitus were also very similar (28.5% to 30.6% vs 28.3% to 30.3% from 2003 to 2008). Therefore, we assumed that the Hct levels of incident peritoneal dialysis patients in our study were similar to those reported in the government documents. After loosening the erythropoietin payment criteria, the Hct level of both prevalent and incident peritoneal dialysis patients increased from 28%-29% to 30%-31% [19-21].

In this study, the median erythropoietin dosage in cohort 2 (12,739 U) was significantly higher than that in cohort 1 (10,588 U); that is, there was a more than 20% increase in the dosage after loosening the erythropoietin reimbursement criteria. Given that the usage rates of iron supplements (both oral and intravenous) and red cell transfusions were comparable in the two cohorts, increased erythropoietin usage supports the assumption that the Hct of incident peritoneal dialysis patients also increased after loosening the erythropoietin payment criteria.

Because the reduction in cardiovascular risk was observed only in patients with diabetes, the difference in cardiovascular event risk reduction between patients with and without diabetes might not be the result of the Hct difference; indeed, the Hct was similar between peritoneal dialysis patients with (28.5%-30.6%) and without (28.3%-30.3%) diabetes from 2003 to 2008 [21].

Therefore, rather than analyzing the two subgroups (with and without diabetes) separately through a Cox proportional hazards model, we reanalyzed the nonstratified data through a Cox proportional hazards model with the addition of two more variables: one dichotomous variable for differentiating patients according to diabetes status and another interaction term between diabetes status and cohort. The estimate of diabetes status represented the cardiovascular risk of patients with diabetes relative to that of patients without diabetes in the time period of cohort 1, and the estimate of the interaction term measured the change in cardiovascular risk of patients with diabetes relative to that of patients without diabetes in the time period of cohort 2 compared to the time period of cohort 1. These results showed that the incident peritoneal dialysis patients with diabetes had a significant 78% higher cardiovascular risk than those of patients without diabetes. Although there was no significant difference in cardiovascular risk observed for our peritoneal dialysis patients without diabetes in cohort 2 (HR 0.974, 95% CI 0.84-1.05), the cardiovascular risk of the patients with diabetes in cohort 2 was significantly reduced by 22% (HR 0.78, 95% CI 0.61-0.94). This means that the cardiovascular risk of incident peritoneal dialysis patients with diabetes mellitus was 39% (1.78×0.78=1.39) higher than that of patients without diabetes in the time period of cohort 2, and was reduced by 78% in the time period of cohort 1. There was no significant difference in the erythropoietin dosages used for patients in the two cohorts according to diabetes status in either cohort (diabetes vs no diabetes median 10,726 U vs 10,525 U, *P*=.09 in cohort 1; 12,254 U vs 12,310 U, *P*=.17 in cohort 2). Given these findings and the similar Hct levels between the patients with and without diabetes, the observed increases in erythropoietin dosage and the Hct levels from below 30% to above 30% might benefit peritoneal dialysis patients with diabetes in terms of reducing the cardiovascular risk but would have no impact on the cardiovascular risk of patients without diabetes.

This finding has an important implication for policymakers for making decisions as to how to allocate health care resources and improve patient care in a cost-efficient manner, which is a major challenge for policymakers worldwide, including Taiwan and the United States. Based on these findings, Taiwan's NHI policymakers should reconsider the relaxation of NHI's reimbursement criteria to target only peritoneal dialysis patients with diabetes rather than applying these criteria universally. In this way, the NHI could spend less while improving diabetic peritoneal dialysis patient care by reducing the cardiovascular risk. With respect to policy decisions in the United States, it is possible that more patients would have an Hb level below 10 g/dL (ie, Hct 30%) and thus a higher cardiovascular risk might be incurred for ESRD patients with diabetes after eliminating the QIP requirement of an Hb level <10 g/dL. Thus, determining whether a lower bound of the Hct/Hg level should be restored for ESRD patients with diabetes mellitus to reach a balance between cost reduction and improvement of patient care is a critical issue to be examined by US policymakers.

Limitations

A more clinically oriented inquiry may explain why the peritoneal dialysis patients with diabetes showed a stronger response to the increase in erythropoietin dosage and Hct levels in terms of reducing cardiovascular risk. Our data do not enable directly testing this clinical issue and thus more research to this end is warranted. There are also limitations of this study. No blood pressure or laboratory data, including serum albumin and lipid profile, were available from the NHI claim database, which prevented performing a comprehensive comparison of baseline characteristics between the two cohorts. Although this might have constrained detailed matching of patients in the two cohorts, the patients matched in the two cohorts were considerably comparable with respect to comorbid conditions and concomitant medication related to cardiovascular risk.

Conclusions

After loosening the erythropoietin payment criteria, a significantly lower risk of cardiovascular events, stroke, and heart failure hospitalization was observed in matched cohort 2, in particular for those with diabetes mellitus. This risk reduction may be related to the higher erythropoietin dosage and maintenance of an adequate Hct range. Further research is needed to investigate why peritoneal dialysis patients with diabetes mellitus are more sensitive to the increase in erythropoietin dosage and Hct levels. Our findings support that for these patients, maintaining an Hct level above 30% is crucial for reducing the cardiovascular risk. This finding has implications for policymakers to determine the allocation of health care resources in a cost-effective manner while reducing the potential cardiovascular risk for patients receiving peritoneal dialysis.

Acknowledgments

This work was partly supported by grants from the Ministry of Science and Technology (MOST 103-2410-H-002 -205-). This work is based in part on data obtained from the National Health Insurance Research Database provided by the National Health Insurance Administration, Ministry of Health and Welfare, and managed by National Health Research Institutes. The views are solely those of the authors and do not represent those of the National Health Insurance Administration, Ministry of Health and Welfare, or National Health Research Institutes. The authors would like to thank Mr. Shin-hung Meng for data management and Dr. Ya-Chi Wu for excellent statistical support.

Authors' Contributions

IL contributed to the conception and design of the study, data interpretation, drafting the article, and final approval of the version to be published. RC contributed to the conception and design of the study, acquisition and interpretation of data, article revision,

and final approval of the version to be published. SL contributed to analysis and interpretation of the data, drafting the article, and final approval of the version to be published. YH, FY, and TW contributed to analysis and interpretation of the data, article revision, and final approval of the version to be published.

Conflicts of Interest

None declared.

References

1. Besarab A, Bolton WK, Browne JK, Egrie JC, Nissenson AR, Okamoto DM, et al. The effects of normal as compared with low hematocrit values in patients with cardiac disease who are receiving hemodialysis and epoetin. N Engl J Med 1998 Aug 27;339(9):584-590. [doi: 10.1056/NEJM199808273390903] [Medline: 9718377]
2. Pfeffer MA, Burdmann EA, Chen CY, Cooper ME, de Zeeuw D, Eckardt KU, TREAT Investigators. A trial of darbepoetin alfa in type 2 diabetes and chronic kidney disease. N Engl J Med 2009 Nov 19;361(21):2019-2032. [doi: 10.1056/NEJMoa0907845] [Medline: 19880844]
3. Singh AK, Szczech L, Tang KL, Barnhart H, Sapp S, Wolfson M, CHOIR Investigators. Correction of anemia with epoetin alfa in chronic kidney disease. N Engl J Med 2006 Nov 16;355(20):2085-2098. [doi: 10.1056/NEJMoa065485] [Medline: 17108343]
4. KDOQI. KDOQI Clinical Practice Guideline and Clinical Practice Recommendations for anemia in chronic kidney disease: 2007 update of hemoglobin target. Am J Kidney Dis 2007 Sep;50(3):471-530. [doi: 10.1053/j.ajkd.2007.06.008] [Medline: 17720528]
5. Renal anemia. In: Taiwan Chronic Kidney Disease Clinical Guidelines. Miaoli County, Taiwan: National Health Research Institutes; 2015:393-397.
6. Tsintis P. Public statement: Epoetins and the risk of tumour growth progression and thromboembolic events in cancer patients and cardiovascular risks in patients with chronic kidney disease. European Medicines Agency Post-authorisation Evaluation of Medicines for Human Use. 2007 Oct 23. URL: https://www.ema.europa.eu/en/documents/public-statement/public-statement-epoetins-risk-tumour-growth-progression-thromboembolic-events-cancer-patients_en.pdf [accessed 2014-02-07]
7. Modified dosing recommendations to improve the safe use of Erythropoiesis-Stimulating Agents (ESAs) in chronic kidney disease. US Food and Drug Administration Drug Safety Communication. 2011 Jun 26. URL: https://www.fda.gov/drugs/drug-safety-and-availability/fda-drug-safety-communication-modified-dosing-recommendations-improve-safe-use-erythropoiesis [accessed 2014-02-07]
8. Kidney Disease: Improving Global Outcomes (KDIGO) Work Group. Chapter 1: Diagnosis and evaluation of anemia in CKD. Kidney Int Suppl (2011) 2012 Aug;2(4):288-291 [FREE Full text] [doi: 10.1038/kisup.2012.33] [Medline: 25018948]
9. Medicare program; end-stage renal disease prospective payment system. Final rule. Federal Register. 2010 Aug 12. URL: https://www.govinfo.gov/content/pkg/FR-2012-11-09/pdf/2012-26903.pdf [accessed 2020-12-01]
10. Medicare program; end-stage renal disease prospective payment system and quality incentive program; ambulance fee schedule; durable medical equipment; and competitive acquisition of certain durable medical equipment prosthetics, orthotics and supplies. Final rule. Federal Register. 2011 Nov 10. URL: http://www.gpo.gov/fdsys/pkg/FR-2011-11-10/pdf/2011-28606.pdf [accessed 2020-11-01]
11. Medicare program; end-stage renal disease prospective payment system, quality incentive program, and durable medical equipment, prosthetics, orthotics, and supplies. Federal Register. 2013 Dec 02. URL: http://www.gpo.gov/fdsys/pkg/FR-2013-12-02/pdf/2013-28451.pdf [accessed 2020-12-01]
12. Chambers JD, Weiner DE, Bliss SK, Neumann PJ. What can we learn from the U.S. expanded end-stage renal disease bundle? Health Policy 2013 May;110(2-3):164-171. [doi: 10.1016/j.healthpol.2013.01.011] [Medline: 23419419]
13. Canadian Erythropoietin Study Group. Association between recombinant human erythropoietin and quality of life and exercise capacity of patients receiving haemodialysis. Canadian Erythropoietin Study Group. BMJ 1990 Mar 03;300(6724):573-578 [FREE Full text] [doi: 10.1136/bmj.300.6724.573] [Medline: 2108751]
14. Nissenson AR, Korbet S, Faber M, Burkart J, Gentile D, Hamburger R, et al. Multicenter trial of erythropoietin in patients on peritoneal dialysis. J Am Soc Nephrol 1995 Jan;5(7):1517-1529 [FREE Full text] [Medline: 7703390]
15. Bahlmann J, Schöter KH, Scigalla P, Gurland HJ, Hilfenhaus M, Koch KM, et al. Morbidity and mortality in hemodialysis patients with and without erythropoietin treatment: a controlled study. Contrib Nephrol 1991;88:90-106. [doi: 10.1159/000419519] [Medline: 2040200]
16. Collins A, Foley RN, Herzog C, Chavers B, Gilbertson D, Ishani A, et al. United States Renal Data System 2008 Annual Data Report. Am J Kidney Dis 2009 Jan;53(1 Suppl):S1-S374. [doi: 10.1053/j.ajkd.2008.10.005] [Medline: 19111206]
17. Statistics of Organ Donation in Year 2010.: Taiwan Organ Registry and Sharing Center; 2010. URL: https://www.torsc.org.tw/FileUploads/docatt/f650a9e5-f836-b099-eb21-ba5280745a1d.doc [accessed 2014-04-01]
18. Collins AJ, Foley RN, Herzog C, Chavers BM, Gilbertson D, Ishani A, et al. Excerpts from the US Renal Data System 2009 Annual Data Report. Am J Kidney Dis 2010 Jan;55(1 Suppl 1):S1-420, A6 [FREE Full text] [doi: 10.1053/j.ajkd.2009.10.009] [Medline: 20082919]

19. Quality Report of Outpatient Dialysis Global Budget 2007 Q4. National Health Insurance Administration.: Ministry of Health and Welfare; 2008. URL: https://tinyurl.com/y43m2b2c [accessed 2014-02-07]
20. Quality Report of Outpatient Dialysis Global Budget 2010 Q3. National Health Insurance Administration.: Ministry of Health and Welfare; 2011. URL: https://www.mohw.gov.tw/dl-44682-2644622d-3e42-495a-bbdc-ebb6e6d55092.html [accessed 2014-02-07]
21. Evaluation of the dialysis payment policies of the National Health Insurance. In: Commissioned research projects of Ministry of Health and Welfare. Taiwan: Ministry of Health and Welfare; 2012:1-94.
22. Lin L, Warren-Gash C, Smeeth L, Chen P. Data resource profile: the National Health Insurance Research Database (NHIRD). Epidemiol Health 2018;40:e2018062. [doi: 10.4178/epih.e2018062] [Medline: 30727703]
23. 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]
24. Austin PC. A comparison of 12 algorithms for matching on the propensity score. Stat Med 2014 Mar 15;33(6):1057-1069. [doi: 10.1002/sim.6004] [Medline: 24123228]
25. Liu J, Huang Z, Gilbertson DT, Foley RN, Collins AJ. An improved comorbidity index for outcome analyses among dialysis patients. Kidney Int 2010 Jan;77(2):141-151 [FREE Full text] [doi: 10.1038/ki.2009.413] [Medline: 19907414]
26. Kan W, Wang J, Wang S, Sun Y, Hung C, Chu C, et al. The new comorbidity index for predicting survival in elderly dialysis patients: a long-term population-based study. PLoS One 2013;8(8):e68748 [FREE Full text] [doi: 10.1371/journal.pone.0068748] [Medline: 23936310]
27. Brenner and Rector's The Kidney E-Book 9th Edition. Philadelphia: Saunders; Nov 01, 2011:2095.

Abbreviations

CKD: chronic kidney disease
CMS: US Centre for Medicare and Medicaid
ESRD: end-stage renal disease
Hb: hemoglobin
Hct: hematocrit
HR: hazard ratio
ICD-9: International Classification of Diseases, Ninth Revision
NHI: Taiwan National Health Insurance
QIP: quality incentive program
PS: propensity score

Edited by G Eysenbach; submitted 13.03.20; peer-reviewed by Z Chen, K Malale; comments to author 24.08.20; revised version received 12.09.20; accepted 23.11.20; published 17.12.20.

Please cite as:
Hou YH, Yang FJ, Lai IC, Lin SP, Wan TTH, Chang RE
Effects of Erythropoietin Payment Policy on Cardiovascular Outcomes of Peritoneal Dialysis Patients: Observational Study
JMIR Med Inform 2020;8(12):e18716
URL: http://medinform.jmir.org/2020/12/e18716/
doi:10.2196/18716
PMID:33331829

Original Paper

The Correlation of Online Health Information–Seeking Experience With Health-Related Quality of Life: Cross-Sectional Study Among Non–English-Speaking Female Students in a Religious Community

Zahra Kavosi[1*], PhD; Sara Vahedian[2*], BSc; Razieh Montazeralfaraj[2*], PhD; Arefeh Dehghani Tafti[3*], MSc; Mohammad Amin Bahrami[1*], PhD

[1]Health Human Resources Research Center, Department of Health Services Management, School of Management and Medical Informatics, Shiraz University of Medical Sciences, Shiraz, Iran

[2]Healthcare Management Department, Shahid Sadoughi University of Medical Sciences, Yazd, Iran

[3]Department of Biostatistics and Epidemiology, Kerman University of Medical Sciences, Kerman, Iran

[*]all authors contributed equally

Corresponding Author:
Mohammad Amin Bahrami, PhD
Health Human Resources Research Center, Department of Health Services Management
School of Management and Medical Informatics
Shiraz University of Medical Sciences
Almas Building, Alley No. 29, Ghasr e Dasht St.
Shiraz, 7175644654646
Iran
Phone: 98 9132565057
Fax: 98 7132340781
Email: aminbahrami1359@gmail.com

Abstract

Background: Given the increasing availability of the internet, it has become a common source of health information. However, the effect of this increased access on health needs to be further studied.

Objective: This study aimed to investigate the correlation between online health information–seeking behavior and general health dimensions in a sample of high school students in Iran.

Methods: A cross-sectional study was conducted in 2019. A total of 295 female students participated in the study. The data were collected using two validated questionnaires: the e-Health Impact Questionnaire and the 36-Item Short Form Health Survey. The collected data were analyzed through descriptive statistics and Pearson correlation coefficients using SPSS version 23 (IBM Corp).

Results: The participants moderately used online information in their health-related decisions, and they thought that the internet helped people in health-related decision making. They also thought that the internet could be used to share health experiences with others. Participants had moderate confidence in online health information and stated that the information provided by health websites was moderately understandable and reliable and moderately encouraged and motivated them to play an active role in their health promotion. Nevertheless, the results showed that online health information–seeking experience had no significant correlation with health-related quality of life.

Conclusions: This study provides insights into the effect of using internet information on the health of adolescents. It has important implications for researchers and policy makers to build appropriate policies to maximize the benefit of internet access for health.

(JMIR Med Inform 2020;8(12):e23854) doi:10.2196/23854

KEYWORDS

general health; SF-36; information seeking behavior; online health information; high school students; health literacy

Introduction

Adolescence refers to the age range of 10 to 19 years [1]. It is generally supposed that this period is an appropriate time to maintain and promote health and prevent health-related adverse effects in the following decades of life [2]. Despite this potential, adolescents have special needs that are often not well met by health systems [3]. Evidence suggests that many high-risk behaviors that usually begin in adolescence cause an epidemic of noncommunicable diseases in adulthood [4]; an 18-year prospective study has shown that physical activity in adolescence has a significant effect on one's health in adulthood [5]. Today, adolescents are facing multiple health-threatening factors, various questions on different aspects of health, and more complicated health challenges and problems than their parents did [6]. Studies on adolescent health status highlight the necessity of changing the assumption that adolescents are generally healthy and need less attention [7]. Therefore, the question is where adolescents can receive help or information when faced with such challenges. Family, peers, teachers, health specialists, and online resources are common sources from which adolescents seek information and advice on health challenges [8].

In general, people choose different ways to find answers to their questions and doubts about health. Health information–seeking behavior refers to seeking and receiving information to reduce uncertainty and doubts and ensure health status [9]. As Wilson suggests in his model, "information-seeking behaviour arises as a consequence of a need perceived by an information user, who, in order to satisfy that need, makes demands upon formal or informal information sources or services, which result in success or failure to find relevant information" [10].

Similar to most fields, health information seeking has changed from traditional practices such as referring to books and magazines and even direct expert advice to new methods such as the use of the internet and social networks. Online resources play an important role in providing health information, and young people are increasingly using online information in various domains. In their systematic review, Park and Kwon (2018) showed that adolescents used the internet widely in different countries [11].

According to another systematic review, 81% (21/26) of the studies indicated that more than 50% of their samples used the internet to obtain health information [12]. Studies show that adolescents use the internet to find answers to their wide range of health-related questions; on the other hand, they doubt the comprehensibility and validity of online information [12-14].

Johnson et al (2015) found that youth with lower mental quality of life used the internet more to gain health information [14]. Besides, studies have shown that adolescents with more health risk factors and those with worse health status, higher health literacy, and a chronic disease are more likely to use the internet to search for health information [15]. In this regard, a question that has remained as a main concern is whether adolescents have sufficient ability to effectively search for, evaluate, and use online health information in a way that promotes their health [16,17]. Thus, the adolescents' ability to access health information online can be described as a double-edged sword that may have a positive or negative impact on their health.

According to the 2016 census in Iran, adolescents make up 8% of the country's population of 12 million, half of whom are girls [18]. Iran has one of the highest rates of internet access in its region [19]. Since a high percentage of the Iranian population is composed of adolescents and youths, and due to the cultural and religious contexts of the country, some of the challenges that adolescents face are not disclosed to their parents or professionals. Therefore, the internet seems to provide an opportunity through which they can seek answers to their health-related questions. Hence, this study aimed to investigate the relationship between online health information–seeking behavior and general health status on a sample of high school girls in Iran. We were particularly interested in studying the online health information–seeking behavior and its correlations with health outcomes among female students for several reasons. First, according to statistics from the Ministry of Education, girls make up half of all Iranian students [20]. Second, adolescence is a critical period of life regarding health, especially for health-promoting behaviors. Statistics show that one fifth of the world's population is between the ages of 10 and 19 years, and 85% of them live in developing countries. Promoting adolescents' health is one of the national development goals, and satisfying the health needs of this population is among the top priorities of health systems around the world. Changing adolescents' health-related behaviors and their lifestyles requires providing them with appropriate and complete health information [21]. Third, girls play an important role in the health of today's and future society, and investment in improving their health is one of the most important strategies to achieve global health goals [18]. The fourth reason is that there is a growing body of research that explores the significance of context in health information, demonstrating that gender is a determinant of information-seeking behavior. Many authors agree that health information seeking is influenced by gender [22]. In a study by Rowley et al (2016), they confirmed gender as a factor influencing the process of health information seeking and evaluation [23]. In addition, some other studies have reported important gender differences in health information–seeking behavior [22-26]. Therefore, it is crucial for societies to help female students to maintain and promote their health, which was the aim of this study as well.

Methods

Overview

This cross-sectional questionnaire study was conducted in 2019. A total of 295 female high school students in Ardekan city, Yazd province, who had access to the internet and the experience of health information seeking participated in the study. All participants provided informed consent to participate in the study and were assured that their personal information would be kept confidential. The parents of the students were made aware of the participation of their children in the study and had the opportunity to not let their children participate in the study. The school principal and students' teachers approved the study. All the study procedures were conducted in accordance with

the ethical standards of the Declaration of Helsinki. In addition, the ethics committee of Shahid Sadoughi University of Medical Sciences approved the study (approval code: IR.SSU.SPH.REC.1399.023). Questionnaires were completed in class, and any students who were absent on the testing days had the opportunity to participate in the study on the following days. All the data were gathered using two validated questionnaires: the e-Health Impact Questionnaire (eHIQ) and the 36-Item Short Form Health Survey (SF-36).

eHIQ

The eHIQ was used to measure the online health information–seeking behavior of participants. The eHIQ, developed by Kelly et al in 2015 as an instrument to measure the potential consequences of using websites containing different types of material across a range of health conditions, is a 2-part instrument with 37 items. eHIQ-Part 1 consists of 11 items related to general views of using the internet in relation to health. These 11 items have been grouped into 2 subscales named "Attitudes towards online health information" (5 items) and "Attitudes towards sharing health experiences online" (6 items). eHIQ-Part 2 consists of 26 items related to the consequences of using specific health-related online sources. The 26 items have also been grouped into 3 subscales: "Confidence and identification" (9 items), "Information and presentation" (8 items), and "Understanding and motivation" (9 items). In our study, the participants were asked to respond to the 26 items of eHIQ-Part 2 regarding the online sources from which they have sought information in recent months. In addition, the participants were asked to score all items from both parts on a 5-point scale ranging from 1 ("never") to 5 ("always"). We used a standard "forward-backward" procedure to translate the eHIQ from English into Persian. To demonstrate the content validity, we used the content validity ratio to quantify the extent of the experts' agreement. The reliability of the translated version of the eHIQ was confirmed using the Cronbach alpha coefficient, which was calculated as 0.89 for the total scale and 0.81, 0.87, 0.94, 0.83, and 0.91 for "Attitudes towards online health information," "Attitudes towards sharing health experiences online," "Confidence and identification," "Information and presentation," and "Understanding and motivation," respectively.

SF-36

The SF-36 is a popular instrument for assessing the health-related quality of life. The SF-36 has 36 items, which measure 8 subscales (ie, vitality, physical functioning, bodily pain, general health perceptions, physical role functioning, emotional role functioning, social role functioning, and mental health). These 8 subscales of SF-36 are grouped into two distinct dimensions, namely a physical dimension represented by the physical component summary (PCS), which is the sum of physical functioning, bodily pain, general health perceptions, and physical role functioning, and a mental dimension represented by the mental component summary (MCS), which is the sum of vitality, emotional role functioning, social role functioning, and mental health. After completing the questionnaire, each scale is directly transformed into a 0-100 score on the assumption that each question carries equal weight. The lower the score, the greater the disability; the higher the score, the less the disability (ie, a score of 0 is equivalent to maximum disability and a score of 100 is equivalent to no disability). In this study, we used the Persian version of the SF-36, which had been validated by Montazeri et al (2005) [27]. In addition, we used the original scoring system. The collected data were analyzed through descriptive statistics (including means and standard deviations) and Pearson correlation coefficients, using SPSS version 23 (IBM Corp).

Results

Of the participants, 16 students were married, and the rest were single. All of them had access to the internet at their home and the experience of seeking health information in recent months before the study. Demographic characteristics of the participants are presented in Table 1.

Table 1. Demographic characteristics of the participants (N=295).

Variable	n (%)
Marital status	
Single	279 (94.6)
Married	16 (5.4)
Religion	
Muslim	279 (94.6)
Not available	16 (5.4)
Education level of parents	
High school	
Fathers	104 (35.3)
Mothers	116 (39.3)
Diploma and associate degree	
Fathers	122 (41.4)
Mothers	108 (36.6)
Bachelor and higher	
Fathers	69 (23.4)
Mothers	71 (24.1)

The findings regarding information-seeking behavior of the participants are presented in Table 2, showing that the participants have moderate scores on all subscales of eHIQ-Part 1 and Part 2. In this study, mean scores between 1 and 2.33, between 2.34 and 3.66, and higher than 3.66 were defined as low, moderate, and high levels, respectively. The moderate scores obtained by the participants in the 2 subscales of eHIQ-Part 1 indicated that the participants had used the internet moderately in their health-related decisions and thought that internet could be moderately useful to help people in their health-related decision making. They also thought that internet was a moderately good channel to share the health experiences and communicate with some people with the same health problems. In addition, the moderate score of participants regarding confidence and identification revealed that they did not have a sense of solidarity with other internet users in their information-seeking journey; the internet did not give them a sense of confidence to explain their health issues to others, and they thought that online searching did not help them to better manage their health-related conditions. Therefore, they did not highly value the online health information. The moderate scores of the participants regarding the last 2 subscales of eHIQ-Part 2, “Information and presentation” and “Understanding and motivation,” showed that the information provided by health websites had been moderately understandable and reliable for the participants and moderately encouraged and motivated them to play an active role in their health promotion.

Table 2. Mean scores for online health information–seeking behavior of the students.

Item	Mean score (SD)
eHIQ-Part 1	
Attitudes towards online health information	2.46 (0.80)
Attitudes towards sharing health experiences online	2.77 (0.90)
eHIQ-Part 2	
Confidence and identification	2.52 (0.77)
Information and presentation	2.90 (0.79)
Understanding and motivation	2.90 (0.88)
eHIQ (total)	2.71 (0.71)

The descriptive results regarding the students' health statuses on the SF-36 subscales are presented in Table 3. As shown in this table, the participants had moderate to good scores on the SF-36 subscales. They obtained the highest and lowest scores in physical functioning and emotional role functioning, respectively.

Table 3. SF-36 scores of the students.

Item	Mean score (SD)
Physical functioning	83.67 (15.00)
Physical role functioning	75.94 (26.65)
Bodily pain	71.84 (23.27)
General health perception	63.31 (19.53)
Emotional role functioning	56.01 (38.58)
Vitality	75.94 (26.65)
Social role functioning	70.25 (25.34)
Mental health	65.29 (22.54)
Physical component summary	72.90 (16.20)
Mental component summary	63.19 (22.26)

The correlation coefficients of online health information–seeking behavior and its subscales with the main SF-36 subscales are presented in Table 4. Based on the findings presented in this table, eHIQ and its subscales showed no statistical correlation with SF-36 subscales. These findings suggest that seeking health information through online sources does not improve health-related quality of life. This could have several explanations. In the Discussion section, these explanations are discussed and suggestions are provided.

Table 4. Correlations of online health information–seeking subscales with health status.

Item	PCS		MCS	
	r	*P* value	r	*P* value
Attitudes towards online health information	0.04	.51	0.04	.55
Attitudes towards sharing health experiences online	0.05	.42	0.04	.50
Confidence and identification	0.02	.69	0.02	.67
Information and presentation	0.05	.38	0.05	.41
Understanding and motivation	0.03	.65	0.01	.84
eHIQ (total)	0.04	.46	0.04	.53

Discussion

Principal Findings

This study aimed to examine the correlation of online health information–seeking behavior with health-related quality of life in a sample of Iranian female students. Results showed that the participants used online information moderately in their health-related decisions and thought that the internet helped people in health-related decision making and could be used to share health experiences with others. Participants had a moderate amount of confidence in online health information. They stated that the information provided by health websites was moderately understandable and reliable, and it moderately encouraged them to play an active role in their health promotion.

Use of the internet to access health information has increased in recent years for reasons such as accessibility, high volume of information disseminated, confidentiality, low cost, multimedia capabilities, and the ability to interact and gain support [19,21,28]. Reports indicate that adolescents are increasingly spending their time on using the internet. Using the internet is part of young people's daily activities, and they acquire and enhance many life skills, including health management, through online information [28].

A US national survey has found that 75.0% (907/1209) of online teens search health information [29]. A study in the United States has also reported that 98.0% (200/204) of youth 12 years and older use online resources to search for health information [30]. Another survey at two US educational institutes [31], a study at three Ghanaian universities [28], a study involving international students in East Asia [32], and a study at six colleges in Oman have reported similar results [33]. Therefore, although internet access is still limited in some countries [34,35], it seems that the internet is increasingly becoming one of the main information sources in the majority of countries.

In Iran, as in other countries, using the internet for health-related purposes has increased in recent years. A survey of adolescents in Shiraz, Iran, has shown that the internet is among the top sources of the respondents' health information, with 88% (352/400) using the internet to find a kind of health information [36]. Two other studies in Tehran high schools have reported similar rates [37,38]. Another study on students aged 15-18 years from different schools in Isfahan [21], a study involving 430 students from Gonabad University [39], two other studies at Gorgan and Kermanshah universities [40,41], and two other

studies at Tabriz University and Tehran University of Medical Sciences have reported similar results [19,42].

Overall, it seems that use of the internet as a source of health information is expanding; however, the review of the literature shows that searching for online health information is correlated with some variables such as age, gender, education level, skills and experience with internet use, health status, and availability and reliability of sources [1,31,43].

Adolescents often seek health information with different objectives and motives [36,40,42], and they typically seek information related to a variety of health subjects such as healthy eating, physical activity, exercise, weight control, risks and complications of disease treatments, sexual and reproductive health, sexual and physical abuse, consumption of alcohol and other substances, tobacco use, mental health, accidents and injuries, health care providers, and support groups [21,29,31,34,42].

Due to the increasing use of the internet for health purposes, many studies have been conducted on online health information–seeking behavior in different demographic groups, including students. Most of these studies have examined the sources of health information used by different groups, attitudes towards health information seeking, aims and motivations, types of information sought, and factors related to health information–seeking behavior [36]. However, few studies have examined the actual effect of accessing online health information on health status. In fact, the question of whether online health information–seeking behavior significantly affects health status or not has largely remained unanswered. Therefore, this study aimed to explore the online health experience of Iranian female students and its correlation with their health-related quality of life.

The findings showed that the majority of the participants had good or somewhat good general health status. Numerous studies have been conducted on the general health status of adolescents in Iran; most of them have reported approximately similar findings [18].

In addition, the descriptive findings of the study regarding online health information–seeking behavior showed that the participants had moderate scores on all subscales of eHIQ.

Regarding attitudes about online health information and sharing them, a similar study that aimed at explaining health information behavior of adolescents in Shiraz has reported that the participants' general attitude toward health information retrieved from the internet is positive. The majority of the participants also trusted in the quality of information and were interested in retrieving health information from the internet twice [36]. Another study at Tabriz University has reported that the internet is considered one of the trustable sources of health information by participants [42]. At the same time, a study in Isfahan schools has shown that 47.7% (3110/6519) of those who did not use the internet to search for health information reported a lack of trust in the internet information as the main cause of their decision not to be an online health information seeker [21]. Regarding the sharing of health information, a study in United States has found that although 98.0% (200/204) of the participants were online health information seekers, only 51.5% (105) of them shared their health information and only 25% (51) of them thought that social media could provide usable health information. This study also reported that women had shared their health information more than men, and adolescents between the ages of 12 and 14 years had shared more than other age groups. People with poor self-reported health and those who thought online sources could help them in accessing health information were also more likely to share their health information [30]. Another study, which was conducted in India, reported that most of its respondents shared online health information with their friends and family [44]. In summary, based on the available literature, it seems that trust in online health information and interest in sharing it are different across different socioeconomic contexts. The participants of our study also thought that information provided by health websites was moderately understandable. In this regard, many studies have reported poor understandability of internet information as one of the main challenges for online users.

This study was conducted among a non–English-speaking female sample in a developing religious community. The unique features of the research environment may affect the results. Several studies show that contextual factors may affect different aspects of information-seeking behavior. Dankasa (2017) found in a study that geographical location, culture, and religious status may influence the information-seeking behavior of the internet users [45]. Lee and Cho (2011) and Chang and Lee (2001) have also reported the same results [46,47]. Based on the findings of these studies, contextual factors may encourage, determine, or prevent information-seeking behavior [45]. In addition, Lee and Cho (2011) found that social and cultural affiliations of individuals influence the way they choose to exchange information [46]. Therefore, our findings regarding the attitude toward online information and attitude toward sharing the information could be affected by the specific context of the study. Furthermore, this study was conducted among a sample of female students. Various studies have demonstrated that demographic variables such as gender and age, together with other factors such as income and education level, may influence health information behavior. Among these factors, gender has been widely identified as a factor affecting health information behavior. Accordingly, most studies suggest that being female and younger is associated with more frequent health-related use of the internet, although a few studies have reported contradictory findings [23,25,26,48]. The findings of this study can also be discussed based on the participants' native language. Few studies have investigated information-seeking behavior of non–English language speakers or information-seeking behavior using non-native language. Although an increasing number of databases have now been created and made available in other languages, including Persian, English is still the dominant language of online information. Searching in different languages might affect different aspects of information-seeking behavior such as understanding of retrieved information, interpretation, evaluation, and the relevant judgment [49]. In this regard, some studies have reported differences between information seeking in different languages [49], while some have not confirmed the same differences [50]. There is no doubt that the users' language skills can affect their information-seeking behavior. In this

study, it seems that all the studied subscales of information-seeking behavior include attitude toward online information, attitude toward sharing of information, confidence and trust, attitude toward the presentation, and understanding of information. It is notable that many studies have identified that the users' attitude has a positive effect on their health information–seeking behavior, similar to the trust they placed in the information [23,26]. Therefore, it should be a priority to improve the attitudes of our participants toward their confidence in online health information.

Statistical tests also showed that different dimensions of online health information–seeking behavior had no significant correlations with health-related quality of life. On this subject, in a survey of 400 school-age adolescents in Shiraz, respondents stated that they believed that the retrieved online health information affected their health status positively [36]. In another study at Tabriz University, the participants approved of the effects of their online health information seeking on some health-related behaviors [43]. A study among Nigerian students found that only 50% (20/400) of participants consulted with a physician about their health after searching online health information [34]. A study at three universities in Ghana also reported that 72.4% (315/435) of respondents used retrieved online information as a basis for lifestyle modifications, and 73.6% (320) of the students stated that access to online health information improved or partially improved their health status, while 1.1% (5) said that using the internet had no effect on their overall health [28].

Overall, it seems that although internet technology has provided a good opportunity to access health information, its practical impact on health status is still controversial. This can have many explanations. Challenges such as the lack of appropriate information, inadequate quality of information, poor health literacy of internet users, insufficient skills in searching for information, lack of trust in online health information sources, and concerns about security and confidentiality reduce the potential of the internet in serving the health of population [21,29,42]. The production and dissemination of health misinformation is also a serious concern. Today, a great deal of health misinformation is also produced and published online, which is potentially a threat to public health [51]. Low internet access is also an infrastructure challenge in some parts of the world [34]. Therefore, it is necessary to formulate and apply improvement strategies to maximize the health benefits of internet. These strategies can be formulated in two levels: supply-side strategies (eg, expanding internet access; providing high-quality, appropriate, and understandable information; monitoring online health content; engaging health professionals in producing evidence-based information; ensuring safety; paying attention to legal issues; and focusing on adolescent health priorities [21,28,34,36]) and demand-side strategies (eg, investigating the patterns of use, improving health literacy, training search and information validation skills, and enhancing information behavior [29,34,36,40]).

Based on the findings of this study, interventions such as encouraging students to make more use of the internet as a source of health information; expanding their access to reliable online health sources; launching specific students' health websites containing relevant, reliable, and understandable information by health authorities, especially in native language; improving the English language skills of students (since it could be a barrier for most of the participants in searching activities); improving students' internet skills; and familiarizing them with search methods and specialized sources can be prioritized in order to maximize the potential of use of the internet in promoting the students' health. It is also helpful to strengthen the online culture by using social marketing in the school environment. This study has several strengths. Few studies have been conducted in Iran to investigate the correlations between online health information–seeking behavior and the health status of students. In addition, there are few studies investigating the health information–seeking behavior of Persian language speakers. Therefore, the study has implications for research and practice. It contributes to research on health information–seeking behavior as it brings out the association of health information seeking with health outcomes that has not been given much attention in the literature. In addition, the study provides health and information professionals with information needed to make health information understandable, available, and accessible for students. The findings could also be used to develop appropriate interventions to enhance the students' internet skills, so that they can make the best use of internet technology to promote their health. The study, however, has some limitations; first, it used a sample of female students, while some studies have reported gender-based differences in health information–seeking behavior that may affect the generalization of our findings to other population groups. Also, the study was done in a specific geographical, cultural, and religious context, which also makes it difficult to generalize the findings to different contexts. The results described have been extracted from research in a developing country, and it is likely that there are differences between countries.

Conclusion

Students have a variety of health issues and have an increased demand for health information [36]. In the online era, the landscape of health information has changed, and the internet has increasingly become the main source of health information [52]. As Smith et al have pointed out, the question is no longer whether the internet can be an important source of information or not, but how its potential can be maximized [29].

Although students' access to online sources has increased substantially, they can only gain the most benefit from this information source by being able to effectively search for, evaluate, and use online information [29]. Moving forward, various stakeholders, including policymakers, information producers, health professionals, teachers, parents, and students themselves, should play their role well. Our study demonstrated the online health information–seeking behavior of a sample of female students in an Islamic developing country. Findings reported here have implications for communities with the same sociocultural status, although it can have lessons for other communities as well.

Acknowledgments

The authors acknowledge the participants, their parents, and their teachers.

Conflicts of Interest

None declared.

References

1. WHO. The second decade: improving adolescent health and development. Geneva: World Health Organization; 2001. URL: https://apps.who.int/iris/bitstream/handle/10665/64320/WHO_FRH_ADH_98.18_Rev.1.pdf [accessed 2020-08-19]
2. Centers for Disease Control and Prevention. STD Surveillance. 2004. URL: https://www.cdc.gov/std/stats/archive/2004SurveillanceAll.pdf [accessed 2020-11-18]
3. Beaglehole R, Bonita R, Horton R, Adams C, Alleyne G, Asaria P, Lancet NCD Action Group, NCD Alliance. Priority actions for the non-communicable disease crisis. Lancet 2011 Apr 23;377(9775):1438-1447. [doi: 10.1016/S0140-6736(11)60393-0] [Medline: 21474174]
4. UN. Prevention and control of non-communicable disease. New York, NY: United Nations; 2010. URL: https://www.who.int/westernpacific/activities/preventing-and-controlling-noncommunicable-diseases [accessed 2020-07-29]
5. World Health Organization. HIV and adolescents: HIV testing and counseling, treatment and care for adolescents living with HIV: policy brief. 2013. URL: https://apps.who.int/iris/handle/10665/94561 [accessed 2020-07-16]
6. Raphael D. Determinants of health of North-American adolescents: evolving definitions, recent findings, and proposed research agenda. J Adolesc Health 1996 Jul;19(1):6-16. [doi: 10.1016/1054-139X(95)00233-I] [Medline: 8842855]
7. Dick B, Ferguson BJ. Health for the world's adolescents: a second chance in the second decade. J Adolesc Health 2015 Jan;56(1):3-6. [doi: 10.1016/j.jadohealth.2014.10.260] [Medline: 25530601]
8. Rickwood DJ, Deane FP, Wilson CJ. When and how do young people seek professional help for mental health problems? Med J Aust 2007 Oct 01;187(S7):S35-S39. [doi: 10.5694/j.1326-5377.2007.tb01334.x] [Medline: 17908023]
9. Kitikannakorn N, Sitthiworanan C. Searching for health information on the Internet by undergraduate students in Phitsanulok, Thailand. Int J Adolesc Med Health 2009;21(3):313-318. [doi: 10.1515/ijamh.2009.21.3.313] [Medline: 20014634]
10. Wilson T. Models in information behaviour research. J Doc 1999;55(3):249-270 [FREE Full text] [doi: 10.1108/EUM0000000007145]
11. Park E, Kwon M. Health-Related Internet Use by Children and Adolescents: Systematic Review. J Med Internet Res 2018 Apr 03;20(4):e120 [FREE Full text] [doi: 10.2196/jmir.7731] [Medline: 29615385]
12. Henderson E, Keogh E, Rosser B, Eccleston C. Searching the internet for help with pain: adolescent search, coping, and medication behaviour. Br J Health Psychol 2013 Feb;18(1):218-232. [doi: 10.1111/bjhp.12005] [Medline: 23126577]
13. Selkie EM, Benson M, Moreno M. Adolescents' Views Regarding Uses of Social Networking Websites and Text Messaging for Adolescent Sexual Health Education. Am J Health Educ 2011 Dec;42(4):205-212 [FREE Full text] [Medline: 22229150]
14. Johnson KR, Fuchs E, Horvath KJ, Scal P. Distressed and looking for help: Internet intervention support for arthritis self-management. J Adolesc Health 2015 Jun;56(6):666-671. [doi: 10.1016/j.jadohealth.2015.02.019] [Medline: 26003583]
15. Sbaffi L, Zhao C. Modeling the online health information seeking process: Information channel selection among university students. Journal of the Association for Information Science and Technology 2019 Apr 13;71(2):196-207 [FREE Full text] [doi: 10.1002/asi.24230]
16. Wong DK, Cheung M. Online Health Information Seeking and eHealth Literacy Among Patients Attending a Primary Care Clinic in Hong Kong: A Cross-Sectional Survey. J Med Internet Res 2019 Mar 27;21(3):e10831 [FREE Full text] [doi: 10.2196/10831] [Medline: 30916666]
17. WHO. Health for the world's adolescents: A second chance in the second decade. URL: https://www.who.int/maternal_child_adolescent/documents/second-decade/en/ [accessed 2020-07-11]
18. Arsanjani SA, Javadifar N, Javadnoori M, Haghighi ZM. A Study of Health-Related Quality of Life among Female High Schools Adolescents in Ahvaz in 2014. Journal of Rafsanjan University of Medical Sciences 2016;14(8):643-654 [FREE Full text]
19. Fayaz-bakhsh A, Khajeh KR, Soleymani NM, Rahimi F, Jahangiri L, Heydari S. The Internet Using and Health: Students' Knowledge, Attitude and Lifestyle Related to the Internet. Hakim Research Journal 2011;14(2):96-105 [FREE Full text]
20. Iran Ministry of Education. The students' statistics. URL: https://www.medu.ir/fa/ [accessed 2020-11-26]
21. Esmaeilzadeh S, Ashrafi-Rizi H, Shahrzadi L, Mostafavi F. A survey on adolescent health information seeking behavior related to high-risk behaviors in a selected educational district in Isfahan. PLoS One 2018;13(11):e0206647 [FREE Full text] [doi: 10.1371/journal.pone.0206647] [Medline: 30403763]
22. Hiebert B, Leipert B, Regan S, Burkell J. Rural Men's Health, Health Information Seeking, and Gender Identities: A Conceptual Theoretical Review of the Literature. Am J Mens Health 2018 Jul;12(4):863-876 [FREE Full text] [doi: 10.1177/1557988316649177] [Medline: 27170674]
23. Roley J, Johnson J, Sbaffi L. Gender as an influencer of online health information?seeking and evaluation behavior. Journal of the Association for Information Science and Technology 2016;68(1):36-47 [FREE Full text] [doi: 10.1002/asi.23597]

24. Al‐Muomen N, Morris A, Maynard S. Modelling information‐seeking behaviour of graduate students at Kuwait University. Journal of Documentation 2012 Jul 20;68(4):430-459. [doi: 10.1108/00220411211239057]
25. Ashkanani H, Asery R, Bokubar F, AlAli N, Mubarak S, Buabbas A, et al. Web-Based Health Information Seeking Among Students at Kuwait University: Cross-Sectional Survey Study. JMIR Form Res 2019 Oct 31;3(4):e14327 [FREE Full text] [doi: 10.2196/14327] [Medline: 31473592]
26. Wang J, Xiu G, Shahzad F. Exploring the Determinants of Online Health Information-Seeking Behavior Using a Meta-Analytic Approach. Sustainability 2019;11(17):4604 [FREE Full text] [doi: 10.3390/su11174604]
27. Montazeri A, Ghastasebi A, Vahdaninia MS. Validity and reliability of Persian version of standard SF-36 questionnaire. Payesh 2005;5(1):49-56 [FREE Full text]
28. Osei Asibey B, Agyemang S, Boakye Dankwah A. The Internet Use for Health Information Seeking among Ghanaian University Students: A Cross-Sectional Study. Int J Telemed Appl 2017;2017:1756473 [FREE Full text] [doi: 10.1155/2017/1756473] [Medline: 29225620]
29. Gray NJ, Klein JD, Noyce PR, Sesselberg TS, Cantrill JA. The Internet: a window on adolescent health literacy. J Adolesc Health 2005 Sep;37(3):243. [doi: 10.1016/j.jadohealth.2004.08.023] [Medline: 16109345]
30. Hausmann JS, Touloumtzis C, White MT, Colbert JA, Gooding HC. Adolescent and Young Adult Use of Social Media for Health and Its Implications. J Adolesc Health 2017 Jun;60(6):714-719 [FREE Full text] [doi: 10.1016/j.jadohealth.2016.12.025] [Medline: 28259620]
31. Escoffery C, Miner KR, Adame DD, Butler S, McCormick L, Mendell E. Internet use for health information among college students. J Am Coll Health 2005 Jan;53(4):183-188. [doi: 10.3200/JACH.53.4.183-188] [Medline: 15663067]
32. Chuang C. Effect of health information seeking behavior on anxiety and loss among east Asian international students. A Dissertation presented to the Faculty of the Graduate School at the University of Missouri. URL: https://www.semanticscholar.org/paper/Effects-of-health-information-seeking-behaviors-on-Chuang/90229ecfea633da619ea92b8894cdf34d3db24bb [accessed 2020-07-28]
33. Sultan K, Joshua V, Misra U. Health information seeking behavior of college students in the sultanate of Oman. Khyber Med Univ J 2017;9(1):8-14 [FREE Full text]
34. Obasola OI, Agunbiade OM. Online Health Information Seeking Pattern Among Undergraduates in a Nigerian University. SAGE Open 2016 Mar 09;6(1):215824401663525-215824401663550. [doi: 10.1177/2158244016635255]
35. Ybarra ML, Emenyonu N, Nansera D, Kiwanuka J, Bangsberg DR. Health information seeking among Mbararan adolescents: results from the Uganda Media and You survey. Health Educ Res 2008 Apr;23(2):249-258. [doi: 10.1093/her/cym026] [Medline: 17639121]
36. Bigdeli Z, Hayati Z, Heidari G, Jowkar T. Place of internet in health information seeking behavior: Case of young internet users in Shiraz. Human information interaction 2017;3(1):68-78 [FREE Full text]
37. Ranjbar Z, Darvizeh Z, Naraghizadeh A. The comparison of the rate and type of internet use regarding the psychological health and educational performance among the students of Tehran city. Psychological studies 2012;7(2):11-35 [FREE Full text] [doi: 10.22051/psy.2011.1547]
38. Nainian M, Adabbdoust F, Khatibi S, Ghomian F. Internet use and its relationship with mental health and quality of life among secondary school students. Clinical psychology and personality 2017;14(2):103-113 [FREE Full text] [doi: 10.22070/14.2.103]
39. Dastani M, Mokhtarzadeh M, Nasirzadeh A, Delshad A. Health information seeking behavior among students of Gonabad University of Medical Sciences? Library Philosophy and Practice (e-journal) 2019 [FREE Full text]
40. Zare A, Rahimi S, Soofi K. The study of the information seeking behavior of health literacy among students of Razi University of Kermanshah. Journal of Health Literacy 2017;2(2):63-72 [FREE Full text]
41. Soleymani NM, Shams M, Charkazi A, Rahimi F, Fayaz-bakhsh A, Goudarzi F. Effects of Internet Use on Lifestyle of University Students in Gorgan, Iran. Health research journal 2013;8(5):834-843 [FREE Full text]
42. Kalankesh LR, Mohammadian E, Ghalandari M, Delpasand A, Aghayari H. Health Information Seeking Behavior (HISB) among the University Students. Front Health Inform 2019 Jul 08;8(1):13. [doi: 10.30699/fhi.v8i1.189]
43. Esmaeilzadeh S, Ashrafi-Rizi H, Shahrzadi L, Mostafavi F. A survey on adolescent health information seeking behavior related to high-risk behaviors in a selected educational district in Isfahan. PLoS One 2018;13(11):e0206647 [FREE Full text] [doi: 10.1371/journal.pone.0206647] [Medline: 30403763]
44. Shubha H. Relationship between internet use and health orientation: a study among university students. J Commun Media Technol 2015;5:103-113 [FREE Full text]
45. Dankasa J. The Effects of Cultural, Geographical and Religious Factors on Information Seeking: A Contextual Study. International Journal of Information Science and Management 2017;15(1):127-147 [FREE Full text]
46. Lee J, Cho H. Factors affecting information seeking and evaluation in a distributed learning environment. Journal of Educational Technology & Society 2011;14(2):213-223 [FREE Full text]
47. Chang S, Lee Y. Conceptualizing context and its relationship to the information behavior in dissertation research process. Journal of Library and Information Science 2001;26(2):4-18 [FREE Full text]

48. Lee YJ, Boden-Albala B, Larson E, Wilcox A, Bakken S. Online health information seeking behaviors of Hispanics in New York City: a community-based cross-sectional study. J Med Internet Res 2014 Jul 22;16(7):e176 [FREE Full text] [doi: 10.2196/jmir.3499] [Medline: 25092120]
49. Al-Wreikat A, Rafferty P. Cross-language information seeking behavior English vs. Arabic. Library Review 2015;64(6/7):446-467.
50. Mahmoud EM, Khafaga SA. Information seeking behavior in Arabic and English: A case study of scholars at Shaqra University. Information Development 2019;35(3):351-361 [FREE Full text] [doi: 10.1177/0266666917721059]
51. Bahrami MA, Nasiriani K, Dehghani A, Zarezade M, Kiani P. Counteracting Online Health Misinformation: A Qualitative Study. MSHSJ 2019 Dec 20;4(3):230-239 [FREE Full text] [doi: 10.18502/mshsj.v4i3.2056]
52. Jacobs W, Amuta AO, Jeon KC. Health information seeking in the digital age: An analysis of health information seeking behavior among US adults. Cogent Social Sciences 2017 Mar 13;3. [doi: 10.1080/23311886.2017.1302785]

Abbreviations

eHIQ: e-Health Impact Questionnaire
MCS: mental component summary
PCS: physical component summary
SF-36: 36-Item Short Form Health Survey

Edited by C Lovis; submitted 26.08.20; peer-reviewed by H Jafari, CF Yen; comments to author 20.09.20; revised version received 20.10.20; accepted 15.11.20; published 02.12.20.

Please cite as:
Kavosi Z, Vahedian S, Montazeralfaraj R, Dehghani Tafti A, Bahrami MA
The Correlation of Online Health Information–Seeking Experience With Health-Related Quality of Life: Cross-Sectional Study Among Non–English-Speaking Female Students in a Religious Community
JMIR Med Inform 2020;8(12):e23854
URL: https://medinform.jmir.org/2020/12/e23854
doi:10.2196/23854
PMID:33263546

Original Paper

Predictors of Internet Use Among Older Adults With Diabetes in South Korea: Survey Study

Sunhee Park[1*], RN, PhD; Beomsoo Kim[2*], PhD

[1]Barun ICT Research Center, Yonsei University, Seoul, Republic of Korea

[2]Graduate School of Information, Yonsei University, Seoul, Republic of Korea

[*]all authors contributed equally

Corresponding Author:
Sunhee Park, RN, PhD
Barun ICT Research Center
Yonsei University
Yonsei-samsung Bldg, 7th Floor
50 Yonsei-ro, Seodamun-gu
Seoul, 03722
Republic of Korea
Phone: 82 2 2123 6694
Fax: 82 2 2123 8095
Email: sunny372@hanmail.net

Abstract

Background: Internet access in Korea has grown dramatically over the past two decades. However, disparities in internet use, referred to as the second level of the digital divide, persist.

Objective: This study aims to examine opportunity, motivation, and health variables that indicate internet use among older adults with diabetes.

Methods: Data were sourced from a nationally representative sample of people 65 years and older with diabetes (N=1919). Logistic regression was used to explore potential differences in predictor variables between internet users and nonusers.

Results: Only 306 of the 1919 (15.95%) participants in the sample used the internet. They were more likely to be younger (odds ratio [OR] 0.89, 95% CI 0.87-0.92), well-educated (OR 1.20, 95% CI 1.16-1.26), and able to afford leisure expenditures (OR 1.02, 95% CI 1.01-1.04). Additionally, they had more information and communications technology (ICT) training experience, were motivated to learn, volunteered, and reported good physical and cognitive function. Participation in ICT education and better health more positively correlated with a higher rate of internet use than did years of education or economic standing in older adults with diabetes.

Conclusions: To support older adults with diabetes in the internet age, policies and health care providers should focus on digital competency training as well as physical and cognitive function.

(JMIR Med Inform 2020;8(12):e19061) doi:10.2196/19061

KEYWORDS

digital divide; internet use; older adults; diabetes; health; internet; Korea

Introduction

Internet access has grown dramatically over the past two decades in Korea. However, disparities in internet use still persist [1,2]. This disparity is known as the second level of the digital divide, which refers to a gap in access (the first level), use (the second level), and outcomes (the third level) of information and communications technology (ICT). Digital competency enables older adults to live more convenient lives and plays an important role in maintaining quality of life, health care, independent living, and relationships and in reducing isolation [3,4].

With a rapid increase in Korea's older adult population, in which chronic diseases are prevalent, addressing aging-related problems is important [5]. Diabetes mellitus is one of the most common chronic diseases affecting lifestyle, and its prevalence is increasing worldwide. In Korea, 25.1% of older adults 65 year and older have diabetes, and their mortality rate due to diabetes or cerebrovascular disease is higher than the

Organisation for Economic Co-operation and Development average, partly because of the vulnerability related to preventing deaths from treatable conditions [2]. An unhealthy lifestyle contributes to diabetes to a great extent, and one of the mainstays of diabetes treatment and prevention is adopting a healthy lifestyle. As there is no cure for diabetes, recently, self-management by mobile health or eHealth has begun to play a vital role in the digital era.

Many systematic reviews and meta-analyses have indicated that eHealth tools are effective in self-management both for disease management and lifestyle changes in daily life [5-7], and limited internet use and low eHealth literacy can indirectly cause health problems [8]. Problems with eHealth literacy due to low cognitive function make it difficult for older adults to manage, prevent, and treat diseases. This in turn leads to health problems [9], poor management of chronic diseases [10], and lower participation in treatment interventions. Low eHealth literacy is also associated with medical service misuse, which can be fatal [11]. Furthermore, the second digital divide, the gap in internet use, alienates older adults, leading to losses in self-employment opportunities, social exchanges, advantageous purchases, and investments. It also contributes to health problems caused by social network loss [12].

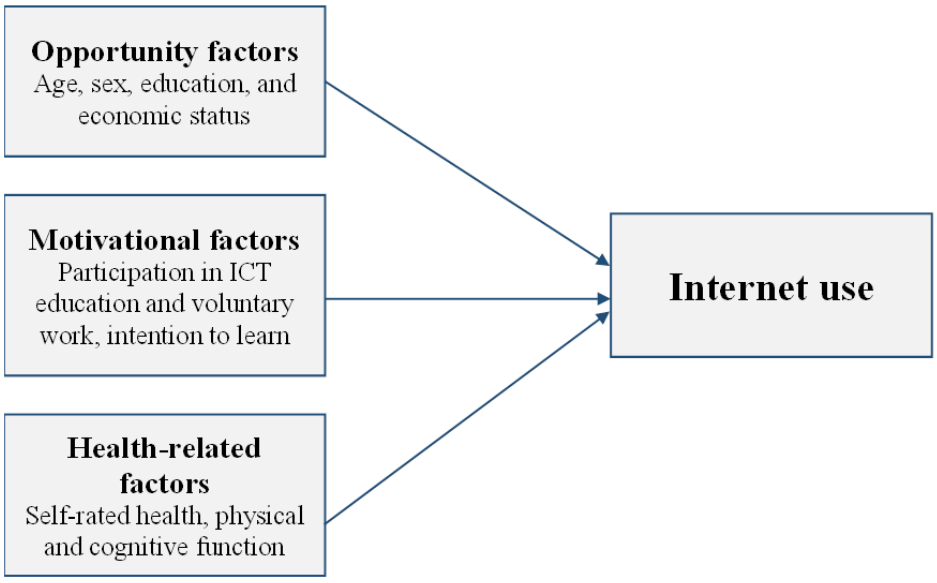
Internet underutilization by older adults is due primarily to limited opportunity and motivation [13]. Limited opportunity affects individuals who do not access the internet due to socioeconomic problems or lack of information. In a study of urban dwellers, only 27% of older adults were found to use computers, and age, years of education, occupation, income level, self-rated health, and volunteer work were the affecting factors [14]. Limited motivation indicates individuals who have not voluntarily chosen internet use and do not accept new technologies because they have no incentive or interest in them. In general, older adults lack ICT knowledge and skills and are often unaware of the need for it [15]. Moreover, older adults lack the confidence or support needed to learn how to use new equipment or acquire new knowledge. This low intention to acquire new knowledge results in a low level of internet use [16]. In addition to opportunities and motivation, aging and health problems involving physiological and cognitive functions also determine internet use, as do daily activities and chronic diseases [17,18]. Internet use has increased in Medicare-eligible patients but remains very low among the frailest older adults. Therefore, functional ability is more indicative of internet avoidance than chronic illness, self-rated health, or age [19].

Barriers to access and use include financial restrictions (ie, equipment and subscription costs are too high), medical and disability-related constraints (ie, the technology is not accessible or intuitive), and digital complexity (ie, accessing and navigating the internet is too complex) [20]. Scheerder et al [21] systematically reviewed 126 papers and distinguished 7 factors contributing to the digital divide: demographics, economics, social networks, cultural context, physical activity, home access and device availability, and attitudes toward online technology. Leisure activity and voluntary work were the affecting factors of internet use, and low levels of internet use affected social networking [12,21].

Although internet use among older adults is less prevalent than in the general public and is associated with aging or health problems [22], some older adults, such as those in the baby boomer generation, use the internet effectively because they are highly educated and were gradually exposed to smartphones and digital devices [20]. They use the internet to search for health-related information and exhibit confidence and satisfaction regarding eHealth [23].

As older adults are vulnerable to aging-related issues and chronic diseases, studies of internet usage among older adults with health problems or chronic diseases are needed. Furthermore, there is limited information on the predictors of internet use among older adults with diabetes, a chronic disease that demands continuous lifestyle modification and self-care. The aim of this study was to examine opportunity, motivation, and health-related factors that determine internet use among older adults with diabetes in South Korea (Figure 1).

Figure 1. Factors related to internet use. ICT: information and communications technology.



Methods

Design and Sample

The data for this study came from the 2017 Survey of Living Conditions and Welfare Needs of Korean Older Persons from the Korea Institute for Health and Social Affairs, which was based on a nationally representative sample of participants 65 years and older who were recruited using a stratified 2-stage cluster sampling design. The survey collected information through face-to-face interviews, and all participants provided written informed consent [2]. The sample for this study included 1919 (of 10,299) respondents with diabetes who reside in the community. The inclusion criteria were age of 65 years or older, official diagnosis of diabetes for more than 3 months with treatment, and response to a survey on internet use. We excluded individuals who did not respond to the survey on internet use and those younger than 65 years. The design was considered exempt from ethical review by the institutional review board of Yonsei University (approval no. 7001988-202001-HR-777-01E), as the data were anonymized.

Measurements

Internet use was assessed with 1 item: the use or nonuse of the internet or mobile phones to browse for information. Participants were asked, “Is it possible for you to use smart phones, computers, tablet PCs, and internet television to search for information?” to which they answered with either “yes” or “no.” Participants provided demographic and socioeconomic information such as age, sex, and years of education. Leisure activity expenditure was assessed according to monthly average expenditure on leisure activities in Korean won to determine participants’ economic standing [24]. Having previous or current volunteer experience was classified as either “yes” or “no.” Participation in ICT education was assessed by the question “Have you participated in ICT education during the last five years?” Participants responded with either “yes” or “no.” Intention to learn was measured on a 5-point Likert scale (1=no intention; 5=very eager to learn). Health-related factors included self-rated health, physical function, and cognitive function. Self-rated health was assessed by one question: “How do you feel about your health?” It was scored from 1 to 5 (1=not good at all; 5=very good). The higher the number, the higher the self-rated health score. Physical function was assessed using the 11-item Korean Instrumental Activities of Daily Living (K-IADL) questionnaire (ability to use a telephone, go shopping, prepare food, perform housekeeping and laundry, handle medication and finances, use transportation, and drive); total scores range from 11 and 33. The higher the score on the K-IADL, the lower the physical function [25]. A score of 33 on the K-IADL represents physical dependency. Cognitive function was assessed using the Korean version of the Mini-Mental State Examination for Dementia Screening (MMSE-DS); total scores range from 0 to 30. The higher the score, the better the cognitive function [26].

Data Analysis

Stata 15.1 (StataCorp) was used to conduct data analyses. Univariate analyses were performed to identify associations between internet use and factors related to opportunity, motivation, and health. Independent variables with significant group differences in the univariate analyses were included in a multivariate logistic regression analysis, which was performed to calculate the adjusted odds ratios (ORs) for internet users and nonusers.

Results

Of the 1919 respondents, only 306 (15.95%) used the internet to search for information (Table 1). Internet users were more likely to be male, younger, and more educated; have a higher leisure activity expenditure; volunteer more; have ICT education experience; have a lower intention to learn; and have better self-rated health, physical function, and cognitive function than internet nonusers.

Table 1. Participant characteristics (N=1919).

Characteristics and variables	Total	Internet users	Internet nonusers	*t* test[a] (*df*)	*F* test (*df*)	*P* value
Dependent variable						
Internet use, n (%)	1919 (100.0)	306 (16.0)	1613 (84.0)	N/A[b]	N/A	N/A
Opportunity factors						
Age (years), mean (SD)[c]	75.15 (5.66)	71.96 (0.27)	75.75 (0.14)	11.06 (1)	N/A	<.001
Gender, n (%)				N/A	64.27 (1)	<.001
Male	1312 (68.4)	269 (87.9)	1043 (64.7)			
Female	307 (31.6)	37 (12.1)	570 (35.3)			
Education (years), mean (SD)[d]	7.40 (4.64)	11.06 (0.21)	6.71 (0.11)	–16.03 (1)	N/A	<.001
Leisure expenditure (in ₩10,000)[e], mean (SD)	6.08 (11.00)	14.18 (1.06)	4.55 (0.20)	–14.83 (1)	N/A	<.001
Motivational factors						
Participation in ICT[f] education, n (%)				N/A	41.98 (1)	<.001
Yes	16 (0.8)	12 (3.9)	4 (0.3)			
No	1903 (99.2)	294 (96.1)	1609 (99.7)			
Intention to learn, mean (SD)	2.01 (0.98)	3.49 (0.06)	4.09 (0.02)	9.96 (1)	N/A	<.001
Voluntary work, n (%)				N/A	84.36 (1)	<.001
Yes	290 (15.1)	99 (32.4)	191 (11.8)			
No	1629 (84.9)	207 (67.6)	1422 (88.2)			
Health-related factors						
Self-rated health, mean (SD)[g]	2.57 (0.92)	2.97 (0.05)	2.49 (0.02)	–8.50 (1)	N/A	<.001
K-IADL[h] dependency, mean (SD)	11.12 (2.42)	10.21 (0.06)	11.29 (0.06)	7.30 (1)	N/A	<.001
MMSE-DS[i], mean (SD)	24.94 (3.78)	27.47 (0.13)	24.46 (0.10)	–13.32 (1)	N/A	<.001

[a]2-tailed *t* tests.

[b]N/A: not applicable.

[c]Age range was 69 to 95 years.

[d]Education range was 0 to 20 years.

[e]A currency exchange rate of ₩1084.74=US $1 is applicable.

[f]ICT: information and communications technology.

[g]Self-rated health range was 1 to 5.

[h]K-IADL: Korean Instrumental Activities of Daily Living (range of 11-33**).**

[i]MMS-DS: Mini-Mental State Examination for Dementia Screening (range of 0-30).

Prior to multivariate logistic regression, multicollinearity was assessed and the variance inflation factor of all the individual variables did not exceed 10.0 (1.06-2.10). The logistic regression analysis (Table 2) revealed that internet use was independently associated with younger age (OR 0.89, 95% CI 0.87-0.92), higher educational level (OR 1.20, 95% CI 1.16-1.26), and higher leisure activity expenditure (OR 1.02, 95% CI 1.01-1.04). Internet users had more experience with ICT education and were more motivated to learn than nonusers. The ORs showed that the odds of participation in ICT education were about 10 times higher (OR 9.75, 95% CI 2.39-39.84) and the odds of voluntary work were over 2 times higher (OR 2.09, 95% CI 1.48-2.94) for internet users compared with nonusers. Users were also more likely to have better K-IADL scores (OR 0.78, 95% CI 0.66-0.92), higher MMSE-DS scores (OR 1.19, 95% CI 1.12-1.27), and better perceived health status (OR 1.27, 95% CI 1.08-1.50).

Table 2. Logistic regression model predicting internet use among older adults with diabetes mellitus (N=1919).

Characteristics and variables	OR[a] (95% CI)	*P* value
Opportunity factors		
Age (years)	0.89 (0.87-0.92)	<.001
Education (years)	1.20 (1.16-1.26)	<.001
Leisure expenditure (₩)	1.02 (1.01-1.04)	<.001
Motivational factors		
Participation in ICT[b] education (reference: none)	9.75 (2.39-39.84)	.002
Intention to learn	1.39 (1.20-1.60)	<.001
Voluntary work (reference: no)	2.09 (1.48-2.94)	<.001
Health-related factors		
Self-rated health	1.27 (1.08-1.50)	.004
K-IADL[c] dependency	0.78 (0.66-0.92)	.003
MMSE-DS[d] (score)	1.19 (1.12-1.27)	<.001

[a]OR: odds ratio.

[b]ICT: information and communications technology.

[c]K-IADL: Korean Instrumental Activities of Daily Living.

[d]MMSE-DS: Mini-Mental State Examination for Dementia Screening.

Discussion

Principal Findings

This study attempted to provide basic data on indicators of internet use among older adults with diabetes in South Korea by identifying relevant variables related to opportunity, motivation, and health. Age, years of education, economic standing, ICT education, volunteer experience, physical function, and cognitive function were identified as major predictors of ICT use among older adults with diabetes.

Only 15.95% (306/1919) of the participants used the internet to search for information in this study. In South Korea, 38.5% of people aged 60 to 89 years use ICT [27]. A study on US residents showed that 27% of urban residents used computers and 38% of patients receiving kidney transplants used the internet [13,28]. These results are in line with studies showing that older adults with chronic diseases use the internet less than younger populations [22]. Some studies have shown that individuals frequently use the internet to search for health information, even when patients had chronic diseases [28]. It is necessary to exercise caution in interpreting whether chronic diseases predict internet use. In this study, more than 80% (1613/1919, 84.05%) of the participants did not use the internet, indicating a need for social policies to bridge the digital divide and improve internet use among older adults with diabetes.

Internet use among older adults is closely related to age, sex, and years of education [29], and the same results were demonstrated for the older adults with diabetes in this study; age was a predictor of internet use in older adults with diabetes.

In Korea, internet access has grown over the past two decades (Multimedia Appendix 1). Over 90% of the population has internet access through national support and various policies [1]. In this study, according to the leisure activity expenditure, the economic predictor of internet use signifies that a digital divide still exists among older adults with diabetes. Therefore, it is important to approach the digital divide in older adults with diabetes from the perspective of accessibility.

Participation in ICT education can be a possible predictor of internet use among older adults with diabetes. This result was in line with previous research, which found that older adults who knew how to use computers before they were 65 years old were 9 times more likely to use the internet than those who did not [30]. Therefore, the capabilities of using the internet and the ICT skills of older adults with diabetes should be assessed by health care providers prior to digital interventions or individualized education programs.

The focus of research on the digital divide has recently shifted from accessibility to utilization and outcomes. Many studies have shown that personal preferences and motivations, in addition to opportunities and structural aspects, influence active internet use [20]. This study revealed that internet nonusers were more willing to receive information on service education. It could thus be inferred that internet nonuse correlates with fewer technology training opportunities and that more training is needed for frail older adults and their caregivers to effectively use the internet to engage in care [24]. Therefore, individualized education programs for older adults with diabetes should include disease-related and ICT education.

In this study, volunteer activities as a type of social participation or activity predicted internet use. The results are consistent with studies that show that internet or mobile phone use by older adults is strongly related to social activities, social support, and self-esteem [27]. Leisure activity expenditure is a good proxy for economic status [24] and was a good predictor of internet use among older adults with diabetes in this study. Oh [31]

encouraged leisure activities among older adults, such as shopping and watching entertainment shows and performances, cultural activities, videos, and movies, because these activities significantly influenced active internet use and search capabilities among older generations. It is necessary to encourage older adults with diabetes to engage in leisure and hobby activities because it may improve their digital health literacy.

In this study, physical and cognitive function were identified as predictors of internet use; internet use decreases when health and instrumental activities of daily living are degraded by physical function [19]. Instrumental activities of daily living require high levels of physical function in everyday behavior to live independently and indicate the possibility of returning to society [25]. The results showed that K-IADL score is a predictor of internet use. Having good physical functional status could encourage older adults with diabetes to participate in social activities, making them more likely to have a chance to use the internet in society [22]. Thus, functional limitations should be considered in strategies to reduce the digital divide among older adults with chronic diseases.

Cognitive function was one of the predictors of internet use among older adults with diabetes. With age, adults experience a decline in both cognitive and physical function and become restricted in activities such as delicate muscle movement, reading, and interpreting large quantities of information. Internet use requires extensive cognitive information processing and learning and can therefore burden older adults [32]. Thus, developing functions and programs that can be more easily accessed and handled by older adults with reduced cognitive function is essential in enhancing internet use and reducing the digital divide.

Implications

Although the digital divide can be defined based on various aspects, such as access, usability, and utilization, this study focused on predictors of internet use among older adults with diabetes. We expect that improved internet use will improve self-care among this population; however, there is still a gap in internet use due to economic, social, physical, and cognitive factors [6]. In the current information age, health care systems are increasingly embracing eHealth and digital services. South Korea has created a national patient portal to provide health information through electronic devices. Meanwhile, other countries have developed digital aids using health-related applications, virtual reality, and games [33]. The weaknesses and strengths among older adults with diabetes should be properly identified to assist in the creation of individualized mediation plans. This will prevent the digital divide from separating older adults with diabetes from digital health care trends.

Due to the limitations of secondary data analysis, this study did not reflect the characteristics of the participants' diabetes, so future research should include the relationship between diabetes characteristics and internet use. Another limitation of this study is that although it used nationally representative data, there may be errors in generalization due to the small number of participants; therefore, it is necessary to be cautious when interpreting the results.

Conclusions

Internet use has dramatically increased in South Korea during the past two decades but remains very low among older adults with diabetes. Our results suggest that years of education, leisure activity expenditure, participation in education, intention of education, voluntary work, self-rated health, and MMSE-DS scores were positively correlated predictors of internet use, while age and K-IADL dependency were negatively correlated predictors of internet use. While prior studies of the digital divide in health care have highlighted demographics and socioeconomic status, our study demonstrates the additional impact of motivational factors and health-related factors in older adults with diabetes. Health care providers need to formulate digital health interventions to prevent the most frail and vulnerable older adults from being left out of consideration in online patient portals and eHealth. Policies and health care providers should focus on digital competency training and volunteer activities among older adults with diabetes. For functionally limited older adults, user-friendly digital aids may improve internet use. For cognitively impaired older adults, caregivers or family members should be included in the intervention. Future studies should examine more strategies to reduce the digital divide among older adults with diabetes.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary table.
[DOCX File , 14 KB - medinform_v8i12e19061_app1.docx]

References

1. Choi D, Park H, Lim H. The report on digital divide. Report No. NIA VI-RBE-C-18031. Seoul, Korea: National Information Society Agency; 2018.
2. Jung K, Oh Y, Lee Y, Oh M, Kang E, Kim K, et al. 2017 National survey of older Koreans. Report No. 11-1352000-00. Korea Institute for Health and Social Affairs. Yeongi-gun, South Korea; 2017. URL: http://www.mohw.go.kr/react/jb/sjb030301vw.jsp?PAR_MENU_ID=03&MENU_ID=032901&page=1&CONT_SEQ=344953 [accessed 2020-12-18]

3. Kim M. The effects of smartphone use on life satisfaction, depression, social activity and social support of older adults. J Korea Acad Industr Coop Soc 2018;19(11):264-277. [doi: 10.5762/KAIS.2018.19.11.264]
4. McGaughey RE, Zeltmann SM, McMurtrey ME. Motivations and obstacles to smartphone use by the elderly: developing a research framework. IJEF 2013;7(3/4):177. [doi: 10.1504/ijef.2013.058601]
5. An S, Lee J. Older Adults' Health Promotion via Mobile Application: The effect of Self-efficacy and Social Stigma. Korean J Journalism Commun Stud 2019 Apr 30;63(2):113-142 [FREE Full text] [doi: 10.20879/kjjcs.2019.63.2.004]
6. Estacio EV, Whittle R, Protheroe J. The digital divide: Examining socio-demographic factors associated with health literacy, access and use of internet to seek health information. J Health Psychol 2019 Oct;24(12):1668-1675. [doi: 10.1177/1359105317695429] [Medline: 28810415]
7. Park JY, June KJ. Influencing Factors on Functional Health Literacy among the Rural Elderly. J Korean Acad Community Health Nurs 2011;22(1):75. [doi: 10.12799/jkachn.2011.22.1.75]
8. Chesser A, Keene Woods N, Smothers K, Rogers N. Health Literacy and Older Adults: A Systematic Review. Gerontol Geriatr Med 2016;2:1-13 [FREE Full text] [doi: 10.1177/2333721416630492] [Medline: 28138488]
9. Matsuoka RL, Marass M, Avdesh A, Helker CS, Maischein HM, Grosse AS, et al. Radial glia regulate vascular patterning around the developing spinal cord. E Life 2016 Nov 17;5:e20253 [FREE Full text] [doi: 10.7554/eLife.20253] [Medline: 27852438]
10. Peterson A, Soberón J, Pearson R, Anderson R, Martínez-Meyer E, Nakamura M, et al. Ecological Niches and Geographic Distributions. Vol 49. Princeton, NJ: Princeton University Press; 2011.
11. Ko M, Kang K. Influence of health literacy and health empowerment on health behavior practice in elderly outpatients with coronary artery disease. J Korean Clin Nurs Res 2018;24(3):293-302. [doi: 10.22650/JKCNR.2018.24.3.293]
12. Kim J, No Y, Choi D, Jung B, Kim J. Aging and digital divide: determinant of divide. Premium Report KISDI issue Report No. 07-10. Jincheon County, South Korea: Korea Information Society Development Institute; 2007.
13. Yu R, Ellison N, McCammon R, Langa K. Mapping the two levels of digital divide: Internet access and social network site adoption among older adults in the USA. Inf Commun Soc 2015 Nov 19;19(10):1445-1464 [FREE Full text] [doi: 10.1080/1369118x.2015.1109695]
14. Cresci MK, Yarandi HN, Morrell RW. The Digital Divide and urban older adults. Comput Inform Nurs 2010;28(2):88-94. [doi: 10.1097/NCN.0b013e3181cd8184] [Medline: 20182159]
15. Choi N, Dinitto DM. The digital divide among low-income homebound older adults: Internet use patterns, eHealth literacy, and attitudes toward computer/Internet use. J Med Internet Res 2013 May 02;15(5):e93 [FREE Full text] [doi: 10.2196/jmir.2645] [Medline: 23639979]
16. Anderson M, Perrin A. Tech adoption climbs among older adults. Pew Research Center. 2017 May 17. URL: https://www.pewresearch.org/internet/2017/05/17/tech-adoption-climbs-among-older-adults/ [accessed 2020-12-09]
17. Andrews JA, Brown LJ, Hawley MS, Astell AJ. Older Adults' Perspectives on Using Digital Technology to Maintain Good Mental Health: Interactive Group Study. J Med Internet Res 2019 Feb 13;21(2):e11694 [FREE Full text] [doi: 10.2196/11694] [Medline: 30758292]
18. Smith A. Older Adults and Technology Use. Pew Research Center. 2014 Apr 03. URL: https://www.pewresearch.org/internet/2014/04/03/older-adults-and-technology-use/ [accessed 2020-12-09]
19. Greysen SR, Chin GC, Sudore RL, Cenzer IS, Covinsky KE. Functional impairment and Internet use among older adults: implications for meaningful use of patient portals. JAMA Intern Med 2014 Jul;174(7):1188-1190 [FREE Full text] [doi: 10.1001/jamainternmed.2014.1864] [Medline: 24839165]
20. Connolly KK, Crosby ME. Examining e-Health literacy and the digital divide in an underserved population in Hawai'i. Hawaii J Med Public Health 2014 Feb;73(2):44-48 [FREE Full text] [Medline: 24567867]
21. Scheerder A, van Deursen A, van Dijk J. Determinants of Internet skills, uses and outcomes. A systematic review of the second- and third-level digital divide. Telematics Inform 2017 Dec;34(8):1607-1624 [FREE Full text] [doi: 10.1016/j.tele.2017.07.007]
22. Kim M, Park CS, Kwon SJ. Babyboomer's Use of Information Technology and Its Effect on the Digital Life Satisfaction: The Mediating Effect of the Self-mastery. Korean J Gerontol Soc Welfare 2012 Sep;57:113-137. [doi: 10.21194/kjgsw..57.201209.113]
23. Tennant B, Stellefson M, Dodd V, Chaney B, Chaney D, Paige S, et al. eHealth literacy and Web 2.0 health information seeking behaviors among baby boomers and older adults. J Med Internet Res 2015 Mar 17;17(3):e70 [FREE Full text] [doi: 10.2196/jmir.3992] [Medline: 25783036]
24. Dardis R, Soberon-Ferrer H, Patro D. Analysis of Leisure Expenditures in the United States. J Leisure Res 2018 Feb 13;26(4):309-321. [doi: 10.1080/00222216.1994.11969964]
25. Won C, Yang K, Rho Y, Kim S, Lee E, Yoon J, et al. The development of Korean activities of daily living and Korean instrumental activities of daily living scale. Korean Geriatrics 2002;6(2):107-120. [doi: 10.1037/t06803-000]
26. Kim TH, Jhoo JH, Park JH, Kim JL, Ryu SH, Moon SW, et al. Korean version of mini mental status examination for dementia screening and its short form. Psychiatry Investig 2010 Jun;7(2):102-108 [FREE Full text] [doi: 10.4306/pi.2010.7.2.102] [Medline: 20577618]

27. Nam-Gung HK, Kim IH, Chun H. Study on the Correlates of Digital Disparity among Older Seoul Residents. J Digital Convergence 2017 Apr 28;15(4):73-81. [doi: 10.14400/jdc.2017.15.4.73]
28. Lockwood M, Saunders M, Josephson M, Becker Y, Lee C. Determinants of frequent Internet use in an urban kidney transplant population in the United States: characterizing the digital divide. Prog Transplant 2015 Mar;25(1):9-17. [doi: 10.7182/pit2015957] [Medline: 25758795]
29. Kim MY, Jun HY. The Influences of IT Use and Satisfaction with IT Use on Depression among Older Adults. Korean J Gerontol Soc Welfare 2016 Mar;71(1):85-110. [doi: 10.21194/kjgsw.71.1.201603.85]
30. Friemel T. The digital divide has grown old: Determinants of a digital divide among seniors. New Media Soc 2014 Jun 12;18(2):313-331 [FREE Full text] [doi: 10.1177/1461444814538648]
31. Oh J. The effect of baby boom generation' leisure activities on ICT skills. J Digit Convergence 2018;16(3):1-12. [doi: 10.14400/JDC.2018.16.3.001]
32. An J, Park K. Smartphone Utilization and Satisfaction in Community Dwelling Elderly. J Korean Soc Living Environ Syst 2019 Aug 31;26(4):540-549. [doi: 10.21086/ksles.2019.08.26.4.540]
33. Perski O, Jackson SE, Garnett C, West R, Brown J. Trends in and factors associated with the adoption of digital aids for smoking cessation and alcohol reduction: A population survey in England. Drug Alcohol Depend 2019 Dec 01;205:107653 [FREE Full text] [doi: 10.1016/j.drugalcdep.2019.107653] [Medline: 31675544]

Abbreviations

ICT: information and communications technology
K-IADL: Korean Instrumental Activities of Daily Living
MMSE-DS: Mini-Mental State Examination for Dementia Screening
OR: odds ratio

Edited by C Lovis; submitted 02.04.20; peer-reviewed by K Kim, D Reinwand, M El Tantawi; comments to author 21.04.20; revised version received 06.05.20; accepted 24.11.20; published 23.12.20.

Please cite as:
Park S, Kim B
Predictors of Internet Use Among Older Adults With Diabetes in South Korea: Survey Study
JMIR Med Inform 2020;8(12):e19061
URL: http://medinform.jmir.org/2020/12/e19061/
doi:10.2196/19061
PMID:33277232

Original Paper

Automatically Explaining Machine Learning Prediction Results on Asthma Hospital Visits in Patients With Asthma: Secondary Analysis

Gang Luo[1], DPhil; Michael D Johnson[2], MD; Flory L Nkoy[2], MPH, MSc, MD; Shan He[3], DPhil; Bryan L Stone[2], MSc, MD

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

[2]Department of Pediatrics, University of Utah, Salt Lake City, UT, United States

[3]Care Transformation and Information Systems, Intermountain Healthcare, Salt Lake City, UT, United States

Corresponding Author:
Gang Luo, DPhil
Department of Biomedical Informatics and Medical Education
University of Washington
Building C, Box 358047
850 Republican Street
Seattle, WA, 98195
United States
Phone: 1 2062214596
Fax: 1 2062212671
Email: gangluo@cs.wisc.edu

Abstract

Background: Asthma is a major chronic disease that poses a heavy burden on health care. To facilitate the allocation of care management resources aimed at improving outcomes for high-risk patients with asthma, we recently built a machine learning model to predict asthma hospital visits in the subsequent year in patients with asthma. Our model is more accurate than previous models. However, like most machine learning models, it offers no explanation of its prediction results. This creates a barrier for use in care management, where interpretability is desired.

Objective: This study aims to develop a method to automatically explain the prediction results of the model and recommend tailored interventions without lowering the performance measures of the model.

Methods: Our data were imbalanced, with only a small portion of data instances linking to future asthma hospital visits. To handle imbalanced data, we extended our previous method of automatically offering rule-formed explanations for the prediction results of any machine learning model on tabular data without lowering the model's performance measures. In a secondary analysis of the 334,564 data instances from Intermountain Healthcare between 2005 and 2018 used to form our model, we employed the extended method to automatically explain the prediction results of our model and recommend tailored interventions. The patient cohort consisted of all patients with asthma who received care at Intermountain Healthcare between 2005 and 2018, and resided in Utah or Idaho as recorded at the visit.

Results: Our method explained the prediction results for 89.7% (391/436) of the patients with asthma who, per our model's correct prediction, were likely to incur asthma hospital visits in the subsequent year.

Conclusions: This study is the first to demonstrate the feasibility of automatically offering rule-formed explanations for the prediction results of any machine learning model on imbalanced tabular data without lowering the performance measures of the model. After further improvement, our asthma outcome prediction model coupled with the automatic explanation function could be used by clinicians to guide the allocation of limited asthma care management resources and the identification of appropriate interventions.

(JMIR Med Inform 2020;8(12):e21965) doi:10.2196/21965

KEYWORDS

asthma; forecasting; machine learning; patient care management

Introduction

Background

About 8.4% of Americans have asthma [1]. Each year in the United States, asthma costs over US $50 billion and results in more than 2 million emergency department (ED) visits, about half a million inpatient stays, and more than 3000 deaths [1,2]. A major goal in managing patients with asthma is to reduce their hospital visits, including ED visits and inpatient stays. As employed by health plans in 9 of 12 metropolitan communities [3] and by health care systems such as Intermountain Healthcare, Kaiser Permanente Northern California [4], and the University of Washington Medicine, the state-of-the-art method for achieving this goal is to employ a predictive model to predict which patients with asthma are highly likely to have poor outcomes in the future. Once identified, such patients are enrolled in care management. Care managers then call these patients on the phone regularly and help them make appointments for health and related services. By offering such tailored preventive care properly, up to 40% of future hospital visits by patients with asthma can be avoided [5-8].

A care management program has limited enrollment capacity [9]. As a result, the effectiveness of the program depends critically on the accuracy of the predictive model. Not enrolling a patient who is likely to have future hospital visits in the program is a missed opportunity to improve the patient's outcomes. Unnecessarily enrolling a patient who is likely to have no future hospital visit would increase health care costs and waste scarce care management resources with no potential benefit. The current models for predicting hospital visits in patients with asthma are inaccurate, with published sensitivity of ≤49% and an area under the receiver operating characteristic curve (AUC) ≤0.81 [4,10-22]. When employed for care management, these models miss more than half of the patients who will have future hospital visits and erroneously label many other patients as likely to have future hospital visits [23]. To address these issues, we recently built an extreme gradient boosting (XGBoost) [24] machine learning model to predict asthma hospital visits in the subsequent year in patients with asthma [23]. Compared with previous models, our model raised the AUC by at least 0.049. However, like most machine learning models, our model offers no explanation of its prediction results. This creates a barrier for use in care management, where care managers need to understand why a patient is at risk for poor outcomes to make care management enrollment decisions and identify suitable interventions for the patient.

Objectives

To overcome the abovementioned barrier, this study aims to develop a method to automatically explain the prediction results of our model and recommend tailored interventions without lowering any of the performance measures of our model, such as AUC, accuracy, sensitivity, specificity, positive predictive value, and negative predictive value.

In the following sections, we describe our methods and the evaluation results. A list of abbreviations adopted in this paper is provided at the end of the paper.

Methods

We used the same patient cohort, data set, prediction target, cutoff threshold for binary classification, method for data preprocessing, including data cleaning and data normalization, and method for partitioning the whole data set into the training and test sets that we described in our prior paper [23].

Ethics Approval and Study Design

This study consists of a secondary analysis of retrospective data and was evaluated and approved by the institutional review boards of the University of Washington Medicine, University of Utah, and Intermountain Healthcare.

Patient Population

Our patient cohort included all patients with asthma who received care at any Intermountain Healthcare facility between 2005 and 2018 and resided in Utah or Idaho as recorded at the visit. Intermountain Healthcare is the largest health care system in Utah and southeastern Idaho. It operates 185 clinics and 22 hospitals and provides care for approximately 60% of people living in that region. A patient was considered asthmatic in a specific year if in the encounter billing database, the patient had one or more asthma diagnosis codes during that year (International Classification of Diseases, ninth revision [ICD-9]: 493.0x, 493.1x, 493.8x, 493.9x; International Classification of Diseases, tenth revision [ICD-10]: J45.x) [12,25,26]. The only exclusion criterion from the analysis in any given year was patient death during that year.

Data Set

We used a structured clinical and administrative data set provided by the enterprise data warehouse of Intermountain Healthcare. The data set covered all visits by the patient cohort within Intermountain Healthcare between 2005 and 2018.

Prediction Target (the Dependent or Outcome Variable)

For each patient identified as asthmatic in a specific year, the outcome was whether any asthma hospital visit occurred in the subsequent year. In this paper, an asthma hospital visit refers to an ED visit or an inpatient stay at an Intermountain Healthcare facility with a principal diagnosis of asthma (ICD-9: 493.0x, 493.1x, 493.8x, 493.9x; ICD-10: J45.x). For training and testing the XGBoost model and automatic explanation method, data of every patient with asthma up to the end of every year were used to predict the patient's outcome in the subsequent year.

Predictive Model and Features (Independent Variables)

Our recent XGBoost model [23] uses 142 features to predict asthma hospital visits in the subsequent year in patients with asthma. As listed in the multimedia appendix in our previous study [23], these features were computed from the structured attributes in our data set covering a wide range of categories, such as patient demographics, visits, medications, laboratory tests, vital signs, diagnoses, and procedures. Each input data instance for our model has these 142 features, targets a pair of a patient with asthma and a year, and is employed to predict the

patient's outcome in the subsequent year. We set the cutoff threshold for binary classification at the top 10% of patients with asthma having the largest predicted risk. These patients were predicted to incur asthma hospital visits in the subsequent year.

Automatic Explanation Method

Previously, we developed an automated method to offer rule-formed explanations for any machine learning model's prediction results on tabular data and recommend tailored interventions without lowering the performance measures of the model [27,28]. Our method was initially demonstrated to predict the diagnosis of type 2 diabetes [27]. Later, other researchers successfully applied our method to predict death or lung transplantation in patients with cystic fibrosis [29], predict cardiac death in patients with cancer, and use predictions to manage preventive care, heart transplant waiting list, and posttransplant follow-ups in patients with cardiovascular diseases [30]. In our method, each rule used for providing explanations has a performance measure termed confidence that must be greater than or equal to a given minimum confidence threshold c_{min}. Our original automatic explanation method [27] was designed for reasonably balanced data, where distinct values of the outcome variable appear with relatively similar frequencies. Recently, we outlined an extension of this method [31,32] to handle imbalanced data, where one value of the outcome variable appears much less often than another. This data imbalance exists when predicting asthma hospital visits in patients with asthma, where only about 4% of the data instances are linked to future asthma hospital visits [23]. In our extended method, each rule used for providing explanations has a second performance measure termed commonality, which must be greater than or equal to a given minimum commonality threshold m_{min}. To date, no technique has been developed to efficiently mine the rules with commonality greater than or equal to m_{min}, compute their confidence, and eliminate those rules with confidence less than c_{min} in the extended method, despite such techniques being essential for handling large data sets. No guideline exists for setting the values of the parameters used in the extended method, although they greatly impact the performance of the extended method. The extended method has never been implemented in computer code. Moreover, the effectiveness of the extended method has not been evaluated or demonstrated.

In this study, we made the following innovative contributions:

1. We provide several techniques for efficiently mining the rules with commonality greater than or equal to m_{min}, computing their confidence, and eliminating those rules with confidence less than c_{min} in the extended automatic explanation method. This completes our extended method. Although our extended method was designed for imbalanced data, it can also be used on reasonably balanced data to improve the efficiency of mining the rules needed to provide automatic explanations. Among the existing automatic explanation methods for machine learning prediction results, our method is the only one that can automatically recommend tailored interventions [33,34]. This capability is desired for many medical applications.
2. We present a guideline to set the values of the parameters used in the extended method (see the Discussion section).
3. We completed the first computer coding implementation of the extended method and explained it in this paper.
4. We demonstrate the effectiveness of the extended method in predicting asthma hospital visits in patients with asthma.

Review of Our Original Automatic Explanation Method

Main Idea

Our automatic explanation method separates explanation and prediction by employing 2 models concurrently, each for a distinct purpose. The first model is used to make predictions and can be any model that takes continuous and categorical features as its inputs. Usually, we adopt the most accurate model as the first model to avoid lowering the performance measures of the model. The second model uses class-based association rules [35,36] mined from historical data to explain the prediction results of the first model rather than to make predictions. Before using a standard association rule mining method like Apriori to mine the rules [36], each continuous feature is first transformed into a categorical feature through automatic discretization [35,37]. Each rule shows a feature pattern associated with a value w of the outcome variable in the form of q_1 AND q_2 AND … AND $q_n \rightarrow w$. The values of n and w can change across rules. For binary classification distinguishing poor versus good outcomes, w is usually the poor outcome value. Every item q_i ($1 \leq i \leq n$) is a feature-value pair (f, u) showing feature f has value u or a value within u, depending on whether u is a value or a range. The rule points out that a patient's outcome variable is inclined to have value w if the patient fulfills q_1, q_2, ..., and q_n. An example rule is as follows:

- The patient had ≥12 ED visits in the past year

AND the patient had ≥21 distinct medications in all asthma medication orders in the past year

→the patient will incur one or more asthma hospital visits in the subsequent year.

The Association Rule Mining and Pruning Processes

The association rule mining process is controlled by 2 parameters: the minimum support threshold s_{min} and the minimum confidence threshold c_{min} [36]. For any rule l: q_1 AND q_2 AND … AND $q_n \rightarrow w$, the percentage of data instances satisfying q_1, q_2, ..., and q_n and linking to w is termed l's support showing l's coverage. Among all data instances satisfying q_1, q_2, ..., and q_n, the percentage of data instances linking to w is termed l's confidence reflecting l's precision. Our original automatic explanation method uses rules with support $\geq s_{min}$ and confidence $\geq c_{min}$. For binary classification distinguishing poor versus good outcomes, we usually focus on the rules that have right-hand sides containing the poor outcome value.

Usually, numerous association rules have support and confidence $\geq s_{min}$ and $\geq c_{min}$, respectively. To avoid overwhelming the users of the automatic explanation function with too many

rules, we used 4 techniques to reduce the number of rules in the second model. First, only features adopted by the first model are used to form rules. Second, a clinician in the automatic explanation function's design team checks all possible values and value ranges of these features and marks those that could possibly have a positive correlation with the values of the outcome variable reflecting poor outcomes. Only those marked values and value ranges of these features are allowed to show up in the rules. Third, the rules are limited to having no more than a given small number of items on their left-hand sides, as long rules are hard to understand. A typical value of this number is 4. Fourth, each more specific rule is dropped when there exists a more general rule with confidence that is not lower by more than a given threshold $\tau \geq 0$. More specifically, consider 2 rules, l_1 and l_2, whose right-hand sides have the same value. The items on the left-hand side of l_2 are a superset of those on the left-hand side of l_1. We drop l_2 if l_1's confidence is $\geq l_2$'s confidence-τ.

For the association rules remaining after the rule-pruning process, a clinician in the automatic explanation function's design team gathers zero or more interventions targeting the reason the rule presents. A rule is called actionable if one or more interventions are compiled for it. Usually, each intervention links to one of the feature-value pair items on the rule's left-hand side. Such an item is called actionable. Thus, an actionable rule contains at least 1 actionable item. To expedite the intervention compilation process, the clinician can identify all of the actionable items and compile interventions for each of them. All of the interventions linking to the actionable items on a rule's left-hand side are automatically connected to the rule.

Our automatic explanation method uses 2 types of knowledge manually compiled by a clinician: the values and value ranges of the features that could possibly have a positive correlation with the outcome variable's values reflecting poor outcomes and the interventions for the actionable items. Our automatic explanation method is fully automatic, except for the knowledge compilation step.

The Explanation Method

For each patient for whom the first model predicts a poor outcome, we explain the prediction result by listing the association rules in the second model whose right-hand sides have the corresponding poor outcome value and whose left-hand sides are fulfilled by the patient, whereas ignoring the rules in the second model whose right-hand sides have a value that differs from the corresponding poor outcome value and whose left-hand sides are fulfilled by the patient. Every rule listed offers a reason why the patient is predicted to have a poor outcome. For each actionable rule listed, the linked interventions are displayed next to it. This helps the user of the automatic explanation function find tailored inventions suitable for the patient. Typically, the rules in the second model describe common reasons for poor outcomes. However, some patients will have poor outcomes for rare reasons not covered by these rules. Consequently, the second model can provide explanations for most, but not all, of the patients for whom the first model predicts poor outcomes.

The Previously Outlined Extension of the Original Automatic Explanation Method

Our original automatic explanation method was designed for reasonably balanced data and is unsuitable for imbalanced data, where one value of the outcome variable appears much less often than another. If the minimum support threshold s_{min} is large on imbalanced data, we cannot obtain enough association rules for the outcome variable's rare values. Consequently, for a large portion of the first model's prediction results on these values, we cannot give any explanation. Conversely, if s_{min} is too small, the rule mining process will generate too many rules as intermediate results, most of which will be filtered out in the end. This easily exhausts computer memory and makes the rule mining process extremely slow. In addition, many overfitted rules will be produced in the end, making it difficult for clinicians to examine the mined rules.

In our recently outlined extension of the original automatic explanation method [31,32] to handle imbalanced data, we replace support with value-specific support termed commonality [38]. For any rule l: q_1 AND q_2 AND ... AND $q_n \rightarrow w$, among all data instances linking to w, the percentage of data instances satisfying q_1, q_2, ..., and q_n is termed l's commonality showing l's coverage within the context of w. Moreover, we replace the minimum support threshold s_{min} with the minimum commonality threshold m_{min}. Instead of using rules whose support is $\geq s_{min}$ and whose confidence is $\geq$ the minimum confidence threshold c_{min}, we used rules whose commonality is $\geq m_{min}$ and whose confidence is $\geq c_{min}$.

Each value of the outcome variable falls into one of 2 possible cases. In the first case, the value is interesting and represents an abnormal case. The prediction results of this value require attention and explanations. In the second case, the value is uninteresting and represents a normal case. The prediction results of this value require neither special attention nor explanation. Typically, each interesting value is a rare one reflecting poor outcomes. The second model contains only the association rules related to interesting values. To mine these rules, we proceeded in 2 steps:

- Step 1: For each interesting value w, we applied a standard association rule mining method like Apriori [36] to the set S_w of data instances linking to w to mine the rules related to w and with support on $S_w \geq$ the minimum commonality threshold m_{min}. These rules have commonality $\geq m_{min}$ on the set S_{all} of all data instances. As S_w is much smaller than S_{all}, mining these rules from S_w is much more efficient than first applying the association rule mining method to S_{all} to obtain the rules with support on $S_{all} \geq m_{min} \times |S_w|/|S_{all}|$, and then filtering out those rules unrelated to w. Here, $|S|$ denotes the cardinality of set S.
- Step 2: For each rule mined from S_w, we compute its confidence on S_{all}. We keep it only if its confidence on S_{all} is $\geq$ the minimum confidence threshold c_{min}.

Techniques for Efficiently Mining the Association Rules Whose Commonality is $\geq m_{min}$, Computing Their

Confidence, and Eliminating Those Rules Whose Confidence is <c_{min} in the Extended Automatic Explanation Method

When the set S_{all} of all data instances includes many data instances and features, we often find that the set S_w of data instances linking to an interesting value w contains many data instances, and the first model adopts many features. Without limiting the number of data instances in S_w and the number of features, numerous (eg, several billion) association rules would be mined from S_w in Step 1. This makes the computer easily run out of memory and the rule mining process extremely slow. In addition, many rules will be produced at the end, making it difficult for clinicians to examine them. To address this issue, we can use one or more of the following approaches:

1. We take a random sample of data instances S_{sample} from S_{all} and use S_{sample} rather than S_{all} to mine the rules [39].
2. Before the rule mining process starts, each data instance is transformed into a transaction. To reduce its size, we remove from the transaction those values and value ranges that the clinician in the automatic explanation function's design team marks as not allowed to show up in any of the rules.
3. Instead of using all of the features adopted by the first model, we use only the top features to mine the rules. Usually, the top features contain most of the predictive power possessed by all features adopted by the first model [23]. If the machine learning algorithm used to build the first model is like XGBoost [24] or random forest, which automatically computes each feature's importance value, the top features are those with the highest importance values. Otherwise, if the machine learning algorithm used to build the first model does not automatically compute each feature's importance value, we can use an automatic feature selection method [40] such as the information gain method to choose the top features. Alternatively, we can use XGBoost or random forest to construct a model, automatically compute each feature's importance value, and choose the top features with the highest importance values.

In the following, we focus on the case of using the set S_{all} of all data instances to mine the association rules. The case of using a random sample of data instances S_{sample} from S_{all} to mine the rules can be handled in a similar way. To compute the rules' confidence values, we transformed S_{all} to the matrix format, with each row of the matrix linking to a distinct data instance and each column of the matrix linking to a distinct value or value range of a feature. For medical data, the matrix is often not very sparse. In this case, we can use a separate bitmap to represent each column of the matrix in a condensed manner. For each rule l: q_1 AND q_2 AND ... AND $q_n \rightarrow w$, we performed efficient bitmap operations to pinpoint the data instances satisfying q_1, q_2, ..., and q_n and needed for computing l's confidence.

Among all the mined association rules related to an interesting value w, we needed to identify those whose confidence on the set S_{all} of all data instances is ≥ the minimum confidence threshold c_{min}. To expedite the identification process, we proceeded as follows: for each rule l: q_1 AND q_2 AND ... AND $q_n \rightarrow w$, let l_w denote the number of data instances satisfying q_1, q_2, ..., and q_n and linking to w, and $l_{\neg w}$ denote the number of data instances satisfying q_1, q_2, ..., and q_n and not linking to w. Our key insight was that l's confidence on S_{all} $l_w/(l_w+l_{\neg w})$ is $<c_{min}$ if and only if $l_{\neg w}$ is $>T_l$ $l_w \times (1-c_{min})/c_{min}$. We partitioned S_{all} into 2 subsets: S_w containing all of the data instances linking to w and $S_{\neg w}$ containing all of the data instances not linking to w. Using the bitmap method mentioned above, we went over all of the data instances in S_w to compute l_w. Then, we went over the data instances in $S_{\neg w}$ one by one to count the data instances satisfying q_1, q_2, ..., and q_n and not linking to w. Once this count is $>T_l$, we know l's confidence on S_{all} is $<c_{min}$, stop the counting process, and drop l. This saves the overhead of going through the remaining data instances in $S_{\neg w}$ to compute $l_{\neg w}$. Otherwise, if this count is $\leq T_l$ when we reach the last data instance in $S_{\neg w}$, we keep l, obtain $l_{\neg w}$, and compute l's confidence on S_{all}, which must be $\geq c_{min}$.

Computer Coding Implementation

We implemented our extended automatic explanation method in computer code, using a hybrid of the C and R programming languages. As R is an interpreted language and inefficient at handling certain operations on large data sets, we wrote several parts of our code in C to improve our code's execution speed. Considering that our asthma outcome variable is hard to predict, we limited the association rules to have at most 5 items on their left-hand sides (see the guideline in the *Discussion* section). We set the minimum confidence threshold c_{min} to 50% and the minimum commonality threshold m_{min} to 0.2%.

Data Analysis

The Training and Test Set Partitioning

As outcomes came from the subsequent year, our data set included 13 years of effective data (2005-2017) during the 14 years between 2005 and 2018. To mirror the practical use of our XGBoost model and our extended automatic explanation method, the 2005 to 2016 data were used as the training set to train our XGBoost model and mine the association rules used by our extended method. The 2017 data were used as the test set to evaluate the performance of our XGBoost model and extended method. We used the full set of 142 features to make predictions and the top 50 features that our XGBoost model [23] ranked with the highest importance values to mine the association rules. Our XGBoost model reached an AUC of 0.859 using the full set of 142 features [23] and an AUC of 0.857 using the top 50 features.

Presenting 5 Example Association Rules Used in the Second Model

To give the reader a concrete feeling of the association rules used in the second model, we randomly chose 5 example rules to present in this paper.

Performance Metrics

We evaluated the performance of our extended automatic explanation method in several ways. The main performance metric that we used to show our extended method's explanation capability was the percentage of patients for whom our extended method could provide explanations among the patients with asthma whom our XGBoost model correctly predicted to incur asthma hospital visits in the subsequent year. We reported both the average number of rules and the average number of actionable rules fitting such a patient. A rule fits a patient if the patient fulfills all of the items on its left-hand side.

As shown in our previous study [27], multiple rules fitting a patient frequently differ from each other by a single feature-value pair item on their left-hand sides. When many rules fit a patient, the amount of nonredundant information embedded in them is often much less than the number of these rules. To give a full picture of the information richness of the automatic explanations provided for the patients, we present 3 distributions of the patients with asthma whom our XGBoost model correctly predicted to incur asthma hospital visits in the subsequent year: (1) by the number of rules fitting a patient, (2) by the number of actionable rules fitting a patient, and (3) by the number of distinct actionable items appearing in all of the rules fitting a patient.

Results

Our Patient Cohort's Demographic and Clinical Characteristics

Every data instance targets a distinct pair of a patient with asthma and a year. Table 1 lists the demographic and clinical characteristics of our patient cohort between 2005 and 2016, which included 182,245 patients. Table 2 lists the demographic and clinical characteristics of our patient cohort in 2017, which included 19,256 patients. These 2 sets of characteristics are reasonably similar. Between 2005 and 2016, 3.59% (11,332/315,308) of data instances were related to asthma hospital visits in the subsequent year. In 2017, this percentage was 4.22% (812/19,256).

Table 1. Demographic and clinical characteristics of the Intermountain Healthcare patients with asthma between 2005 and 2016.

Characteristics	Data instances related to no asthma hospital visit in the subsequent year (n=303,976), n (%)	Data instances related to asthma hospital visits in the subsequent year (n=11,332), n (%)	Data instances (n=315,308), n (%)
Gender			
Female	181,928 (59.85)	6163 (54.39)	188,091 (59.65)
Male	122,048 (40.15)	5169 (45.61)	127,217 (40.35)
Age (years)			
≥65	46,260 (15.22)	621 (5.48)	46,881 (14.87)
18 to 65	172,436 (56.73)	5003 (44.15)	177,439 (56.27)
6 to <18	50,572 (16.64)	2590 (22.86)	53,162 (16.86)
<6	34,708 (11.42)	3118 (27.52)	37,826 (12.00)
Ethnicity			
Non-Hispanic	244,442 (80.41)	8157 (71.98)	252,599 (80.11)
Hispanic	27,014 (8.89)	2279 (20.11)	29,293 (9.29)
Unknown or not reported	32,520 (10.70)	896 (7.91)	33,416 (10.60)
Race			
White	273,206 (89.88)	9420 (83.13)	282,626 (89.63)
Native Hawaiian or other Pacific Islander	3877 (1.28)	411 (3.63)	4288 (1.36)
Black or African American	5291 (1.74)	460 (4.06)	5751 (1.82)
Asian	2120 (0.70)	77 (0.68)	2197 (0.70)
American Indian or Alaska Native	2295 (0.76)	214 (1.89)	2509 (0.80)
Unknown or not reported	17,187 (5.65)	750 (6.62)	17,937 (5.69)
Duration of asthma (years)			
>3	76,810 (25.27)	3666 (32.35)	80,476 (25.52)
≤3	227,166 (74.73)	7666 (67.65)	234,832 (74.48)
Insurance			
Self-paid or charity	26,611 (8.75)	1902 (16.78)	28,513 (9.04)
Public	76,916 (25.30)	3238 (28.57)	80,154 (25.42)
Private	200,449 (65.94)	6192 (54.64)	206,641 (65.54)
Smoking status			
Never smoker or unknown	251,501 (82.74)	8952 (79.00)	260,453 (82.60)
Former smoker	18,735 (6.16)	569 (5.02)	19,304 (6.12)
Current smoker	33,740 (11.10)	1811 (15.98)	35,551 (11.28)
Comorbidity			
Sleep apnea	20,421 (6.72)	471 (4.16)	20,892 (6.63)
Sinusitis	14,164 (4.66)	592 (5.22)	14,756 (4.68)
Premature birth	5102 (1.68)	440 (3.88)	5542 (1.76)
Obesity	35,215 (11.58)	1076 (9.50)	36,291 (11.51)
Gastroesophageal reflux	54,887 (18.06)	1309 (11.55)	56,196 (17.82)
Eczema	4484 (1.48)	443 (3.91)	4927 (1.56)
Cystic fibrosis	447 (0.15)	11 (0.10)	458 (0.15)
Chronic obstructive pulmonary disease	12,496 (4.11)	391 (3.45)	12,887 (4.09)

Characteristics	Data instances related to no asthma hospital visit in the subsequent year (n=303,976), n (%)	Data instances related to asthma hospital visits in the subsequent year (n=11,332), n (%)	Data instances (n=315,308), n (%)
Bronchopulmonary dysplasia	394 (0.13)	35 (0.31)	429 (0.14)
Anxiety or depression	55,245 (18.17)	1716 (15.14)	56,961 (18.07)
Allergic rhinitis	4534 (1.49)	181 (1.60)	4715 (1.50)
Asthma medication prescription			
Systemic corticosteroid	129,318 (42.54)	7324 (64.63)	136,642 (43.34)
Short-acting, inhaled beta-2 agonist	121,983 (40.13)	7545 (66.58)	129,528 (41.08)
Mast cell stabilizer	114 (0.04)	7 (0.06)	121 (0.04)
Long-acting beta-2 agonist	1744 (0.57)	69 (0.61)	1813 (0.58)
Leukotriene modifier	33,187 (10.92)	2320 (20.47)	35,507 (11.26)
Inhaled corticosteroid/long-acting beta-2 agonist combination	42,796 (14.08)	2196 (19.38)	44,992 (14.27)
Inhaled corticosteroid	73,566 (24.20)	4539 (40.05)	78,105 (24.77)

Table 2. Demographic and clinical characteristics of the Intermountain Healthcare patients with asthma in 2017.

Characteristics	Data instances related to no asthma hospital visit in the subsequent year (n=18,444), n (%)	Data instances related to asthma hospital visits in the subsequent year (n=812), n (%)	Data instances (n=19,256), n (%)
Gender			
Female	11,001 (59.65)	439 (54.06)	11,440 (59.41)
Male	7443 (40.35)	373 (45.94)	7816 (40.59)
Age (years)			
≥65	3833 (20.78)	46 (5.67)	3879 (20.14)
18 to 65	9879 (53.56)	386 (47.54)	10,265 (53.31)
6 to <18	3054 (16.56)	181 (22.29)	3235 (16.80)
<6	1678 (9.10)	199 (24.51)	1877 (9.75)
Ethnicity			
Non-Hispanic	16,242 (88.06)	618 (76.11)	16,860 (87.56)
Hispanic	2020 (10.95)	192 (23.65)	2212 (11.49)
Unknown or not reported	182 (0.99)	2 (0.25)	184 (0.96)
Race			
White	17,025 (92.31)	681 (83.87)	17,706 (91.95)
Native Hawaiian or other Pacific Islander	299 (1.62)	47 (5.79)	346 (1.80)
Black or African American	361 (1.96)	42 (5.17)	403 (2.09)
Asian	195 (1.06)	10 (1.23)	205 (1.06)
American Indian or Alaska Native	146 (0.79)	13 (1.60)	159 (0.83)
Unknown or not reported	418 (2.27)	19 (2.34)	437 (2.27)
Duration of asthma (years)			
>3	7734 (41.93)	389 (47.91)	8123 (42.18)
≤3	10,710 (58.07)	423 (52.09)	11,133 (57.82)
Insurance			
Self-paid or charity	1136 (6.16)	142 (17.49)	1278 (6.64)
Public	4920 (26.68)	208 (25.62)	5128 (26.63)
Private	12,388 (67.17)	462 (56.90)	12,850 (66.73)
Smoking status			
Never smoker or unknown	13,956 (75.67)	583 (71.80)	14,539 (75.50)
Former smoker	2243 (12.16)	83 (10.22)	2326 (12.08)
Current smoker	2245 (12.17)	146 (17.98)	2391 (12.42)
Comorbidity			
Sleep apnea	2925 (15.86)	78 (9.61)	3003 (15.60)
Sinusitis	746 (4.04)	34 (4.19)	780 (4.05)
Premature birth	435 (2.36)	41 (5.05)	476 (2.47)
Obesity	3389 (18.37)	116 (14.29)	3505 (18.20)
Gastroesophageal reflux	3477 (18.85)	71 (8.74)	3548 (18.43)
Eczema	273 (1.48)	34 (4.19)	307 (1.59)
Cystic fibrosis	94 (0.51)	1 (0.12)	95 (0.49)
Chronic obstructive pulmonary disease	1033 (5.60)	23 (2.83)	1056 (5.48)

Characteristics	Data instances related to no asthma hospital visit in the subsequent year (n=18,444), n (%)	Data instances related to asthma hospital visits in the subsequent year (n=812), n (%)	Data instances (n=19,256), n (%)
Bronchopulmonary dysplasia	12 (0.07)	3 (0.37)	15 (0.08)
Anxiety or depression	3815 (20.68)	131 (16.13)	3946 (20.49)
Allergic rhinitis	382 (2.07)	10 (1.23)	392 (2.04)
Asthma medication prescription			
Systemic corticosteroid	11,327 (61.41)	693 (85.34)	12,020 (62.42)
Short-acting, inhaled beta-2 agonist	13,046 (70.73)	739 (91.01)	13,785 (71.59)
Mast cell stabilizer	8 (0.04)	0 (0.00)	8 (0.04)
Long-acting beta-2 agonist	47 (0.25)	5 (0.62)	52 (0.27)
Leukotriene modifier	3364 (18.24)	209 (25.74)	3573 (18.56)
Inhaled corticosteroid/long-acting beta-2 agonist combination	4178 (22.65)	222 (27.34)	4400 (22.85)
Inhaled corticosteroid	6817 (36.96)	424 (52.22)	7241 (37.60)

For each demographic or clinical characteristic, Table 3 presents the statistical test results on whether the data instances related to asthma hospital visits in the subsequent year and those related to no asthma hospital visit in the subsequent year had the same distribution. When the *P* value was ≥.05, the 2 sets of data instances had the same distribution. Otherwise, they had different distributions. All *P* values <.05 are shown in italics in Table 3.

Table 3. For each demographic or clinical characteristic, the statistical test results on whether the data instances related to asthma hospital visits in the subsequent year and those related to no asthma hospital visit in the subsequent year had the same distribution.

Characteristics	*P* value for the 2005-2016 data	*P* value for the 2017 data
Gender	*<.001* [a, b]	*.002* [a]
Age (years)	*<.001* [c]	*<.001* [c]
Ethnicity	*<.001* [a]	*<.001* [a]
Race	*<.001* [a]	*<.001* [a]
Duration of asthma (years)	*<.001* [c]	*<.001* [c]
Insurance category	*<.001* [a]	*<.001* [a]
Smoking status	*<.001* [a]	*<.001* [a]
Comorbidity		
Sleep apnea	*<.001* [a]	*<.001* [a]
Sinusitis	*.006* [a]	.91[a]
Premature birth	*<.001* [a]	*<.001* [a]
Obesity	*<.001* [a]	*.004* [a]
Gastroesophageal reflux	*<.001* [a]	*<.001* [a]
Eczema	*<.001* [a]	*<.001* [a]
Cystic fibrosis	.21[a]	.20[a]
Chronic obstructive pulmonary disease	*<.001* [a]	*<.001* [a]
Bronchopulmonary dysplasia	*<.001* [a]	*.02* [a]
Anxiety or depression	*<.001* [a]	*.002* [a]
Allergic rhinitis	.38[a]	.13[a]
Asthma medication prescription		
Systemic corticosteroid	*<.001* [a]	*<.001* [a]
Short-acting, inhaled beta-2 agonist	*<.001* [a]	*<.001* [a]
Mast cell stabilizer	.29[a]	>.99[a]
Long-acting beta-2 agonist	.67[a]	.11[a]
Leukotriene modifier	*<.001* [a]	*<.001* [a]
Inhaled corticosteroid/long-acting beta-2 agonist combination	*<.001* [a]	*.002* [a]
Inhaled corticosteroid	*<.001* [a]	*<.001* [a]

[a]*P* values obtained by performing the chi-square two-sample test.

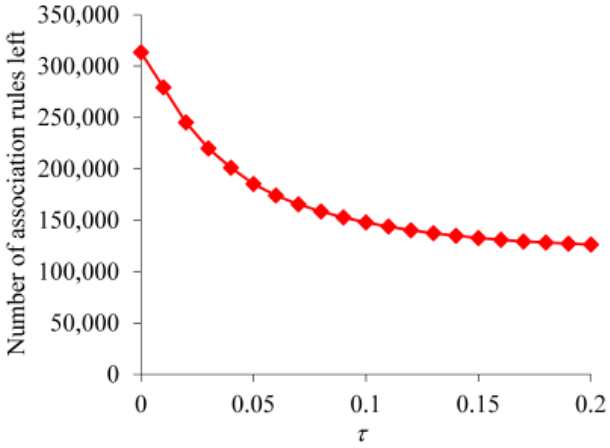
[b]*P* values <.05 marked in italics.

[c]*P* values obtained by performing the Cochran-Armitage trend test [41].

The Number of Association Rules Left at Different Phases of Rule Mining and Pruning Processes

The association rules used in the second model were mined on the training set. Using the top 50 features that were ranked by our XGBoost model with the highest importance values, we obtained 559,834 association rules. Figure 1 presents the number of rules left versus the confidence difference threshold τ. Recall that each more specific rule is dropped when there exists a more general rule whose confidence is not lower by more than τ. Initially, when τ is small, the number of rules left decreases quickly as τ increases. Once τ becomes 0.15 or larger, the number of rules left approaches an asymptote. Accordingly, in our computer coding implementation, we set τ to 0.15, resulting in 132,816 remaining rules.

Figure 1. The number of association rules left versus τ.



A clinical expert on asthma (MJ) in our team marked the values and value ranges of the top 50 features that could possibly have a positive correlation with future asthma hospital visits. After dropping the rules including any other value or value range, 124,506 rules were left. Each rule explains why a patient is predicted to incur one or more asthma hospital visits in the subsequent year. Almost all (124,502/124,506, 100.00%) of these rules were actionable. The left-hand sides of these rules contain various combinations of 208 distinct items related to 50 features.

Example Association Rules in the Second Model

Table 4 presents 5 sample association rules randomly chosen from the 124,502 actionable rules used in the second model.

Table 4. Five sample association rules.

Item on the left-hand side of the rule	Implication of the item	Intervention compiled for the item
Rule 1: The patient had ≥12 ED[a] visits in the past year AND the patient had ≥21 distinct medications in all of the asthma medication orders in the past year → the patient will incur one or more asthma hospital visits in the subsequent year.		
The patient had ≥12 ED visits in the past year	Having many ED visits reflects poor asthma control	Implement control strategies to avoid the need for emergency care
The patient had ≥21 distinct medications in all of the asthma medication orders in the past year	Using many asthma medications reflects poor asthma control	Tailor prescribed asthma medications and help the patient maximize asthma control medication adherence
Rule 2: The patient had ≥9 distinct asthma medication prescribers in the past year AND the block group where the patient lives has a national health literacy score [42] ≤244 AND the patient had ≥21 distinct medications in all of the asthma medication orders in the past year → the patient will incur one or more asthma hospital visits in the subsequent year.		
The patient had ≥9 distinct asthma medication prescribers in the past year	Having many asthma medication prescribers reflects poor care continuity, which often leads to poor outcomes	Provide the patient with social resources to address social chaos that leads to ineffective access to health care
The block group where the patient lives has a national health literacy score ≤244	Having low health literacy is correlated with poor outcomes	Improve education access in the area where the patient lives to help increase health literacy
Rule 3: The patient had a total of ≥25 units of systemic corticosteroids ordered in the past year AND the patient had ≥12 ED visits in the past year AND the patient is Hispanic → the patient will incur one or more asthma hospital visits in the subsequent year.		
The patient had a total of ≥25 units of systemic corticosteroids ordered in the past year	Systemic corticosteroids are one type of asthma medication intended for short-term use to relieve acute asthma exacerbations. Using a lot of systemic corticosteroids reflects poor asthma control	Tailor prescribed asthma medications and help the patient maximize asthma control medication adherence
The patient is Hispanic	In the US, Hispanic people have a disproportionately high rate of poor asthma outcomes	—[b]
Rule 4: The patient had ≥4 major visits for asthma in the past year AND the patient is between 11 and 35 years old AND the patient had no outpatient visit in the past year AND the average length of an inpatient stay of the patient in the past year is >1.75 and ≤2.95 days → the patient will incur one or more asthma hospital visits in the subsequent year.		
The patient had ≥4 major visits for asthma in the past year	As defined in our paper [23], a major visit for asthma is an inpatient stay or ED visit having an asthma diagnosis code, or an outpatient visit having a primary diagnosis of asthma. Intuitively, all else being equal, a patient having major visits for asthma has a higher likelihood of incurring future asthma hospital visits than a patient having only outpatient visits with asthma as a secondary diagnosis	Implement control strategies to avoid the need for emergency care
The average length of an inpatient stay of the patient in the past year is >1.75 and ≤2.95 days	Having inpatient stays reflects poor asthma control	Implement control strategies to avoid the need for emergency care
The patient had no outpatient visit in the past year	For good asthma management, a patient with asthma is supposed to see the primary care provider regularly. Having no outpatient visit often implies that the patient has no primary care provider	Help the patient obtain a primary care provider if the patient does not already have one
Rule 5: The patient had ≥4 major visits for asthma in the past year AND the patient's last ED visit is within the last 49 days AND the patient had between 6 and 8 distinct asthma medication prescribers in the past year AND the patient had a total of ≥36 units of asthma medications ordered in the past year AND >23.7% and ≤33.3% of families in the block group where the patient lives are below 150% of the federal poverty level → the patient will incur one or more asthma hospital visits in the subsequent year.		
The patient's last ED visit is within the last 49 days	Having a recent ED visit reflects poor asthma control	Implement control strategies to avoid the need for emergency care
The patient had a total of ≥36 units of asthma medications ordered in the past year	Taking many asthma medications reflects poor asthma control	Tailor prescribed asthma medications and help the patient maximize asthma control medication adherence
>23.7% and ≤33.3% of families in the block group where the patient lives are below 150% of the federal poverty level	Poverty correlates with poor outcomes	Provide living wage programs in the area where the patient lives to increase resources for health care

[a]ED: emergency department.

[b]Not applicable.

Performance Measures Reached by the Extended Automatic Explanation Method

Our extended automatic explanation method was assessed on the test set. This method explained the prediction results for 92.4% (182/197) of the adults with asthma (age ≥18 years) and 87.5% (209/239) of the children with asthma (age <18 years) for whom our XGBoost model correctly predicted the occurrence of asthma hospital visits in the subsequent year. Combined, our extended method explained the prediction results for 89.7% (391/436) of the patients with asthma whom our XGBoost model correctly predicted to incur asthma hospital visits in the subsequent year. For each such patient, our extended method offered an average of 974.01 (SD 1600.48) explanations, 974.00 (SD 1600.47) of which were actionable. Each explanation came from 1 rule. When confined to using actionable rules, our extended method explained the prediction results for 89.7% (391/436) of the patients with asthma for whom our XGBoost model correctly predicted the occurrence of asthma hospital visits in the subsequent year.

For the patients for whom our extended automatic explanation method could offer explanations of our XGBoost model's correct prediction results of incurring asthma hospital visits in the subsequent year, the average number of distinct actionable items appearing in all of the rules fitting a patient was 21.50 (SD 8.71). This number is much less than 974.01, the average number of actionable rules fitting such a patient.

For the patients with asthma whom our XGBoost model correctly predicted to incur asthma hospital visits in the subsequent year, Figure 2 shows the distribution of patients by the number of rules fitting a patient. This distribution has a long tail and is highly skewed toward the left. As the number of rules fitting a patient becomes larger, the number of patients to each of whom this number of rules apply is inclined to drop nonmonotonically. The largest number of rules fitting a patient is high, 9223, although only 1 patient fits such a high number of rules.

For the patients with asthma whom our XGBoost model correctly predicted to incur asthma hospital visits in the subsequent year, Figure 3 shows the distribution of patients by the number of actionable rules fitting a patient. This distribution is similar to that shown in Figure 2. The largest number of actionable rules fitting a patient is high, 9223, although only 1 patient fits such a high number of actionable rules.

Figure 2. Distribution of patients by the number of rules fitting a patient for the patients with asthma whom our extreme gradient boosting model correctly predicted to incur asthma hospital visits in the subsequent year. (a) When no limit is placed on the number of rules fitting a patient. (b) When the number of rules fitting a patient is ≤250.

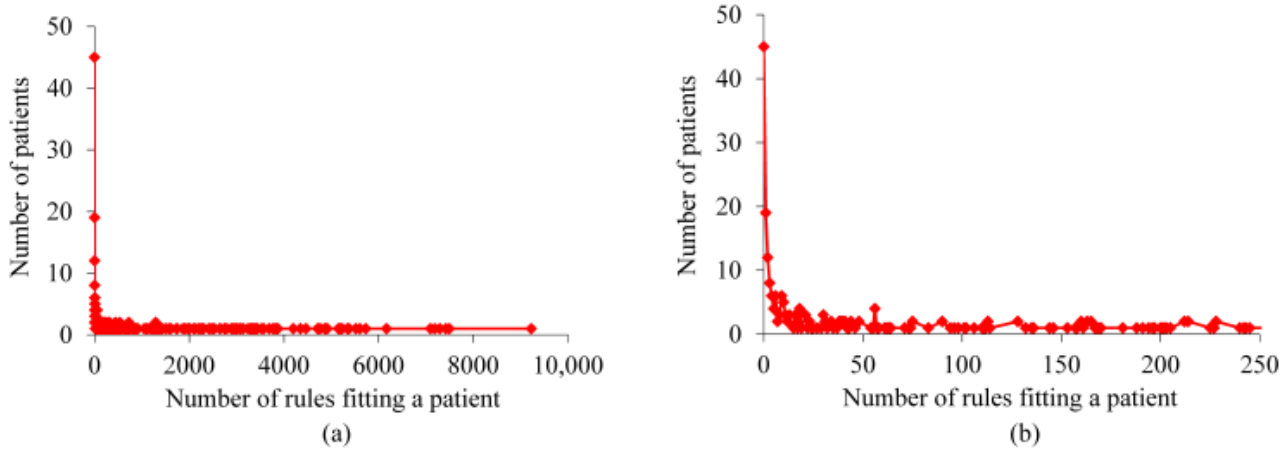
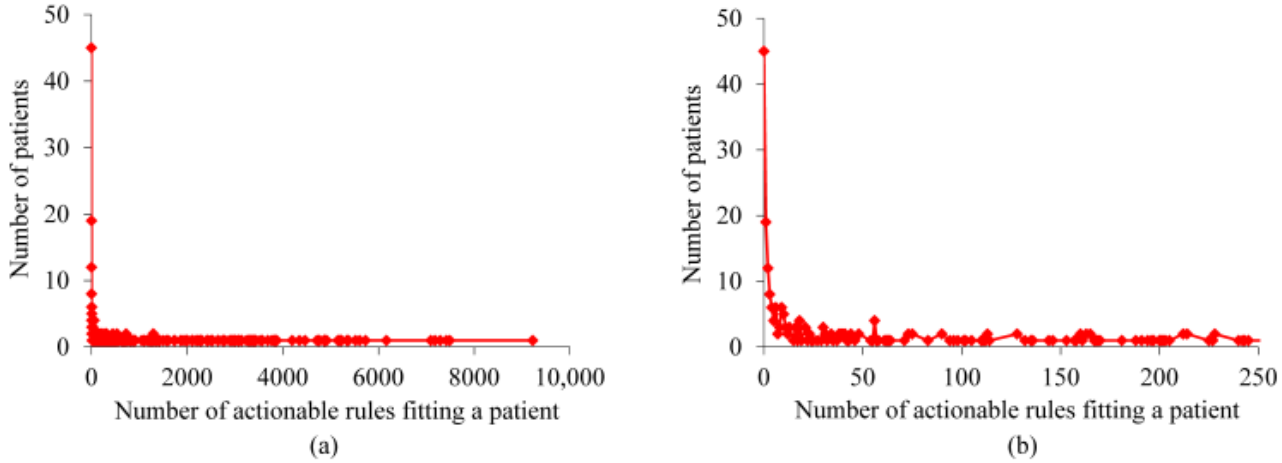


Figure 3. Distribution of patients by the number of actionable rules fitting a patient for the patients with asthma whom our extreme gradient boosting model correctly predicted to incur asthma hospital visits in the subsequent year. (a) When no limit is placed on the number of actionable rules fitting a patient. (b) When the number of actionable rules fitting a patient is ≤250.

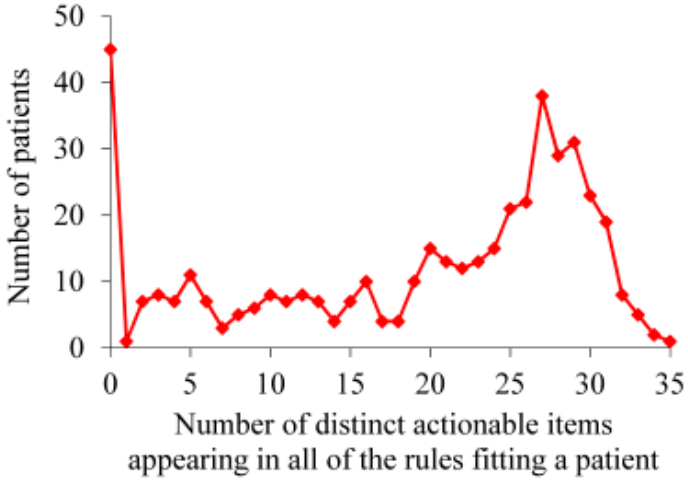


For the patients with asthma whom our XGBoost model correctly predicted to incur asthma hospital visits in the subsequent year, Figure 4 exhibits the distribution of patients by the number of distinct actionable items appearing in all of the rules fitting a patient. The largest number of distinct actionable items appearing in all of the rules fitting a patient is 35, much smaller than the largest number of (actionable) rules fitting a patient. Frequently, 2 or more actionable items appearing in the rules fitting a patient link to the same set of interventions. For example, the intervention of tailoring prescribed asthma medications and helping the patient maximize asthma control medication adherence links to several value ranges of multiple medication-related features.

Our extended automatic explanation method could offer explanations for 69.2% (562/812) of patients with asthma who will incur asthma hospital visits in the subsequent year.

To evaluate the generalizability of our extended automatic explanation method for predicting asthma hospital visits, we tested our method on the University of Washington Medicine data and Kaiser Permanente Southern California data. The results we obtained are similar to the abovementioned results and are detailed in 2 separate papers [43,44].

Figure 4. Distribution of patients by the number of distinct actionable items appearing in all of the rules fitting a patient for the patients with asthma whom our extreme gradient boosting model correctly predicted to incur asthma hospital visits in the subsequent year.



Discussion

Principal Findings

We developed a method to automatically offer rule-formed explanations for any machine learning model's prediction results on imbalanced tabular data without lowering the performance measures of the model. We showed that this method explained the prediction results for 89.7% (391/436) of the patients with asthma whom our XGBoost model correctly predicted to incur asthma hospital visits in the subsequent year. This percentage is high enough for routine clinical use of this method. After further improvement of its accuracy, our asthma outcome prediction model coupled with the automatic explanation function could be used for decision support to guide the allocation of limited asthma care management resources. This could help boost asthma outcomes and reduce resource use and costs.

Our extended automatic explanation method could offer explanations for 69.2% (562/812) of the patients with asthma who will incur asthma hospital visits in the subsequent year.

This percentage is smaller than the success rate of 89.7% (391/436) for our extended automatic explanation method to explain the correct prediction results of our XGBoost model of incurring asthma hospital visits in the subsequent year. One possible reason is that the prediction results of the association rules are correlated with the prediction results of our XGBoost model. Among the patients with asthma who will incur asthma hospital visits in the subsequent year and on whom our XGBoost model gave incorrect predictions, many are difficult cases for any model to correctly predict or explain their outcomes. Among the patients with asthma whom our XGBoost model correctly predicted to incur asthma hospital visits in the subsequent year, many are easy cases for using association rules to explain the outcomes of these cases.

Asthma in adults differs from asthma in children. As shown in a previous study [23], the AUC of our XGBoost model for adults with asthma was 0.034 higher than that for children with asthma, that is, the outcome is easier to predict for adults with asthma than for children with asthma. Intuitively, the degree of difficulty in predicting the outcome is positively correlated with that of using association rules to explain the prediction results of the model, as each rule is a small predictive model. Hence, our extended automatic explanation method explained the prediction results for a larger portion of the adults with asthma than the children with asthma for whom our XGBoost model correctly predicted the occurrence of asthma hospital visits in the subsequent year.

A Guideline for Setting the Values of the Parameters Used in Our Extended Automatic Explanation Method

Our extended automatic explanation method has 4 parameters: the maximum number of items l_{max} allowed on the left-hand side of an association rule, the minimum commonality threshold m_{min}, the minimum confidence threshold c_{min}, and the confidence difference threshold τ. These parameters significantly affect the performance of the method. Our previous papers [31,32] outlined the method but gave no guideline for setting the values of these parameters. We offer such a guideline here.

The maximum number of items l_{max} allowed on the left-hand side of an association rule is usually small, as long rules are difficult to understand [35]. Our previous study [27] showed that for an outcome variable that is relatively easy to predict, an l_{max} of 4 works well for automatic explanation. When the outcome variable is hard to predict, we can increase l_{max} slightly to a number such as 5. Without making the rules too complex to understand, this helps ensure that the second model can provide explanations for a large portion of the data instances that the first model correctly predicts to take one of the interesting values of the outcome variable.

In the original paper [38] that proposed the concept of commonality for class-based association rules, mined rules were used to build a classifier. To maximize the accuracy of the classifier, the minimum commonality threshold m_{min} was set to 14%. However, this value is too high for automatic explanation. With such a high value, we cannot obtain enough rules for the outcome variable's rare values. Consequently, for a large portion of the first model's prediction results on these values, we cannot give any explanation. In addition, the mined rules tend to be too general and have low confidence, causing the users of the automatic explanation function to have little trust in the automatically generated explanations. To avoid these problems, for automatic explanation, we recommend setting m_{min} to a value much smaller than 14%. More specifically, our paper [27] showed that on reasonably balanced data, a minimum support threshold s_{min} of 1% and a minimum confidence threshold c_{min} of 50% work well for automatic explanation. By definition, commonality is a value-specific support. Thus, we would expect m_{min} and s_{min} to have relatively similar optimal values. Accordingly, we set m_{min} to a value close to 1% and c_{min} to a value close to 50%. Although a value close to 50% may not seem so high, it is already much larger than the percentage of data instances linking to an interesting value of the outcome variable. For instance, in our case of predicting asthma hospital visits in patients with asthma, this percentage is 4% [23]. Moreover, a value close to 50% is also much larger than our XGBoost model's positive predictive value of 22.65%. The concrete values of m_{min} and c_{min} depend on the data set and are chosen to meet 2 goals simultaneously and as much as possible. First, the second model can provide explanations for a large portion of the data instances that the first model correctly predicts to take one of the interesting values of the outcome variable. Often, the harder the outcome variable is to predict, the smaller m_{min} and c_{min} need to be to meet this goal. Second, c_{min} is high enough for users of the automatic explanation function to trust the automatically generated explanations.

Recall that during the rule-pruning process, each more specific rule is dropped when there is a more general rule whose confidence is not lower by more than the confidence difference threshold τ. To determine the value of τ, we plot the number of rules left versus τ. As our previous paper [27] shows, initially when τ is small, the number of rules left decreases quickly as τ increases. Once τ becomes sufficiently large, the number of rules left approaches an asymptote. This is the place to set the value of τ to strike a balance between cutting the number of rules and retaining high-confidence rules.

Five Clarifications on Using the Automatic Explanation Function

In practice, our automatic explanation method could produce a paradox. Two patients both fulfilled the left-hand side of the same rule linking to a poor outcome. The first model correctly predicts one of them to have a poor outcome. The automatic explanation function displays the rule to explain this prediction result. Simultaneously, the first model correctly predicts a good outcome on the other patient, for whom the automatic explanation function shows nothing. In this case, one should not think that the automatic explanation function acts incorrectly because it behaves differently in these 2 patients; rather, this difference occurs because the second patient fulfills some items that are not in the rule. These items counter the risk induced by those on the rule's left-hand side and reduce the second patient's risk of having a poor outcome to a low level.

When using the automatic explanation function, one needs to remember that the function is intended to serve as a reminder

system for decision support rather than a replacement for clinical judgment. The function is used to help the user quickly identify some reasons why a patient is predicted to have a poor outcome and some tailored interventions suitable for the patient. If successful, this helps the clinical user avoid substantial time laboriously reviewing the records of the patient to assess risk factors and devise interventions. This also helps reduce the number of interventions that are suitable for the patient, but the user forgets to consider. In the end, it is still the user who uses his or her own judgment to decide whether to use the prediction result of the first model and apply suggested interventions to the patient. If there is doubt about the appropriateness of the output of the function, the clinical user can always check the records of the patient to resolve the doubt before making the final decisions with the patient.

Different health care systems have different properties and practice patterns. Consequently, the association rules mined from the data of one health care system may or may not directly apply to or work well for another health care system. However, our automatic explanation method is general. It relies on no special property of a specific disease, patient cohort, prediction target, or health care system and can be applied to various predictive modeling problems and health care systems [27,29,30,43,44], regardless of whether the rules mined from the data of 1 health care system generalize to the data of another health care system. For any health care system, we would recommend mining rules from its own data whenever possible, rather than reusing the rules mined from the data of another health care system.

In our test case, the second model contained 124,506 association rules. The left-hand sides of these rules contain various combinations of 208 distinct items related to 50 features. Within 1 day, a clinician in our team (MJ) finished manually compiling the 2 types of knowledge needed by the automatic explanation function: the values and value ranges of the top 50 features that could possibly have a positive correlation with future asthma hospital visits and the interventions for the actionable items. The amount of time needed to perform this manual compilation is moderate and acceptable to the clinicians in our team.

Although many association rules could fit a patient, the total number of distinct items included on their left-hand sides is not large: at most 35. To avoid overwhelming the automatic explanation function's user, we can use the rule diversification method in our paper [27] to rank these rules. The top few rules are likely to contain nonredundant information and are displayed by default.

Related Work

As described in a survey paper [33] and a book [34], other researchers previously proposed various methods for automatically explaining the prediction results of machine learning models. These methods often lower the performance measures of the model by replacing the original model with a less accurate model and usually give nonrule-formed explanations. Many of these methods work for only a specific machine learning algorithm rather than for all algorithms. Moreover, none of these methods can automatically recommend tailored interventions. In comparison, our extended automatic explanation method not only offers rule-formed explanations for the prediction results of any machine learning model on tabular data but also recommends tailored interventions without lowering the performance measures of the model [27]. Compared with nonrule-formed explanations, rule-formed explanations are easier to comprehend and can more directly recommend tailored interventions.

Hatwell et al [45] proposed a method to automatically provide rule-formed explanations for the prediction results of an AdaBoost model. This method does not work for non-AdaBoost machine learning algorithms. The rules are unknown before the prediction time and hence cannot be used to automatically recommend tailored interventions at prediction time. In comparison, the rules used in our extended automatic explanation method are precompiled beforehand and used to automatically recommend tailored interventions at prediction time.

Limitations

This study has 2 limitations that give interesting directions for future work:

1. Our data set contained no information on health care use of the patients outside of Intermountain Healthcare. Consequently, the features were computed using incomplete clinical and administrative data [46-49]. In addition, the prediction target was limited to asthma hospital visits at Intermountain Healthcare rather than asthma hospital visits anywhere. It would be interesting to see how the automatically generated explanations of the prediction results of the model would differ if we have access to more complete clinical and administrative data [50].
2. Our study used 1 predictive modeling problem, predicting asthma hospital visits as the test case. Although our original automatic explanation method [27] has been successfully applied to several predictive modeling problems [29,30], the generalizability of our extended automatic explanation method to other predictive modeling problems beyond predicting asthma hospital visits has not been evaluated. Conducting such evaluations would help inform the utility and refine the implementation of our extended method.

Conclusions

Using asthma outcome prediction as a demonstration case, this study shows for the first time the feasibility of automatically offering rule-formed explanations for the prediction results of any machine learning model on imbalanced tabular data without lowering the performance measures of the model. After further improvement, our asthma outcome prediction model coupled with the automatic explanation function could be used for decision support to guide the allocation of limited asthma care management resources. This could simultaneously help improve asthma outcomes and reduce resource use and cost.

Acknowledgments

The authors would like to thank Katherine A Sward for the useful discussions. GL, MJ, FN, SH, and BS were 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 the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' Contributions

GL was responsible for the paper. GL conceptualized and designed the study, performed the literature review and data analysis, and wrote the paper. MJ, FN, and BS provided feedback on various medical issues, contributed to conceptualizing the presentation, and revised the paper. SH took part in retrieving the Intermountain Healthcare data set and interpreting its detected peculiarities.

Conflicts of Interest

None declared.

References

1. Moorman JE, Akinbami LJ, Bailey CM, Zahran HS, King ME, Johnson CA, et al. National surveillance of asthma: United States, 2001-2010. Vital Health Stat 3 2012 Nov(35):1-58. [Medline: 24252609]
2. Nurmagambetov T, Kuwahara R, Garbe P. The economic burden of asthma in the United States, 2008-2013. Ann Am Thorac Soc 2018 Mar;15(3):348-356. [doi: 10.1513/AnnalsATS.201703-259OC] [Medline: 29323930]
3. Mays GP, Claxton G, White J. Managed care rebound? Recent changes in health plans' cost containment strategies. Health Aff (Millwood) 2004;Suppl Web Exclusives:W4-427. [doi: 10.1377/hlthaff.w4.427] [Medline: 15451964]
4. Lieu TA, Quesenberry CP, Sorel ME, Mendoza GR, Leong AB. Computer-based models to identify high-risk children with asthma. Am J Respir Crit Care Med 1998 Apr;157(4 Pt 1):1173-1180. [doi: 10.1164/ajrccm.157.4.9708124] [Medline: 9563736]
5. Caloyeras JP, Liu H, Exum E, Broderick M, Mattke S. Managing manifest diseases, but not health risks, saved PepsiCo money over seven years. Health Aff (Millwood) 2014 Jan;33(1):124-131. [doi: 10.1377/hlthaff.2013.0625] [Medline: 24395944]
6. Greineder DK, Loane KC, Parks P. A randomized controlled trial of a pediatric asthma outreach program. J Allergy Clin Immunol 1999 Mar;103(3 Pt 1):436-440. [doi: 10.1016/s0091-6749(99)70468-9] [Medline: 10069877]
7. Kelly CS, Morrow AL, Shults J, Nakas N, Strope GL, Adelman RD. Outcomes evaluation of a comprehensive intervention program for asthmatic children enrolled in medicaid. Pediatrics 2000 May;105(5):1029-1035. [doi: 10.1542/peds.105.5.1029] [Medline: 10790458]
8. Axelrod RC, Zimbro KS, Chetney RR, Sabol J, Ainsworth VJ. A disease management program utilizing life coaches for children with asthma. J Clin Outcomes Manag 2001;8(6):38-42.
9. Axelrod RC, Vogel D. Predictive modeling in health plans. Dis Manag Health Out 2003;11(12):779-787. [doi: 10.2165/00115677-200311120-00003]
10. Loymans RJ, Debray TP, Honkoop PJ, Termeer EH, Snoeck-Stroband JB, Schermer TR, et al. Exacerbations in adults with asthma: a systematic review and external validation of prediction models. J Allergy Clin Immunol Pract 2018;6(6):1942-1952.e15. [doi: 10.1016/j.jaip.2018.02.004] [Medline: 29454163]
11. Loymans RJ, Honkoop PJ, Termeer EH, Snoeck-Stroband JB, Assendelft WJ, Schermer TR, et al. Identifying patients at risk for severe exacerbations of asthma: development and external validation of a multivariable prediction model. Thorax 2016 Sep;71(9):838-846. [doi: 10.1136/thoraxjnl-2015-208138] [Medline: 27044486]
12. Schatz M, Cook EF, Joshua A, Petitti D. Risk factors for asthma hospitalizations in a managed care organization: development of a clinical prediction rule. Am J Manag Care 2003 Aug;9(8):538-547 [FREE Full text] [Medline: 12921231]
13. Eisner MD, Yegin A, Trzaskoma B. Severity of asthma score predicts clinical outcomes in patients with moderate to severe persistent asthma. Chest 2012 Jan;141(1):58-65. [doi: 10.1378/chest.11-0020] [Medline: 21885725]
14. Sato R, Tomita K, Sano H, Ichihashi H, Yamagata S, Sano A, et al. The strategy for predicting future exacerbation of asthma using a combination of the asthma control test and lung function test. J Asthma 2009 Sep;46(7):677-682. [doi: 10.1080/02770900902972160] [Medline: 19728204]
15. Osborne ML, Pedula KL, O'Hollaren M, Ettinger KM, Stibolt T, Buist AS, et al. Assessing future need for acute care in adult asthmatics: the Profile of Asthma Risk Study: a prospective health maintenance organization-based study. Chest 2007 Oct;132(4):1151-1161. [doi: 10.1378/chest.05-3084] [Medline: 17573515]
16. Miller MK, Lee JH, Blanc PD, Pasta DJ, Gujrathi S, Barron H, TENOR Study Group. TENOR risk score predicts healthcare in adults with severe or difficult-to-treat asthma. Eur Respir J 2006 Dec;28(6):1145-1155 [FREE Full text] [doi: 10.1183/09031936.06.00145105] [Medline: 16870656]
17. Peters D, Chen C, Markson LE, Allen-Ramey FC, Vollmer WM. Using an asthma control questionnaire and administrative data to predict health-care utilization. Chest 2006 Apr;129(4):918-924. [doi: 10.1378/chest.129.4.918] [Medline: 16608939]
18. Yurk RA, Diette GB, Skinner EA, Dominici F, Clark RD, Steinwachs DM, et al. Predicting patient-reported asthma outcomes for adults in managed care. Am J Manag Care 2004 May;10(5):321-328 [FREE Full text] [Medline: 15152702]

19. Lieu TA, Capra AM, Quesenberry CP, Mendoza GR, Mazar M. Computer-based models to identify high-risk adults with asthma: is the glass half empty of half full? J Asthma 1999 Jun;36(4):359-370. [Medline: 10386500]
20. Schatz M, Nakahiro R, Jones CH, Roth RM, Joshua A, Petitti D. Asthma population management: development and validation of a practical 3-level risk stratification scheme. Am J Manag Care 2004 Jan;10(1):25-32 [FREE Full text] [Medline: 14738184]
21. Grana J, Preston S, McDermott PD, Hanchak NA. The use of administrative data to risk-stratify asthmatic patients. Am J Med Qual 1997;12(2):113-119. [doi: 10.1177/0885713X9701200205] [Medline: 9161058]
22. Forno E, Fuhlbrigge A, Soto-Quirós ME, Avila L, Raby BA, Brehm J, et al. Risk factors and predictive clinical scores for asthma exacerbations in childhood. Chest 2010 Nov;138(5):1156-1165 [FREE Full text] [doi: 10.1378/chest.09-2426] [Medline: 20472862]
23. 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]
24. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: KDD'16; August 13-17, 2016; San Francisco, CA p. 785-794. [doi: 10.1145/2939672.2939785]
25. Desai JR, Wu P, Nichols GA, Lieu TA, O'Connor PJ. Diabetes and asthma case identification, validation, and representativeness when using electronic health data to construct registries for comparative effectiveness and epidemiologic research. Med Care 2012 Jul;50 Suppl:S30-S35. [doi: 10.1097/MLR.0b013e318259c011] [Medline: 22692256]
26. Wakefield DB, Cloutier MM. Modifications to HEDIS and CSTE algorithms improve case recognition of pediatric asthma. Pediatr Pulmonol 2006 Oct;41(10):962-971. [doi: 10.1002/ppul.20476] [Medline: 16871628]
27. 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]
28. Luo G, Stone BL, Sakaguchi F, Sheng X, Murtaugh MA. Using computational approaches to improve risk-stratified patient management: rationale and methods. JMIR Res Protoc 2015 Oct 26;4(4):e128 [FREE Full text] [doi: 10.2196/resprot.5039] [Medline: 26503357]
29. 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. [doi: 10.1038/s41598-018-29523-2] [Medline: 30050169]
30. Alaa AM, van der Schaar M. AutoPrognosis: Automated Clinical Prognostic Modeling via Bayesian Optimization With Structured Kernel Learning. In: Proceedings of 35th International Conference on Machine Learning. 2018 Presented at: ICML'18; July 10-15, 2018; Stockholm, Sweden p. 139-148.
31. 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]
32. Luo G, Stone BL, Koebnick C, He S, Au DH, Sheng X, et al. Using temporal features to provide data-driven clinical early warnings for chronic obstructive pulmonary disease and asthma care management: protocol for a secondary analysis. JMIR Res Protoc 2019 Jun 6;8(6):e13783 [FREE Full text] [doi: 10.2196/13783] [Medline: 31199308]
33. 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]
34. Molnar C. Interpretable Machine Learning. Morrisville, NC: lulu.com; 2020.
35. Liu B, Hsu W, Ma Y. Integrating classification and association rule mining. In: Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. 1998 Presented at: KDD'98; August 27-31, 1998; New York City, NY p. 80-86.
36. Thabtah F. A review of associative classification mining. Knowl Eng Rev 2007 Mar 1;22(1):37-65. [doi: 10.1017/s0269888907001026]
37. Fayyad UM, Irani KB. Multi-interval Discretization of Continuous-valued Attributes for Classification Learning. In: Proceedings of the 13th International Joint Conference on Artificial Intelligence. 1993 Presented at: IJCAI'93; August 28-September 3, 1993; Chambéry, France p. 1022-1029.
38. Paul R, Groza T, Hunter J, Zankl A. Inferring characteristic phenotypes via class association rule mining in the bone dysplasia domain. J Biomed Inform 2014 Apr;48:73-83 [FREE Full text] [doi: 10.1016/j.jbi.2013.12.001] [Medline: 24333481]
39. Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques, 3rd ed. Waltham, MA: Morgan Kaufmann; 2011.
40. Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Burlington, MA: Morgan Kaufmann; 2016.
41. Agresti A. Categorical Data Analysis, 3rd ed. Hoboken, NJ: Wiley; 2012.
42. Health Literacy Data Map. US Health Literacy Scores. 2020. URL: http://healthliteracymap.unc.edu [accessed 2020-12-10]
43. 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 healthcare system. IEEE Access 2020;8:195971-195979. [doi: 10.1109/access.2020.3032683]

44. Luo G, Nau CL, Crawford WW, Schatz M, Zeiger RS, Koebnick C. Assessing the Generalizability of an Automatic Explanation Method for Machine Learning Prediction Results: A Secondary Analysis on Forecasting Asthma-related Hospital Visits in Patients With Asthma. UW Computer Sciences User Pages. URL: http://pages.cs.wisc.edu/~gangluo/explain_predict_hospital_use_for_asthma_KPSC.pdf [accessed 2020-12-10]
45. Hatwell J, Gaber MM, Atif Azad RM. Ada-WHIPS: explaining AdaBoost classification with applications in the health sciences. BMC Med Inform Decis Mak 2020 Oct 2;20(1):250 [FREE Full text] [doi: 10.1186/s12911-020-01201-2] [Medline: 33008388]
46. Bourgeois FC, Olson KL, Mandl KD. Patients treated at multiple acute health care facilities: quantifying information fragmentation. Arch Intern Med 2010 Dec 13;170(22):1989-1995. [doi: 10.1001/archinternmed.2010.439] [Medline: 21149756]
47. Finnell JT, Overhage JM, Grannis S. All health care is not local: an evaluation of the distribution of emergency department care delivered in Indiana. AMIA Annu Symp Proc 2011;2011:409-416 [FREE Full text] [Medline: 22195094]
48. Luo G, Tarczy-Hornoch P, Wilcox AB, Lee ES. Identifying patients who are likely to receive most of their care from a specific health care system: demonstration via secondary analysis. JMIR Med Inform 2018 Nov 5;6(4):e12241 [FREE Full text] [doi: 10.2196/12241] [Medline: 30401670]
49. Kern LM, Grinspan Z, Shapiro JS, Kaushal R. Patients' use of multiple hospitals in a major US city: implications for population management. Popul Health Manag 2017 Apr;20(2):99-102 [FREE Full text] [doi: 10.1089/pop.2016.0021] [Medline: 27268133]
50. Samuels-Kalow ME, Faridi MK, Espinola JA, Klig JE, Camargo CA. Comparing statewide and single-center data to predict high-frequency emergency department utilization among patients with asthma exacerbation. Acad Emerg Med 2018 Jun;25(6):657-667 [FREE Full text] [doi: 10.1111/acem.13342] [Medline: 29105238]

Abbreviations

AUC: area under the receiver operating characteristic curve
ED: emergency department
ICD-9: International Classification of Diseases, ninth revision
ICD-10: International Classification of Diseases, tenth revision
XGBoost: extreme gradient boosting

Edited by C Lovis; submitted 03.07.20; peer-reviewed by L Zhou, M Simons; comments to author 18.10.20; revised version received 25.10.20; accepted 15.11.20; published 31.12.20.

Please cite as:
Luo G, Johnson MD, Nkoy FL, He S, Stone BL
Automatically Explaining Machine Learning Prediction Results on Asthma Hospital Visits in Patients With Asthma: Secondary Analysis
JMIR Med Inform 2020;8(12):e21965
URL: http://medinform.jmir.org/2020/12/e21965/
doi:10.2196/21965
PMID:33382379

Original Paper

Extracting Family History of Patients From Clinical Narratives: Exploring an End-to-End Solution With Deep Learning Models

Xi Yang[1,2], PhD; Hansi Zhang[1], MSc; Xing He[1], MSc; Jiang Bian[1,2], PhD; Yonghui Wu[1,2], PhD

[1]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, United States

[2]Cancer Informatics Shared Resource, University of Florida Health Cancer Center, Gainesville, FL, United States

Corresponding Author:
Yonghui Wu, PhD
Department of Health Outcomes and Biomedical Informatics
College of Medicine
University of Florida
2004 Mowry Rd
Gainesville, FL, 32610
United States
Phone: 1 352 294 8436
Email: yonghui.wu@ufl.edu

Abstract

Background: Patients' family history (FH) is a critical risk factor associated with numerous diseases. However, FH information is not well captured in the structured database but often documented in clinical narratives. Natural language processing (NLP) is the key technology to extract patients' FH from clinical narratives. In 2019, the National NLP Clinical Challenge (n2c2) organized shared tasks to solicit NLP methods for FH information extraction.

Objective: This study presents our end-to-end FH extraction system developed during the 2019 n2c2 open shared task as well as the new transformer-based models that we developed after the challenge. We seek to develop a machine learning–based solution for FH information extraction without task-specific rules created by hand.

Methods: We developed deep learning–based systems for FH concept extraction and relation identification. We explored deep learning models including long short-term memory-conditional random fields and bidirectional encoder representations from transformers (BERT) as well as developed ensemble models using a majority voting strategy. To further optimize performance, we systematically compared 3 different strategies to use BERT output representations for relation identification.

Results: Our system was among the top-ranked systems (3 out of 21) in the challenge. Our best system achieved micro-averaged F1 scores of 0.7944 and 0.6544 for concept extraction and relation identification, respectively. After challenge, we further explored new transformer-based models and improved the performances of both subtasks to 0.8249 and 0.6775, respectively. For relation identification, our system achieved a performance comparable to the best system (0.6810) reported in the challenge.

Conclusions: This study demonstrated the feasibility of utilizing deep learning methods to extract FH information from clinical narratives.

(JMIR Med Inform 2020;8(12):e22982) doi:10.2196/22982

KEYWORDS

family history; information extraction; natural language processing; deep learning

Introduction

Patients' family history (FH) is a critical risk factor associated with numerous diseases [1-3] such as diabetes [4], coronary heart disease [5], and multiple types of cancers [6-9]. For example, a previous study showed that if a female patient has both her mother and sister having breast cancer, her relative risk [10] of having breast cancer increased 3.6 times compared with people without such FH [11]. Knowing the FH of patients can greatly help the prevention, diagnosis, and treatment of various diseases. However, FH is not well structured in current electronic health record databases but often documented as free text in clinical notes. Manually extracting patients' FH information is a labor-intensive and time-consuming procedure that cannot be scaled up. Natural language processing (NLP) is the key technology to build automated computational models

to extract patients' FH from clinical narratives in their electronic health records.

In the past 2 decades, researchers have invested a significant amount of effort into developing various methods and tools to extract patients' information from clinical narratives [12-14]. The clinical NLP community has organized a series of shared tasks for retrieving various patients' information from clinical narratives including diseases or disorders [15-17], adverse drug events [18,19], and medical temporal relations [20]. Both rule-based and machine learning–based methods have been examined, and clinical NLP systems such as MetaMap [21], cTAKES [22], and CLAMP [23] have been developed. More recently, deep learning–based approaches have demonstrated superior performances in many NLP tasks [24]. For example, the long short-term memory-conditional random fields (LSTM-CRFs) architecture [25], which is a modified implementation of the recurrent neural network, has been widely adopted for named entity recognition (NER) tasks in both general and clinical domains. Later, a newly emerged bidirectional encoder representations from transformers (BERT) model achieved state-of-the-art performances in 20 NLP benchmarks in the general English domain [26] and demonstrated promising results in several clinical NLP tasks [27-29]. However, there are only a handful of studies focused on extracting FH of patients [30-32], which is more complicated than merely extracting information of the patients as it relates to various family members of the patient. FH often contains information from different aspects of the patients, including family members, their living status, and their diseases or disorders. Furthermore, patient's family members need to be characterized by family role (eg, mother) and family side (eg, maternal). Besides, there are limited clinical corpora annotated for FH. The 2018 BioCreative/OHNLP Challenge [33,34] is the first shared task focusing on FH extraction. During that challenge, Shi et al [35] explored a joint deep learning approach and achieved the best performance among all participated teams.

In 2019, the National NLP Clinical Challenge (n2c2) organized shared tasks to solicit advanced NLP methods for extracting FH information from clinical text. The 2019 n2c2 open shared task consisted of 2 subtasks: (1) NER for family members and observations (ie, diseases or disorders); and (2) identifying relations between family members, observations, and living status. Participants were required to identify mentions of FH and present a family member as a combination of family role (eg, mother) and family side (eg, maternal) and living status as a score derived from the healthy and alive state.

This paper presents our end-to-end FH extraction system developed during the 2019 n2c2 open shared task as well as new transformer models we developed after the challenge. During this challenge, we adopted an LSTM-CRF model for NER and a BERT-based model for relation identification. Our best submission was ranked fifth in subtask 1 and third in subtask 2. After the challenge, we further explored a BERT-based model for NER and demonstrated better performances in both subtasks.

Methods

Data

This study used the data set developed by the 2019 n2c2 open shared task organizers consisting of 216 clinical notes extracted from the Mayo Clinic data warehouse. The organizers split the corpus into a training set of 99 notes and a test set of 117 notes. Three types of concepts were annotated, including family members, observations (ie, diseases and disorders), and living status. There are also 2 types of relations annotated among family members, observations, and living status. The organizers provided annotations at (1) entity level (ie, the words and phrases about FH), and (2) document level, where the multiple mentions of the same FH were aggregated. Table 1 shows the descriptive statistics of the corpus.

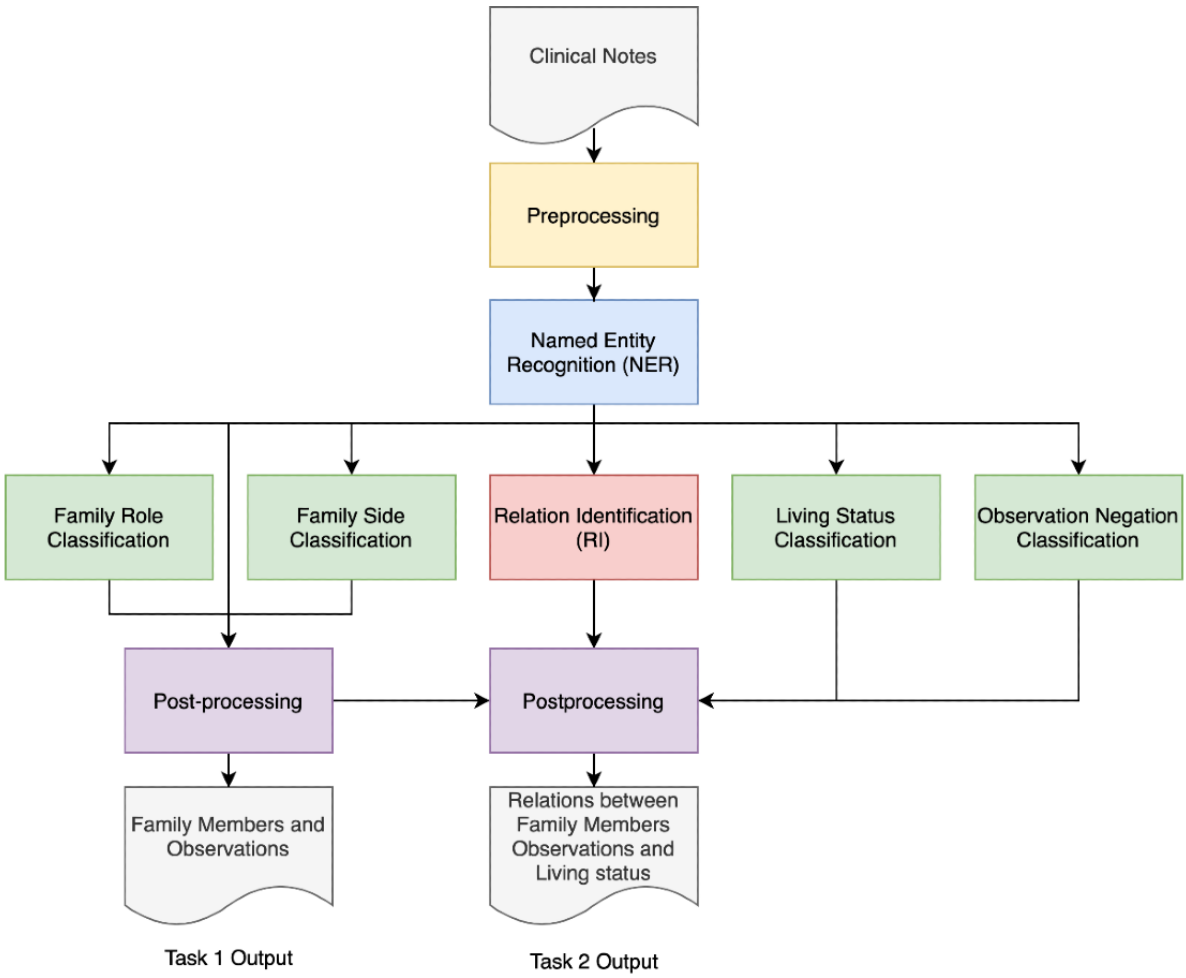
Table 1. Descriptive statistics of the challenge data set.

Corpus information, annotation type, and annotation category	2019 n2c2 family history challenge corpus	
	Training set	Test set
Number of notes	99	117
Entity-level annotation		
Concept		
Family members	803	N/A
Observations	978	N/A
Living status	415	N/A
Document-level annotation		
Concept		
Family members	667	638
Observations	930	983
Relation		
Family members—observations	740	755
Family members—living status	376	349

The Family History Extraction System

Figure 1 shows the system architecture for our end-to-end FH extraction system. Our system has 5 modules including preprocessing, NER, classification, relation identification, and postprocessing. The preprocessing module contains standard NLP procedures including tokenization, sentence boundary detection, and data format transformation. In the NER module, we explored state-of-the-art NLP models, including LSTM-CRFs and BERT to identify FH concepts. The relation identification module applied deep learning models to determine the relations among FH concepts. The postprocessing module aggregated the entity-level results to the document level for both concept extraction and relation identification subtasks.

Figure 1. Overview of our family history extraction system.

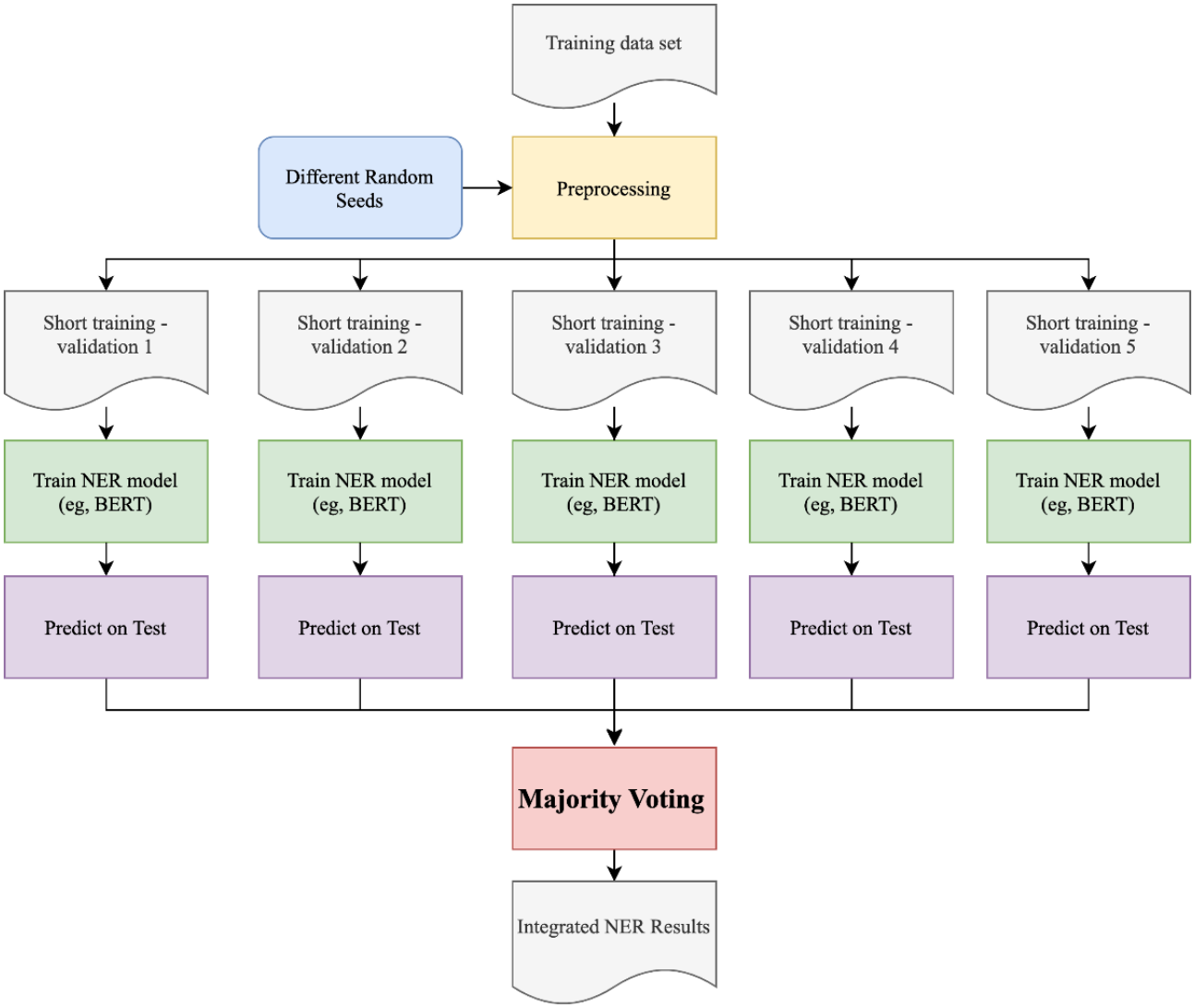


Extracting Family History Concepts

The concept extraction subtask focused on detecting the mentions of family members and observations. We approached this subtask as a typical NER problem and applied deep learning–based models. Following the standard machine learning–based NER procedure, we converted the annotations using the beginning-inside-outside (BIO) tagging scheme [36,37], where "B" indicates the first token of a concept, "I" indicates tokens inside of a concept, and "O" indicates tokens that do not belong to any concepts. Thus, we converted information extraction problem into a sequence labeling task to assign each word with one of the predefined NER labels ("B," "I," or "O"). We explored 2 deep learning–based models including LSTM-CRFs and BERT.

Previous studies [38-41] have shown that adopting an ensemble method could further improve the clinical NER performances. Thus, we adopted the majority voting strategy to integrate the different NER models as shown in Figure 2. More specifically, we randomly (based on a random seed) split the training data into a short training data and a validation data at a 9:1 ratio. We trained deep learning models using the short training data and selected the best checkpoints based on the model performance on the validation data. By repeating the procedure 5 times with different random seeds, we obtained 5 different models. In each training procedure, we used different short training data and validate data but the same hyperparameters (ie, the optimized hyperparameters used for training the single BERT NER model). Then, the majority voting strategy was used to vote among the 5 models. Here, we use a suffix "-EN" to indicate the ensemble method. For example, we used "LSTM-CRFs-EN" to denote the ensemble model of LSTM-CRFs, and "BERT-EN" to denote the ensemble model of using BERT.

Figure 2. The majority voting strategy to ensemble NER models. BERT: bidirectional encoder representations from transformers; NER: named entity recognition.



Determining Family Role and Family Side

This task is to determine the family role and family side for the mentions of FH. There are a total number of 15 types of family roles defined in this challenge, including father, mother, sister, parent, brother, grandmother, grandfather, grandparent, daughter, son, child, cousin, sibling, aunt, and uncle. There are 3 predefined family sides including maternal, paternal, and not applicable. We approached the 2 tasks as classification problems. Previous studies [35,42] approached the 2 tasks using rule-based methods; here, we applied deep learning–based classification methods as machine learning–based methods have shown a better generalizability.

Relation Identification

Typically, relation identification consists of 2 steps: (1) determine whether there is a relation between 2 entities; and (2) classify the correct relation type. In this study, we formulated the relation identification as a binary classification problem. We presented each relation as a pair of 2 entities and used contextual information around the entities to classify these pairs into categories as “in-relation” or “nonrelation” (no relation between entities). Then, we further categorized the “in-relation” entity pairs into either “family member—living status” group or “family member—observation” group based on the entity types: if 1 of the entities in an entity pair is observation, we classify it as “family member—observation”; if one of the entities in an entity pair is living status, we classify it as “family member—living status.”

Candidate Concept Pairs Generation

Theoretically, there might be relations between any pair of FH concepts. Thus, a naïve way is to generate candidate pairs from all combinations of clinical concepts in document level. However, a previous study [43] has reported that this method often generates too many negative samples (ie, nonrelation), causing an extremely imbalanced positive-to-negative sample ratio. To alleviate this issue, we applied the following heuristic rule to reduce the combinations: only keep the concept pairs composed of a family member entity as the first element and a nonfamily member entity as the second element. We also looked into the cross-distance of pairs—defined as the number of sentence boundaries between the 2 entities (eg, 0 for single-sentence relations, and 1 for relations across 2 sentences). In the training set, the cross-distance ranges from 0 to 10 and we found that 96% of the annotated relations have

cross-distances less than 3. Therefore, we only consider candidate pairs with cross-distances less than 3. Previous studies [44,45] developed individual classifiers to handle relations with different cross-distance; here, we developed a unified BERT-based classifier to handle all candidate pairs with various cross-distances as the BERT model is able to learn both token- and sentence-level representations.

Handling Negations

In this study, we approached negation detection as a binary classification problem—classify the observation entity into 2 predefined categories including "negated" and "non-negated." We developed a BERT-based classifier for negation detection. In our system, we performed the negation detection for each observation entity and then integrated the results into relations. We only used the negation annotations from the challenge data set and did not use any external resources.

Assessing the Living Status Scores

For the relations between "family member—living status," the participants were required to assess the living status using scores of 0, 2, or 4, where 0 indicates not alive, 2 indicates alive but not healthy, and 4 indicates alive and healthy. We approached this task as a classification task—to categorize a living status entity into one of 3 score categories (ie, 0, 2, and 4). We developed a BERT-based classifier to classify each living status entity into a category according to its context.

Deep Learning Models

LSTM-CRFs

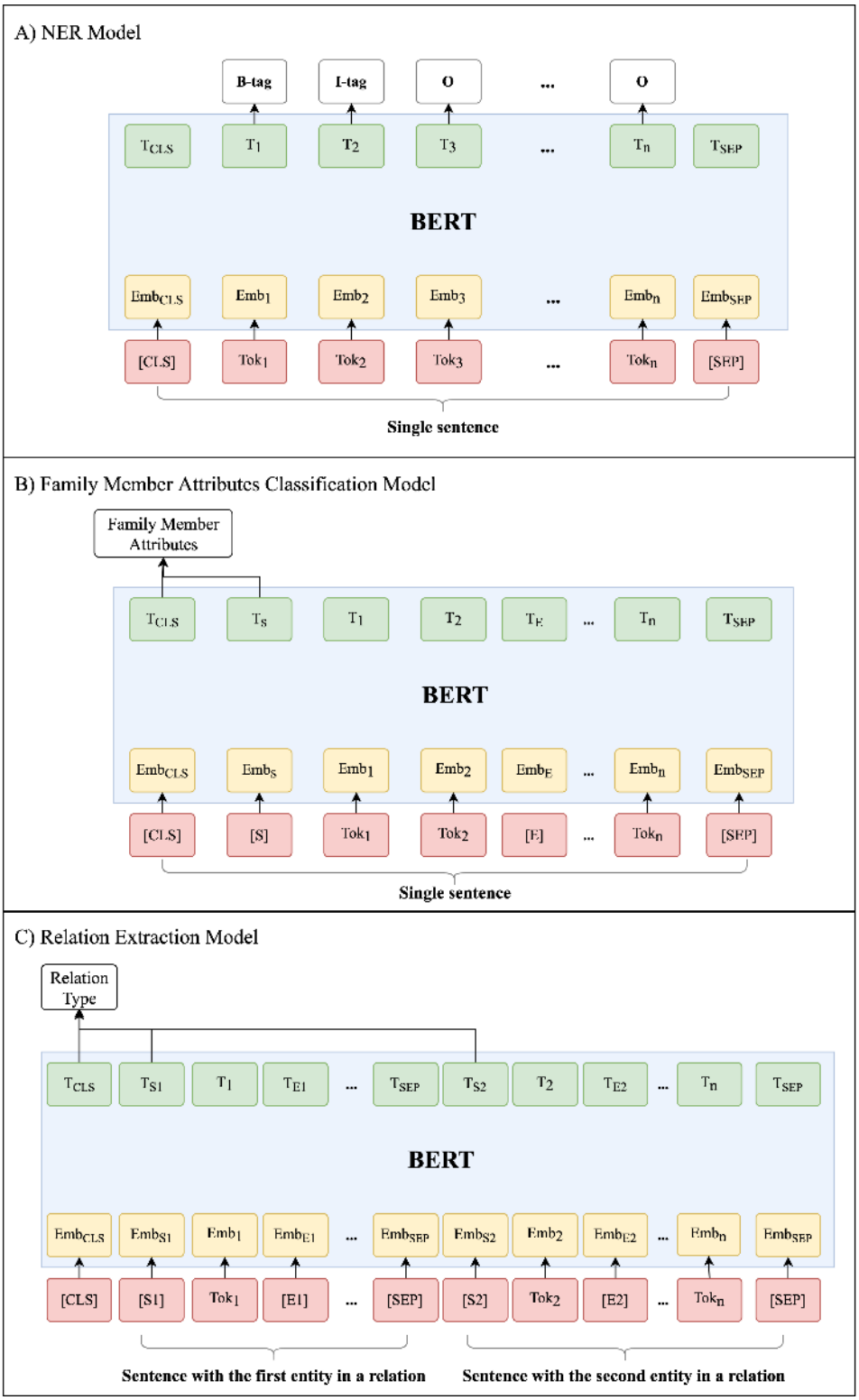
In this study, we adopted an LSTM-CRFs architecture proposed by Lample et al [25]. The model has 2 bidirectional LSTM layers: one for learning representations at the character level and the other for learning those at the word level. The model utilizes a CRFs layer to decode the LSTM hidden states to BIO tags. We screened 4 different word embeddings following a similar procedure reported in our previous study [46] and found that the Common Crawl embeddings—released by Facebook and trained using the fastText on the Common Crawl data set [47]—achieved better performance compared to other embeddings on a validation data set. Thus, we used the Common Crawl embeddings for all LSTM-CRFs models.

BERT

The BERT model is a multilayer transformer encoder model implemented using the self-attention mechanism [48], which is pretrained by combining the masked language modeling method and the next sentence prediction task. BERT has 2 versions featuring different model sizes, including a BASE version with 12 transformer layers and 110 million parameters, and a LARGE version with 24 transformer layers and 340 million parameters [26]. There are 2 steps to apply BERT for various downstream NLP, including (1) pretraining a BERT model using large unlabeled corpora and (2) fine-tuning the pretrained model using task-specific annotated corpora. In this study, we adopted the general pretrained BERT-LARGE model and fine-tuned it individually for each subtask (ie, concept extraction and relation identification) using the annotated data set developed in this challenge. We denoted the BERT-based NER model as BERT-*ner*, and the BERT-based family member attributes (ie, family role, side of family, negation, living status) classification module as BERT-*cls* and relation extraction module as BERT-*rel*.

Figure 3 illustrates the fine-tuning procedure for BERT. For token Tok_i, its input embedding and contextual representation are denoted as Emb_i and T_i. The [CLS] and [SEP] are 2 special symbols designed to format the input sequences. In this study, we also introduced a pair of entity marker including [S] and [E] to differentiate the target entity from other entities in the same sentence, where [S] indicates the start position and [E] indicates the end position. For NER (Figure 3A), the input for BERT model is a sequence of tokens, and the output is a sequence of distributed representation. Then, we used a linear layer to calculate a score for each BIO tag. Based on the entities, we developed classifiers to determine related attributes (Figure 3B). To distinguish between the target entity and other entities in the same sentence, we inserted entity markers (ie, [S] and [E]) in front of and after the target entity. For example, the input sequenced in Figure 3B contains the target entity (ie, Tok_1 and Tok_2) surrounded by the entity markers and other entities (eg, Tok_n). Then, we concatenated the representations corresponding to the [CLS] and [S] tokens and calculated a score for each predefined class label using a linear layer. For relation identification (Figure 3C), we determined the relation type based on the contextual information of 2 concepts in a relation. Therefore, the input consisted of 2 sentences linked by the special token [SEP], where each sentence contains 1 of the 2 entities in the relation. We used 2 sets of entity markers (ie, [S1], [E1], and [S2], [E2]) to label the entities. If the 2 entities of a relation are in the same sentence, then the 2 model-input sentences are the same but with different entity markers. To determine the relation category, we concatenated the representations from [CLS] and 2 start position entity makers ([S1] and [S2]) and used a linear layer to calculate a score for each predefined relation type.

Figure 3. Illustration of BERT models for (A) NER, (B) family member attributes (including side and role of family members, negation of observations, and living scores) classification, and (C) relation extraction. BERT: bidirectional encoder representations from transformers; NER: named entity recognition.



Experiments and Evaluations

In this study, we reused the LSTM-CRFs model developed in our previous study [49] and implemented the BERT-based models on top of the Transformers library [50] implemented in PyTorch [51]. We used the following parameters to initialize the LSTM-CRFs: the character embedding dimension was 25, the word embedding dimension was 100, the character-level bidirectional LSTM layer dimension was 25, the word-level bidirectional LSTM layer was 100 with a dropout probability of 0.5, the learning rate was fixed at 0.005, and the stochastic gradient descending applied a gradient clapping at [–5.0, 5.0]. The character embeddings were randomly initialized and the word embeddings were initiated using embeddings from fastText [47] (ie, containing 2 million word vectors trained on Common

Crawl). We initialized all BERT-based models using the BERT-LARGE pretrained on the general English corpus and fine-tuned them with the default model parameter settings. To train NER models, we randomly (using random seeds for reproducibility) split the original training set (99 notes) into a short training set of 89 notes and a development set of 10 notes. The best NER models were selected according to the performance on the development set. We optimized 2 hyperparameters, including the number of epochs and batch size, via fivefold cross-validation. Table 2 summarizes the optimized hyperparameters. We conducted all experiments using 2 NVIDIA P6000 graphics processing units (GPUs). We used the official evaluation script provided by the 2019 n2c2 open shared task organizers to calculate the evaluation scores on the test set. Evaluation metrics as micro-averaged precision, recall, and F1 score were used for both subtask 1 and subtask 2.

Table 2. The optimized hyperparameters of BERT-based models for various tasks.

Task	Pretrained model	Number of epochs	Batch size	Learning rate
NER[a]	BERT[b]-LARGE	30	4	1.00×10^{-05}
Negation classification	BERT-LARGE	5	8	1.00×10^{-05}
Side of family classification	BERT-LARGE	10	4	1.00×10^{-05}
Role of family classification	BERT-LARGE	5	8	1.00×10^{-05}
Living status classification	BERT-LARGE	6	8	1.00×10^{-05}
Relation identification	BERT-LARGE	12	16	2.00×10^{-05}

[a]NER: named entity recognition.

[b]BERT: bidirectional encoder representations from transformers.

Results

Table 3 compares our 4 systems for conception extraction and relation identification. Our best submission during the original challenge (LSTM-CRFs-EN + BERT-*cls* +BERT-*rel*) achieved F1 scores of 0.7944 and 0.6544 for subtask 1 and subtask 2, respectively, which is the third best system of this challenge among 17 participants. After the challenge, we further explored the BERT model for NER and the combination of BERT-*ner*-EN, BERT-*cls*, and BERT-*rel* achieved better F1 scores of 0.8249 and 0.6775 for the 2 subtasks, respectively. Compared to our best system developed during the challenge (LSTM-CRFs-EN + BERT-*cls* + BERT-*rel*), the new system (BERT-*ner*-EN + BERT-*cls* + BERT-*rel*) improved the F1 scores by 0.0305 and 0.0235 for the 2 subtasks, respectively. Our best relation identification performance was comparable to the best result reported in this challenge (0.6775 from us versus 0.6810 reported in this challenge).

Table 3. The micro-average performances for concept extraction and relation identification.a

Models	Subtask 1 (concept extraction)			Subtask 2 (relation identification)		
	Precision	Recall	F1 score	Precision	Recall	F1 score
LSTM[a]-CRFs[b] + BERT[c]-*cls* + BERT-*rel*	0.7760	0.8087	0.7920	0.7343	0.5465	0.6266
LSTM-CRFs-EN + BERT-*cls* + BERT-*rel*[d]	0.7969	0.7920	0.7944	0.6995	0.6184	0.6544
BERT-*ner* + BERT-*cls* + BERT-*rel*	0.8060	0.8105	0.8083	0.7140	0.6252[e]	0.6667
BERT-*ner*-EN + BERT-*cls* + BERT-*rel*	0.8301[e]	0.8198[e]	0.8249[e]	0.7421[e]	0.6233	0.6775[e]

[a]LSTM: long short-term memory.

[b]CRFs: conditional random fields.

[c]BERT: bidirectional encoder representations from transformers.

[d]Our best system developed during the challenge.

[e]The best performances.

Table 4 compares the detailed performance of LSTM-CRFs and BERT-*ner* for FH extraction. Compared with LSTM-CRFs, the BERT-*ner* model achieved a remarkably higher F1 score for the observation concepts (0.8094 for BERT-*ner* versus 0.7833 for LSTM-CRFs), but marginally lower performance for the family member concepts (0.8066 for BERT-*ner* versus 0.8069 for LSTM-CRFs). Table 4 also demonstrated that our ensemble strategy improved the performance of FH extraction. For example, the BERT-*ner*-EN, which was ensembled from 5 different BERT-*ner* models, outperformed the single BERT-*ner* model by about 2% for family members and about 1.5% for observations.

Table 4. A comparison of LSTM-CRFs and BERT for subtask 1 (concept extraction).

Model and concept	Precision	Recall	F1 score
LSTM-CRFs[a,b]			
Family member	0.8480	0.7686	0.8069
Observation	0.7382	0.8342	0.7833
LSTM-CRFs-EN			
Family member	0.8451	0.7868	0.8149
Observation	0.7685	0.7953	0.7817
BERT[c]-*ner*			
Family member	0.8059	0.8072	0.8066
Observation	0.8061	0.8127	0.8094
BERT-*ner*-EN			
Family member	0.8294	0.8229	0.8261
Observation	0.8306	0.8178	0.8241

[a]LSTM: long short-term memory.

[b]CRFs: conditional random fields.

[c]BERT: bidirectional encoder representations from transformers.

Table 5 compares the performance of relation identification for each relation category. Similar to the concept extraction results, the BERT-*ner*-EN + BERT-*cls* + BERT-*rel* system achieved the best F1 scores of 0.6821 and 0.6760 for the "family member—living status" and "family member—observation" relations, respectively. Compared to the LSTM-CRFs, the BERT-*ner*–based systems achieved better recalls.

Table 5. The category-level performances for subtask 2 (relation identification).

Model and relation	Precision	Recall	F1
LSTM-CRFs[a,b] + BERT[c]-*cls* + BERT-*rel*			
Family member—living status	0.7039	0.6132	0.6554
Family member—observation	0.7452	0.5269	0.6174
LSTM-CRFs-EN + BERT-*cls* + BERT-*rel*			
Family member—living status	0.6773	0.6676	0.6724
Family member—observation	0.7071	0.5993	0.6487
BERT-*ner* + BERT-*cls* + BERT-*rel*			
Family member—living status	0.6583	0.6734	0.6657
Family member—observation	0.7341	0.6111	0.6670
BERT-*ner*-EN + BERT-*cls* + BERT-*rel*			
Family member—living status	0.6912	0.6734	0.6821
Family member—observation	0.7603	0.6086	0.6760

[a]LSTM: long short-term memory.

[b]CRFs: conditional random fields.

[c]BERT: bidirectional encoder representations from transformers.

Discussion

Overview

Patients' FH is a critical risk factor associated with numerous diseases. Clinical NLP systems that automatically extract FH from clinical narrative are needed for many clinical studies and applications. The 2019 n2c2 organized shared tasks to assess current NLP methods for FH information extraction from clinical narratives. We participated in both subtasks and our system (LSTM-CRFs-EN + BERT-*cls* + BERT-*rel*) achieved the third best performance (F1 of 0.6544) among all the 21 submitted systems from 17 teams that participated in subtask 2. After the challenge, we further explored the BERT models for the concept extraction and improved our system in both concept extraction and relation identification.

Principal Findings

We observed that the BERT-*ner* model achieved both better precision (0.8060 versus 0.7760) and recall (0.8105 versus 0.8087) for clinical concept extraction compared to the LSTM-CRFs, which is consistent with a recent study by Si et al [52]. We also noticed that the single BERT-*ner* mode even achieved a higher F1 score of 0.8083 than the ensembled LSTM-CRFs model (LSTM-CRFs-EN with F1 score of 0.7944). Ensemble is an effective strategy to further improve the performance of NER. For example, the ensembled BERT model (ie, BERT-*ner*-EN, which was ensembled from 5 individual BERT-*ner* models) improved the concept extraction performance to 0.8249, compared to the single BERT model (F1 score of 0.8083). The performance improvement of the ensembled model was mainly in precision, suggesting that the ensembled models may reduce the classification errors in NER. However, further studies should examine whether our observation is related to the size of training corpus (relatively small, only 99 notes).

Most of the previous studies applied rule-based solutions to determine the family roles and family sides [34]. In this study, we adopted a pure machine learning–based solution. The experimental results showed that the BERT-based classifiers were feasible to determine the family roles, family sides, negation of observations, and living status scores. Another advantage of our method is that machine learning–based models generally have a better generalizability than rule-based systems and are easy to scale up. FH information has many variations from one patient to another, which makes the development of rules time-consuming and expensive.

In our system, we only used the sentences containing the concepts to classify the family member attributes. We also examined a strategy to include both the proceeding and following sentences. However, the experimental results based on the fivefold cross-validation on the training set showed that adding the context information did not improve the performance. One potential reason may be that most of the key information for classifying the family member attributes is located in the same sentence where the concepts (ie, family member or observation) are located. Besides, there might be potential noises brought in when including the context sentences.

A previous study [53] examined various input encoding and output representation of using BERT for relation extraction, and concluded that using representations aggregated from the start position entity markers (eg, [S1] and [S2] in Figure 3C) was the best practice. In this study, we re-evaluated 3 types of BERT output representations, including (1) the representation of the [CLS] only, (2) the representations aggregated from the start position entity markers, and (3) the representations aggregated from the [CLS] and the start position entity markers. Our results showed that option (3) led to a remarkably higher averaged F1 score (0.8975) compared to the other 2 representations (0.8851 and 0.8904). A possible reason is that the representations captured in the special token [CLS] and the representations of the start position markers contain contextual information that is complement to each other. Further studies are needed to continue examining more efficient methods for encodings and representations.

This study has limitations. First, there are limited clinical corpora for FH-related information extraction as annotating clinical notes is expensive and time-consuming. A potential solution is to use data augmentation techniques such as generative adversarial networks, which have been applied for medical imaging data [54,55]. There are preliminary research works demonstrating that generative adversarial networks could be utilized to synthesize clinical text [56]. Second, our system is a 2-stage pipeline where the errors generated in the NER will be propagated to relation extraction. We will explore potential solutions such as joint learning algorithms to alleviate this issue in our future work.

Error Analysis

Table 6 shows the confusion matrix generated for the concept extraction (subtask 1) based on our best NER model (ie, BERT-*ner*-EN). The confusion matrix showed that our system could efficiently identify family member entities. However, it is challenging for our system to differentiate the nonconcept terms for both family members and observations. For concept extraction, our system had relatively lower performances for “parent,” “grandparent,” “child,” and “siblings.” One possible reason is that the training set contains limited annotations of these entities. For example, the “parent” entity only appeared once and the “grandparent” entities appeared 6 times in the training data set. We also found that our system identified some observations not annotated in the test set. For example, in the sentence “The father also had a history of vascular surgery, a long history of smoking, and has had hip replacement,” our system extracted observations of “vascular surgery,” “smoking,” and “hip replacement,” which were annotated in the challenge corpus.

Table 6. The confusion matrix table for the NER (subtask 1).[a]

Entity type	Model prediction		
	FM[b]	OB[c]	NC[d]
FM	525	0	113
OB	0	799	178
NC	108	163	N/A[e]

[a]FM, OB, and NC are considered gold standard.

[b]FM: family members.

[c]OB: observations.

[d]NC: not a concept.

[e]N/A: not applicable.

Conclusions

Extracting patients' FH information from clinical narratives is a challenging NLP task. This study demonstrated the efficiency of deep learning–based NLP models for extraction of FH. Our system and pretrained models can be accessed at [57]. We believe our system could help other researchers to extract and leverage patient's FH documented in clinical narratives in their studies.

Acknowledgments

Research reported in this publication was supported by (1) the University of Florida Clinical and Translational Science Institute, which is supported in part by the NIH National Center for Advancing Translational Sciences under award number UL1TR001427; (2) the Patient-Centered Outcomes Research Institute (PCORI) under award number ME-2018C3-14754; (3) the Centers for Disease Control and Prevention (CDC) under award number U18DP006512; (4) the NIH National Cancer Institute (NCI) under award number R01CA246418; (5) the NIH National Institute on Aging under award number R21AG061431-02S1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health and Patient-Centered Outcomes Research Institute. We thank the n2c2 organizers for providing the annotated corpus and the guidance for this challenge. We gratefully acknowledge the support of NVIDIA Corporation for the donation of the GPUs used for this research.

Authors' Contributions

XY, JB, and YW were responsible for the overall design, development, and evaluation of this study. HZ and XH were involved in conducting experiments and result analysis. XY, JB, and YW did the writing and editing of this manuscript. All authors reviewed the manuscript critically for scientific content, and all authors gave final approval of the manuscript for publication.

Conflicts of Interest

None declared.

References

1. Scheuner MT, Wang S, Raffel LJ, Larabell SK, Rotter JI. Family history: a comprehensive genetic risk assessment method for the chronic conditions of adulthood. Am J Med Genet 1997 Aug 22;71(3):315-324. [doi: 10.1002/(sici)1096-8628(19970822)71:3<315::aid-ajmg12>3.0.co;2-n] [Medline: 9268102]
2. Rich E, Burke W, Heaton C, Haga S, Pinsky L, Short M, et al. Reconsidering the family history in primary care. J Gen Intern Med 2004 Mar;19(3):273-280 [FREE Full text] [doi: 10.1111/j.1525-1497.2004.30401.x] [Medline: 15009784]
3. Guttmacher AE, Collins FS, Carmona RH. The family history--more important than ever. N Engl J Med 2004 Nov 25;351(22):2333-2336. [doi: 10.1056/NEJMsb042979] [Medline: 15564550]
4. Harrison TA, Hindorff LA, Kim H, Wines RC, Bowen DJ, McGrath BB, et al. Family history of diabetes as a potential public health tool. Am J Prev Med 2003 Feb;24(2):152-159. [doi: 10.1016/s0749-3797(02)00588-3] [Medline: 12568821]
5. Williams RR, Hunt SC, Heiss G, Province MA, Bensen JT, Higgins M, et al. Usefulness of cardiovascular family history data for population-based preventive medicine and medical research (the Health Family Tree Study and the NHLBI Family Heart Study). Am J Cardiol 2001 Jan 15;87(2):129-135. [doi: 10.1016/s0002-9149(00)01303-5] [Medline: 11152826]
6. Ramsey SD, Yoon P, Moonesinghe R, Khoury MJ. Population-based study of the prevalence of family history of cancer: implications for cancer screening and prevention. Genet Med 2006 Oct;8(9):571-575 [FREE Full text] [doi: 10.1097/01.gim.0000237867.34011.12] [Medline: 16980813]

7. Pharoah PD, Ponder BA. The genetics of ovarian cancer. Best Pract Res Clin Obstet Gynaecol 2002 Aug;16(4):449-468. [doi: 10.1053/beog.2002.0296] [Medline: 12413928]
8. Johns LE, Houlston RS. A systematic review and meta-analysis of familial colorectal cancer risk. Am J Gastroenterol 2001 Oct;96(10):2992-3003. [doi: 10.1111/j.1572-0241.2001.04677.x] [Medline: 11693338]
9. Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. Lancet 2001 Oct 27;358(9291):1389-1399. [doi: 10.1016/S0140-6736(01)06524-2] [Medline: 11705483]
10. Porta M. A Dictionary of Epidemiology Internet. Oxford, UK: Oxford University Press; 2016. URL: https://www.oxfordreference.com/view/10.1093/acref/9780199976720.001.0001/acref-9780199976720 [accessed 2020-12-07]
11. Pharoah PD, Day NE, Duffy S, Easton DF, Ponder BA. Family history and the risk of breast cancer: a systematic review and meta-analysis. Int J Cancer 1997 May 29;71(5):800-809 [FREE Full text] [doi: 10.1002/(sici)1097-0215(19970529)71:5<800::aid-ijc18>3.0.co;2-b] [Medline: 9180149]
12. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008:128-144. [Medline: 18660887]
13. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: A literature review. J Biomed Inform 2018 Jan;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
14. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, et al. Deep learning in clinical natural language processing: a methodical review. J Am Med Inform Assoc 2020 Mar 01;27(3):457-470. [doi: 10.1093/jamia/ocz200] [Medline: 31794016]
15. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011;18(5):552-556 [FREE Full text] [doi: 10.1136/amiajnl-2011-000203] [Medline: 21685143]
16. Suominen H, Salanterä S, Velupillai S, Chapman W, Savova G, Elhadad N, et al. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: Forner P, Müller H, Paredes R, Rosso P, Stein B, editors. Information Access Evaluation. Multilinguality, Multimodality, and Visualization. Berlin, Heidelberg: Springer; 2013:212-231.
17. Kelly L, Goeuriot L, Suominen H, Schreck T, Leroy G, Mowery D, et al. Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. In: Information Access Evaluation. Multilinguality, Multimodality, and Interaction. New York City, USA: Springer; 2014:172-191.
18. 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. [doi: 10.1093/jamia/ocz166] [Medline: 31584655]
19. 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]
20. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc 2013;20(5):806-813 [FREE Full text] [doi: 10.1136/amiajnl-2013-001628] [Medline: 23564629]
21. Aronson AR, Lang F. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229-236 [FREE Full text] [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]
22. 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]
23. 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. [doi: 10.1093/jamia/ocx132] [Medline: 29186491]
24. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman S. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. arXiv 2019 [FREE Full text]
25. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural Architectures for Named Entity Recognition. arXiv 2016 [FREE Full text]
26. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv 2018 [FREE Full text]
27. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
28. Alsentzer E, Murphy J, Boag W, Weng W, Jindi D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. arXiv 2019 [FREE Full text]
29. Peng Y, Yan S, Lu Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. arXiv 2019 [FREE Full text]
30. Wang Y, Wang L, Rastegar-Mojarad M, Liu S, Shen F, Liu H. Systematic Analysis of Free-Text Family History in Electronic Health Record. AMIA Jt Summits Transl Sci Proc 2017;2017:104-113 [FREE Full text] [Medline: 28815117]
31. Goryachev S, Kim H, Zeng-Treitler Q. Identification and extraction of family history information from clinical reports. AMIA Annu Symp Proc 2008 Nov 06:247-251 [FREE Full text] [Medline: 18999129]

32. Bill R, Pakhomov S, Chen ES, Winden TJ, Carter EW, Melton GB. Automated extraction of family history information from clinical notes. AMIA Annu Symp Proc 2014;2014:1709-1717 [FREE Full text] [Medline: 25954443]
33. Liu S, Wang Y, Liu H. Selected articles from the BioCreative/OHNLP challenge 2018. BMC Med Inform Decis Mak 2019 Dec 27;19(Suppl 10):262 [FREE Full text] [doi: 10.1186/s12911-019-0994-6] [Medline: 31882003]
34. Sijia L, Majid RM, Yanshan W, Liwei W, Feichen S, Sunyang F, et al. Overview of the BioCreative/OHNLP 2018 Family History Extraction Task. 2018. URL: https://www.researchgate.net/publication/327424806_Overview_of_the_BioCreativeOHNLP_2018_Family_History_Extraction_Task [accessed 2020-12-10]
35. Shi X, Jiang D, Huang Y, Wang X, Chen Q, Yan J, et al. Family history information extraction via deep joint learning. BMC Med Inform Decis Mak 2019 Dec 27;19(Suppl 10):277 [FREE Full text] [doi: 10.1186/s12911-019-0995-5] [Medline: 31881967]
36. Brill E. Automatic Grammar Induction and Parsing Free Text: A Transformation-Based Approach. 1993 Presented at: The 31st Annual Meeting of the Association for Computational Linguistics; 1993; Columbus, OH p. 259-265 URL: https://www.aclweb.org/anthology/P93-1035/ [doi: 10.3115/981574.981609]
37. Ramshaw L, Marcus M. Text Chunking using Transformation-Based Learning. arXiv 1995 [FREE Full text]
38. Nayel H, Shashirekha H. Improving NER for Clinical Texts by Ensemble Approach using Segment Representations. 2017 Presented at: The 14th International Conference on Natural Language Processing (ICON-2017); 2017; Kolkata, West Bengal, India p. 197-204 URL: https://www.aclweb.org/anthology/W17-7525
39. 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]
40. 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]
41. 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]
42. Dai H. Family member information extraction via neural sequence labeling models with different tag schemes. BMC Med Inform Decis Mak 2019 Dec 27;19(Suppl 10):257 [FREE Full text] [doi: 10.1186/s12911-019-0996-4] [Medline: 31881965]
43. Tang B, Wu Y, Jiang M, Chen Y, Denny JC, Xu H. A hybrid system for temporal information extraction from clinical text. J Am Med Inform Assoc 2013;20(5):828-835 [FREE Full text] [doi: 10.1136/amiajnl-2013-001635] [Medline: 23571849]
44. Yang X, Bian J, Gong Y, Hogan WR, Wu Y. MADEx: A System for Detecting Medications, Adverse Drug Events, and Their Relations from Clinical Notes. Drug Saf 2019 Jan 2;42(1):123-133 [FREE Full text] [doi: 10.1007/s40264-018-0761-0] [Medline: 30600484]
45. 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]
46. Yang X, Lyu T, Li Q, Lee C, Bian J, Hogan WR, et al. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. BMC Med Inform Decis Mak 2019 Dec 05;19(Suppl 5):232 [FREE Full text] [doi: 10.1186/s12911-019-0935-4] [Medline: 31801524]
47. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. TACL 2017 Dec;5:135-146. [doi: 10.1162/tacl_a_00051]
48. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is All you Need. arXiv 2017 [FREE Full text]
49. Wu Y, Yang X, Bian J, Guo Y, Xu H, Hogan W. Combine Factual Medical Knowledge and Distributed Word Representation to Improve Clinical Named Entity Recognition. AMIA Annu Symp Proc 2018;2018:1110-1117 [FREE Full text] [Medline: 30815153]
50. 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 2020 [FREE Full text]
51. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv 2019 [FREE Full text]
52. Si Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. J Am Med Inform Assoc 2019 Nov 01;26(11):1297-1304. [doi: 10.1093/jamia/ocz096] [Medline: 31265066]
53. Soares L, FitzGerald N, Ling J, Kwiatkowski T. Matching the Blanks: Distributional Similarity for Relation Learning. arXiv 2019 [FREE Full text]
54. Frangi AF, Tsaftaris SA, Prince JL. Simulation and Synthesis in Medical Imaging. IEEE Trans Med Imaging 2018 Mar;37(3):673-679. [doi: 10.1109/tmi.2018.2800298]
55. Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: A review. Med Image Anal 2019 Dec;58:101552. [doi: 10.1016/j.media.2019.101552] [Medline: 31521965]
56. Guan J, Li R, Yu S, Zhang X. Generation of Synthetic Electronic Medical Record Text. 2018 Presented at: IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2018; Madrid, Spain p. 374-380. [doi: 10.1109/BIBM.2018.8621223]

57. Yang X, He X, Zhang H, Bian J, Wu Y. UF team in 2019 N2C2 challenge Track2 Family history extraction from clinical narratives. 2019. URL: https://github.com/uf-hobi-informatics-lab/2019_N2C2_Track2_FHextraction.git [accessed 2020-12-07]

Abbreviations

BERT: bidirectional encoder representations from transformers
BIO: beginning-inside-outside
CRFs: conditional random fields
FH: family history
LSTM: long short-term memory
n2c2: National NLP Clinical Challenge
NER: named entity recognition
NLP: natural language processing
GPU: graphics processing unit

Edited by Y Wang, F Shen; submitted 28.07.20; peer-reviewed by F Liu, M Huang, M Torii; comments to author 22.09.20; revised version received 05.10.20; accepted 20.11.20; published 15.12.20.

Please cite as:
Yang X, Zhang H, He X, Bian J, Wu Y
Extracting Family History of Patients From Clinical Narratives: Exploring an End-to-End Solution With Deep Learning Models
JMIR Med Inform 2020;8(12):e22982
URL: http://medinform.jmir.org/2020/12/e22982/
doi:10.2196/22982
PMID:33320104

Original Paper

Using Character-Level and Entity-Level Representations to Enhance Bidirectional Encoder Representation From Transformers-Based Clinical Semantic Textual Similarity Model: ClinicalSTS Modeling Study

Ying Xiong[1], PhD; Shuai Chen[1], MS; Qingcai Chen[1,2], PhD, Prof Dr; Jun Yan[3], PhD; Buzhou Tang[1,2], PhD, Prof Dr

[1]Harbin Institute of Technology, Shenzhen, China

[2]Peng Cheng Laboratory, Shenzhen, China

[3]Yidu Cloud Technology Company Limited, Beijing, China

Corresponding Author:
Buzhou Tang, PhD, Prof Dr
Harbin Institute of Technology
HIT Campus, Xili University Town
Shenzhen, 518055
China
Phone: 86 075526033182
Email: tangbuzhou@gmail.com

Abstract

Background: With the popularity of electronic health records (EHRs), the quality of health care has been improved. However, there are also some problems caused by EHRs, such as the growing use of copy-and-paste and templates, resulting in EHRs of low quality in content. In order to minimize data redundancy in different documents, Harvard Medical School and Mayo Clinic organized a national natural language processing (NLP) clinical challenge (n2c2) on clinical semantic textual similarity (ClinicalSTS) in 2019. The task of this challenge is to compute the semantic similarity among clinical text snippets.

Objective: In this study, we aim to investigate novel methods to model ClinicalSTS and analyze the results.

Methods: We propose a semantically enhanced text matching model for the 2019 n2c2/Open Health NLP (OHNLP) challenge on ClinicalSTS. The model includes 3 representation modules to encode clinical text snippet pairs at different levels: (1) character-level representation module based on convolutional neural network (CNN) to tackle the out-of-vocabulary problem in NLP; (2) sentence-level representation module that adopts a pretrained language model bidirectional encoder representation from transformers (BERT) to encode clinical text snippet pairs; and (3) entity-level representation module to model clinical entity information in clinical text snippets. In the case of entity-level representation, we compare 2 methods. One encodes entities by the entity-type label sequence corresponding to text snippet (called entity I), whereas the other encodes entities by their representation in MeSH, a knowledge graph in the medical domain (called entity II).

Results: We conduct experiments on the ClinicalSTS corpus of the 2019 n2c2/OHNLP challenge for model performance evaluation. The model only using BERT for text snippet pair encoding achieved a Pearson correlation coefficient (PCC) of 0.848. When character-level representation and entity-level representation are individually added into our model, the PCC increased to 0.857 and 0.854 (entity I)/0.859 (entity II), respectively. When both character-level representation and entity-level representation are added into our model, the PCC further increased to 0.861 (entity I) and 0.868 (entity II).

Conclusions: Experimental results show that both character-level information and entity-level information can effectively enhance the BERT-based STS model.

(JMIR Med Inform 2020;8(12):e23357) doi:10.2196/23357

KEYWORDS

natural language processing; deep learning; clinical semantic textual similarity; knowledge graph

Introduction

Background

Electronic health record (EHR) systems have been widely used in hospitals all over the world for convenience to health information storage, share, and exchange [1]. In recent years, EHRs have become a key data source for medical research and clinical decision support. Therefore, the quality of EHRs is crucial. However, copy-and-paste and templates are very common in EHR writing [2,3], resulting in EHRs of low quality in content. How to detect copy-and-paste and templates in different documents has become increasingly important for the secondary use of EHRs. This can be regarded as a clinical semantic textual similarity (ClinicalSTS) task, which is also applied to clinical decision support, trial recruitment, tailored care, clinical research [4-6], and medical information services, such as clinical question answering [7,8] and document classification [9].

In the past few years, some shared tasks on STS, such as Semantic Evaluation (SemEval), have been launched by different organizers [10-14]. These shared tasks mainly focus on general domains, including newswire, tutorial dialog system, Wikipedia, among others. There has been almost no study on STS in the clinical domain. To boost the development of ClinicalSTS, Wang et al [15] constructed a clinical STS corpus of 174,629 clinical text snippet pairs from Mayo Clinic. Based on a part of this corpus, BioCreative/OHNLP organizers held the first ClinicalSTS shared pilot task (challenge) in 2018 [16]. A corpus of 1068 clinical text snippet pairs with similarity ranging from 0 to 5 was provided for this shared task. In 2019, the n2c2/OHNLP organizers extended the 2018 shared task corpus and continued to hold ClinicalSTS shared task [17]. The extended corpus is composed of 2055 clinical text snippet pairs.

In this paper, we introduce our system developed for the 2019 n2c2/OHNLP shared task on ClinicalSTS. The system is based on bidirectional encoder representation from transformers (BERT) [18] and includes the 2 other types of representations besides BERT: (1) character-level representation to tackle the out-of-vocabulary (OOV) problem in natural language processing (NLP) and (2) entity-level representation to model clinical entity information in clinical text snippets. In the case of entity-level representation, we apply 2 entity-level representations: one encodes entities in a text snippet by the corresponding entity label sequence (called entity I) and the other one encodes entities with their representation on Mesh [19] (called entity II). Our system achieves the highest Pearson correlation coefficient (PCC) of 0.868 on the corpus of the 2019 n2c2/OHNLP track on ClinicalSTS, which is competitive with other state-of-the-art systems.

Related Work

A model for STS usually consists of 2 modules: a module to encode text snippet (or sentence) pairs and a module for prediction (classification or regression). According to sentence pair encoding, STS models can be classified into the following 2 categories: sentence encoding models and sentence pair interaction models. The sentence encoding models first use Siamese neural network to individually encode 2 sentences with 2 neural networks of the same structure and shared parameters [20-23], then combine the 2 sentences' representation through concatenation, element-wise product, or element-wise difference operations, and finally make a classification or regression prediction via a specific layer such as multilayer perceptron (MLP) [24]. The main limitation of the sentence pair encoding models is that they ignore word-level interactions. The sentence pair interaction models adopt matching-aggregation architectures to encode word-level interactions [25,26]. These models first build an interaction matrix and then use a convolutional neural network (CNN) [27] and long short-term memory [28] with attention mechanism [29,30] and hierarchical architecture [31] to obtain aggregated matching representation for final prediction.

In recent years, pretrained language models good at capturing sentence-level semantic information, such as BERT [18], XLNet [32], RoBERTa [33], have been proved to significantly improve downstream tasks. However, most pretrained language models are at the token level. In order to tackle the inherent OOV problem of NLP, character-level representation is also considered in various NLP tasks, such as named entity recognition [34-36] and entity normalization [37], and brings improvements. Besides, researchers have started investigating how to use entity-level representation in NLP tasks [38,39].

Methods

Data Set

The n2c2/OHNLP organizers manually annotated a total of 2055 clinical text snippet pairs by 2 medical experts for the ClinicalSTS task, where 1643 pairs are used as the training set and 412 as the test set. The similarity of each clinical text snippet pair is measured by PCC ranging from 0 to 5, where 0 means that 2 clinical text snippets are absolutely different, and 5 means that 2 clinical text snippets are entirely semantically equal. All clinical text snippets are selected from deidentified EHRs. Table 1 gives examples of each score.

Table 1. Examples of ClinicalSTS[a].

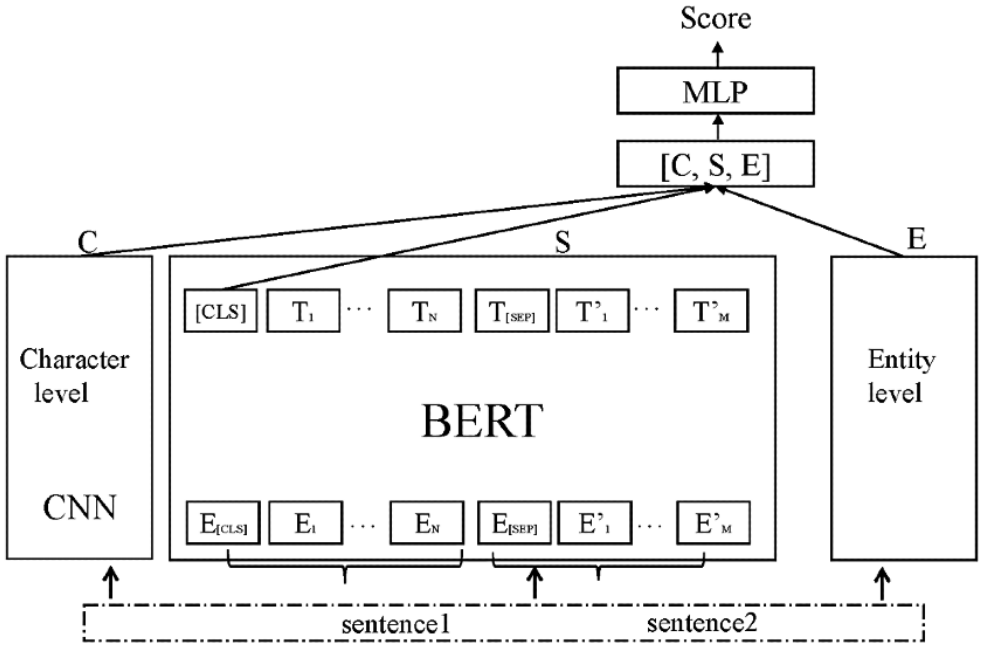
Score	Example of clinical text snippet pair
0	**The 2 sentences are completely dissimilar**
	S1: The patient has missed 0 hours of work in the past seven days for issues not related to depression.
	S2: In the past year the patient has the following number of visits: none in the hospital none in the er and one as an outpatient.
1	**The 2 sentences are not equivalent but have the same topic**
	S1: There is no lower extremity edema present bilaterally.
	S2: There is a 2+ radial pulse present in the upper extremities bilaterally.
2	**The 2 sentences are not equivalent but share some details**
	S1: I met with the charge nurse and reviewed the patient's clinical condition.
	S2: I have reviewed the relevant imaging and medical record.
3	**The 2 sentences are roughly equivalent but some important information differs**
	S1: I explained the diagnosis and treatment plan in detail, and the patient clearly expressed understanding of the content reviewed.
	S2: Began discussion of diagnosis and treatment of chronic pain and chronic fatigue; patient expressed understanding of the content.
4	**The 2 sentences are mostly equivalent and only a little detail is different**
	S1: Albuterol [PROVENTIL/VENTOLIN] 90 mcg/Act HFA Aerosol 2 puffs by inhalation every 4 hours as needed.
	S2: Albuterol [PROVENTIL/VENTOLIN] 90 mcg/Act HFA Aerosol 1-2 puffs by inhalation every 4 hours as needed #1 each.
5	**The 2 sentences mean the same thing, they are absolutely equivalent**
	S1: Goals/Outcomes: Patient will be instructed in a home program, demonstrate understanding, and state the ability to continue independently.
	S2: Patient will be instructed in home program, demonstrate understanding, and state ability to continue independently-ongoing.

[a]ClinicalSTS: clinical semantic textual similarity.

Models

Figure 1 presents an overview architecture of our model. In this model, we first use 3 representation modules at different levels to encode input text snippet pairs, that is, character-level, sentence-level, and entity-level representation modules, and then feed them to MLP for prediction.

Figure 1. Overview architecture of our model for the ClinicalSTS track of the 2019 n2c2/OHNLP challenge. BERT: bidirectional encoder representation from transformers; ClinicalSTS: clinical semantic textual similarity; CNN: convolutional neural network; MLP: multilayer perceptron; PCC: Pearson correlation coefficient; [CLS]: the representation of sentence pair with BERT.



Character-Level Representation

In order to tackle the OOV problem in NLP, following [34-37], given a pair of clinical text snippets (a, b), we first apply character-level CNN on each token to obtain its character-level representation, and then apply max pooling operation on all tokens in a and b to obtain the character-level representation of (a, b), denoted by C. We model the character-level representation with CNN, because there is no significant difference in using CNN and long short-term memory, according to previous studies [40,41].

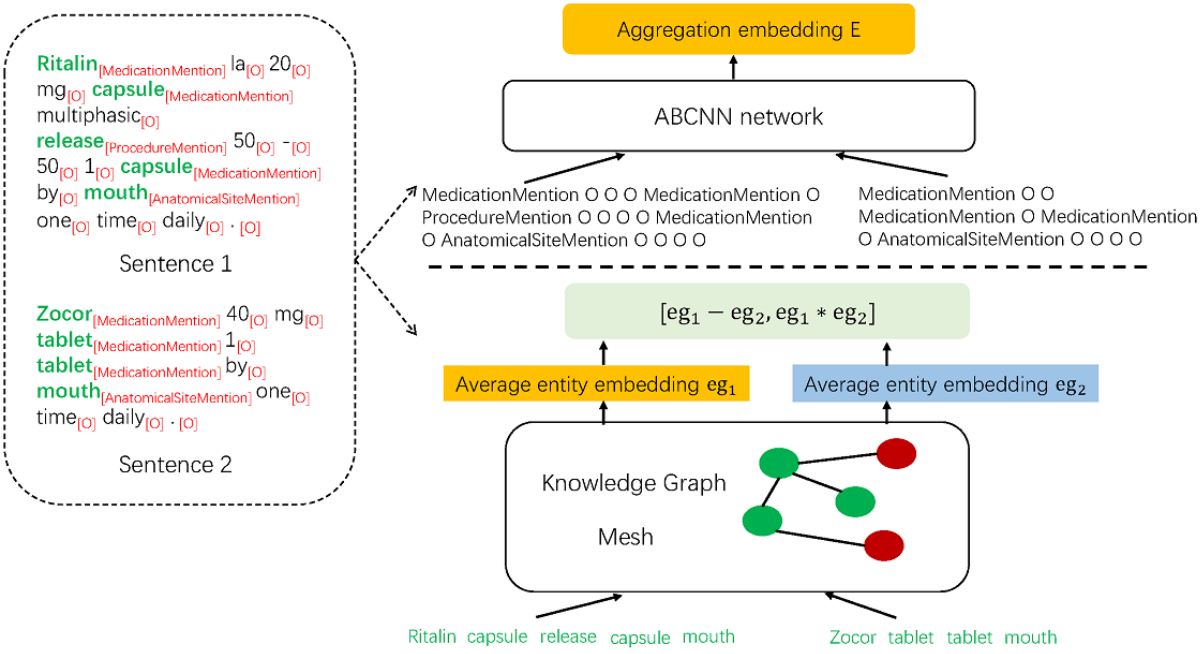
Sentence-Level Representation

We use BERT to encode the input clinical text snippet pair (a, b) and obtain its sentence-level representation, denoted by S = BERT(a, b).

Entity-Level Representation

We first deploy cTAKES [42], a popular clinical NLP tool, to extract entity mentions from text snippets, and then propose 2 methods to obtain the entity-level representations of the text snippets according to the extracted entity mentions, as shown in Figure 2. cTAKES can extract 9 kinds of entities: AnatomicalSiteMention, DiseaseDisorderMention, FractionAnnotation, MedicationMention, Predicate, ProcedureMention, RomanNumeralAnnotation, SignSymptomMention, and Temporal Information.

Figure 2. Entity-level representation.



In the first method for entity-level representation (entity I), we convert text snippet a and b into entity-type sequences corresponding to them, and then deploy attention-based CNN [27] on the pair of the entity-type sequences in the following way:

$$E = BCNN(es_a, es_b) \quad (1)$$

where es_a is the entity label sequence of text snippet a, es_b is the entity label sequence of text snippet b, BCNN is basic bi-CNN, and E is the entity-level representation of (es_a, es_b). For example, given a text snippet b "Zocor 40 mg tablet 1 tablet by mouth one time daily." shown in Figure 2, cTAKES first extracts 3 medication mentions {"Zocor", "tablet", "tablet"} and 1 anatomical mention {"mouth"}, and then we obtain the entity-type sequence corresponding to text snippet b: "MedicationMetion O O MedicationMetion O MedicationMetion O AnatomicalSiteMention O O O O". In this entity-type sequence, "O" stands for "Other."

The second method for entity-level representation (entity II) first directly adopts entity representation learned by TransE [43] on an external knowledge graph (KG; Mesh in this study), and then applies average pooling operation on all entities individually in sentences a and b to get entity-level representations of a (denoted by eg_a) and b (denoted by eg_b) respectively, and finally aggregates their representations using equation 2.

$$E = \tanh(W_e[eg_a - eg_b; eg_a * eg_b] + b_e) \quad (2)$$

where "[;]" denotes concatenation operation, W_e is a weight matrix, and b_e is a bias vector.

MLP Layer

To aggregate the information of 3 modules, we concatenate them together:

$$f = [S; C; E] \quad (3)$$

Then, we use MLP (as shown in equation 4) to predict the STS score p_{score} of (a, b) as follows:

$$p_{score} = MLP(Wf + b) \quad (4)$$

where W is a weight matrix, and b is a bias vector.

The loss function used in our model is the minimum square error (MSE) function:

$$Loss = MSE(p_{score} - g_{score}) \quad (5)$$

where g_{score} is the gold-standard score.

Experimental Setting

Before conducting experiments, we preprocess the corpus using the following simple rules: (1) convert clinical text snippets into lowercase; (2) tokenize clinical text snippets using special symbols, such as "[", "]", "/", ",", and ".", and keep them unstained in some situations such as "." in decimals. The hyperparameters of our model are shown in Table 2. Other parameters are optimized via fivefold cross validation on the training set. The pretrained BERT model used for text snippet pair representation in our experiments is [BERT-Base, Uncased] [44]. We train all model parameters simultaneously, set epochs as 12, and save the last checkpoints as the final models. The performance of all models is measured by PCC.

Table 2. Hyperparameters setting of our model.

Parameters	Value
Learning rate	2×10^{-5}
Sequence length of BERT[a]	380
Epochs	12
Batch size	20
Knowledge graph embedding dimension d	100
Character-level kernel size	3
Convolution kernels of BCNN[b]	50
Kernel size of BCNN	3
Word embedding dimension of entity I	50

[a]BERT: bidirectional encoder representation from transformers.

[b]BCNN: Basic bi-CNN.

Results

Table 3 shows the overall results of our proposed model. Our model achieves the highest PCC of 0.868, which is competitive with other state-of-the-art models proposed for the 2019 n2c2/OHNLP track on ClinicalSTS. The model using entity II is better than that using entity I by 0.007 in PCC, indicating that entity II is a better supplement to BERT than entity I. When character-level representation is removed, the PCC of our model decreases to 0.859 (entity I) and 0.854 (entity II). When entity-level representation is removed, the PCC of our model decreases to 0.858. When both types of representations are removed, the PCC of our model further decreases to 0.848. The results indicate that both character-level representation and entity-level representation are supplementary to BERT. Although the improvements individually from entity I and character-level text snippet representation are more remarkable than entity II, the improvement from the combination of entity I and character-level representation is much smaller than the combination of entity II and character-level representation. It is because both character-level representation and entity I come from text snippets, whereas entity II comes from external KG. The diversity between character-level representation and entity II is much larger than that between character-level representation and entity I. It is interesting that our model is not further improved when both entity I and entity II are considered in our model at the same time, which may be also because of the diversity.

Moreover, we investigate the effect of the domain-specific pretrained BERT models [45,46] on our model. We replace the pretrained BERT model in the general domain, [BERT-Base, Uncased] [44], by the pretrained BERT model in the clinical domain [45] to obtain a new model. The highest PCC of the new model is 0.872, which is slightly better than our previous model, indicating that the domain-specific pretrained BERT model is beneficial to our model.

Table 3. Pearson correlation coefficient of our model on the test set.

Model and setting	PCC[a]
Our model	
Entity I	0.861
Entity II	0.868[b]
Entity I + Entity II	0.862
Without character -level text snippet representation	
Entity I	0.859
Entity II	0.854
Without entity-level representation	0.858
Without both	0.848

[a]PCC: Pearson correlation coefficient.

[b]The highest PCC.

Discussion

Error Analysis

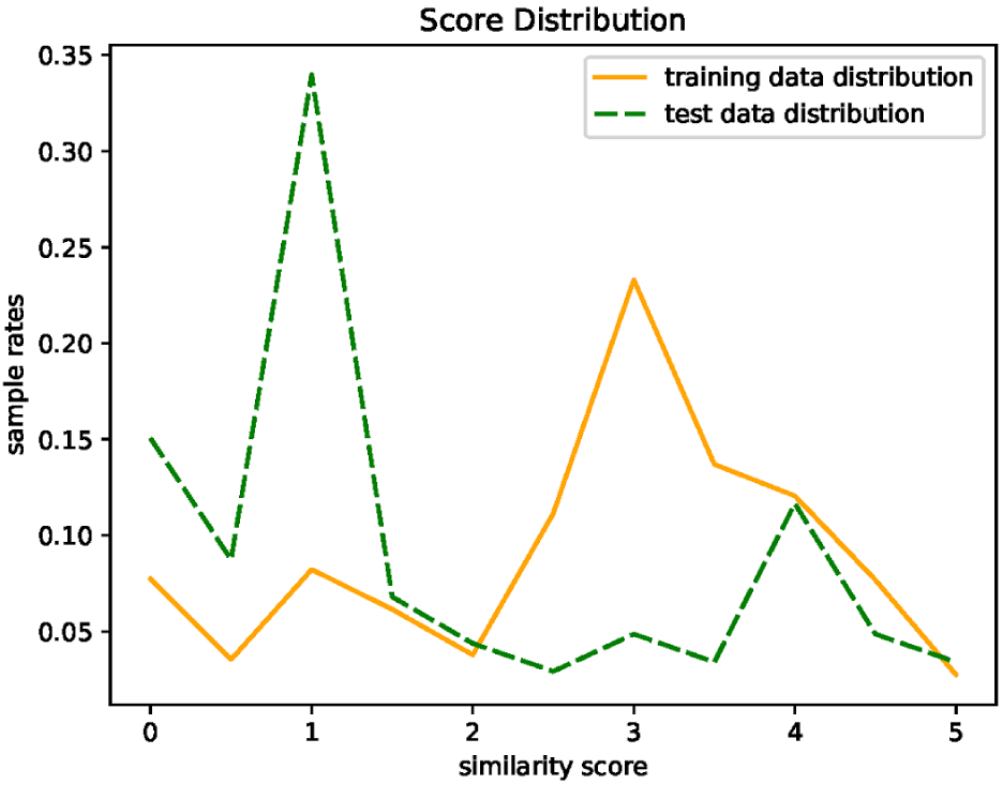
Although the proposed model achieves competitive performance, there are also some errors. To analyze these errors, we look into samples for which the difference between the predicted STS score and gold-standard similarity score is greater than 1.0 and find that the main errors can be classified into 2 types.

The first type of error is related to polarity of clinical text snippets as our model is insensitive to positive and negative words. For example, as shown in Table 4, because both clinical text snippets in example 1 depict coughing up, their STS score predicted by our model is 2.5, but their gold-standard STS score is 1.0 as the polarity of the first text snippet is positive, whereas that of the second text snippet is negative. The second type of error is related to prescriptions that include medication names, usages, and dosages. For example, the gold-standard STS score of example 2 in Table 4 is 1.0 as the medications in the 2 text snippets are completely different, but the STS score of the example predicted by our model is 2.5 as some other words are the same in the 2 text snippets. Just because our model cannot extract medical information comprehensively, there are lots of errors of the second type. For further improvement, we need a comprehensive information extraction module to extract polarity information and medications with usage and dosage attributes besides the current 9 kinds of clinical entities. A possible way is to integrate the existing tools specifically for polarity information extraction (such as SenticNet [47]) or medication extraction (such as MedEx [48]) into our model. We also find that the scores of mispredictions are close to 2.5, which may be caused by the different STS score distributions of the training and test sets. As shown in Figure 3, the STS scores of most sentence pairs in the training set concentrate in [2.5, 3.5], whereas those in the test set concentrate in [0.5, 1.5]. The difference is remarkable. It is reasonable to obtain the STS scores of mispredictions around the average score of the training set.

Table 4. Examples of errors on the test set.

Number	Example
1	• *Sentence 1:respiratory: positive for coughing up mucus (phlegm), dyspnea and wheezing.* • *Sentence 2: negative for coughing up blood and dry cough.* • Gold-standard: 1.0 • Predicted: 2.5
2	• *Sentence 1: ibuprofen [motrin] 800 mg tablet 1 tablet by mouth four time a day as needed.* • *Sentence 2: lisinopril 10 mg tablet 1 tablet by mouth one time daily.* • Gold-standard: 1.0 • Predict: 2.4

Figure 3. Similarity interval distribution in the training and test data sets.



Effect of Entity-Level Representation

Although the results in Table 3 show that any one of the 2 entity-level representations enhances the BERT-based model, some limitations also exist. In the case of entity I, we only consider type semantic information, but no entity semantic information. In the case of entity II, only about 20% (220/1080) of clinical entities recognized by cTAKES [42] can be mapped to Mesh via dictionary look-up. There are 2 directions for improvement: (1) introduce entity semantic information into entity I, and (2) improve entity mapping performance in entity II and find a larger KG instead of Mesh.

Conclusions

In this paper, we propose an enhanced BERT-based model for ClinicalSTS by introducing a character-level representation and an entity-level representation. Experiments on the 2019 n2c2/OHNLP track on ClinicalSTS in 2019 indicate that both the character-level representation and the entity-level representation can enhance the BERT-based ClinicalSTS model, and our enhanced BERT-based model achieves competitive performance with other state-of-the-art models. In addition, domain-specific pretrained BERT models are better than general pretrained BERT models.

Acknowledgments

This paper is supported in part by grants: National Natural Science Foundations of China (U1813215, 61876052, and 61573118), Special Foundation for Technology Research Program of Guangdong Province (2015B010131010), National Natural Science Foundations of Guangdong, China (2019A1515011158), Guangdong Province Covid-19 Pandemic Control Research Fund (2020KZDZX1222), Strategic Emerging Industry Development Special Funds of Shenzhen (JCYJ20180306172232154 and JCYJ20170307150528934), and Innovation Fund of Harbin Institute of Technology (HIT.NSRIF.2017052).

Conflicts of Interest

None declared.

References

1. Evans R. Electronic Health Records: Then, Now, and in the Future. Yearb Med Inform 2016 May 20;Suppl 1:S48-S61 [FREE Full text] [doi: 10.15265/IYS-2016-s006] [Medline: 27199197]
2. Markel A. Copy and paste of electronic health records: a modern medical illness. Am J Med 2010 May;123(5):e9. [doi: 10.1016/j.amjmed.2009.10.012] [Medline: 20399309]

3. Kettl PA. A Piece of My Mind. JAMA 1992 Feb 12;267(6):798. [doi: 10.1001/jama.1992.03480060040014]
4. Wu H, Toti G, Morley K, Ibrahim Z, Folarin A, Jackson R, et al. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. J Am Med Inform Assoc 2018 May 01;25(5):530-537 [FREE Full text] [doi: 10.1093/jamia/ocx160] [Medline: 29361077]
5. Hanauer DA, Wu DT, Yang L, Mei Q, Murkowski-Steffy KB, Vydiswaran VV, et al. Development and empirical user-centered evaluation of semantically-based query recommendation for an electronic health record search engine. J Biomed Inform 2017 Mar;67:1-10 [FREE Full text] [doi: 10.1016/j.jbi.2017.01.013] [Medline: 28131722]
6. Plaza L, Díaz A. Retrieval of Similar Electronic Health Records Using UMLS Concept Graphs. In: Hopfe CJ, Rezgui Y, Métais E, Preece A, Li H, editors. Natural Language Processing and Information Systems. NLDB 2010. Lecture Notes in Computer Science, vol 6177. Berlin, Germany: Springer; 2010:293-303.
7. Cao Y, Liu F, Simpson P, Antieau L, Bennett A, Cimino J, et al. AskHERMES: An online question answering system for complex clinical questions. J Biomed Inform 2011 Apr;44(2):277-288 [FREE Full text] [doi: 10.1016/j.jbi.2011.01.004] [Medline: 21256977]
8. Demner-Fushman D, Lin J. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics Association for Computational Linguistics. 2006 Jul Presented at: 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics Association for Computational Linguistics; 2006; Sydney, Australia p. 841-848 URL: https://dl.acm.org/doi/10.3115/1220175.1220281 [doi: 10.3115/1220175.1220281]
9. Stubbs A, Filannino M, Soysal E, Henry S, Uzuner Ö. Cohort selection for clinical trials: n2c2 2018 shared task track 1. J Am Med Inform Assoc 2019 Nov 01;26(11):1163-1171 [FREE Full text] [doi: 10.1093/jamia/ocz163] [Medline: 31562516]
10. Agirre E, Cer D, Diab M, Gonzalez-Agirre A. Semeval-2012 task 6: A pilot on semantic textual similarity. In: SemEval '12: Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation. New York, NY: ACM; 2012 Jun 7 Presented at: SemEval '12; 2012; Montréal, Canada p. 385-393 URL: https://dl.acm.org/doi/10.5555/2387636.2387697 [doi: 10.18653/v1/s17-2001]
11. Agirre E, Cer D, Diab M, Gonzalez-Agirre A, Guo W. SEM 2013 shared task: Semantic Textual Similarity. 2013 Jun 13 Presented at: Human Language Technology: Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL); 2013; Atlanta, GA p. 32-43. [doi: 10.18653/v1/s17-2001]
12. Agirre E, Banea C, Cardie C, Cer D, Diab M, Gonzalez-Agirre A, et al. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. 2014 Aug 23 Presented at: International Conference on Computational Linguistics (COLING); August 23-24, 2014; Dublin, Ireland p. 81-91. [doi: 10.3115/v1/s14-2010]
13. Agirre E, Banea C, Cardie C, Cer D, Diab M, Gonzalez-Agirre A, et al. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. 2015 Jun 4 Presented at: Human Language Technology: Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL); June 4-5, 2015; Denver, CO p. 252-263. [doi: 10.18653/v1/S15-2045]
14. Agirre E, Banea C, Cer D, Diab M, González-Agirre A, Mihalcea R, et al. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. 2016 Jun 2 Presented at: Human Language Technology: Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL); June 16-17, 2016; San Diego, CA p. 497-511. [doi: 10.18653/v1/s16-1081]
15. 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]
16. 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 Presented at: BioCreative/OHNLP Challenge 2018; December, 2018; Washington, DC. [doi: 10.1145/3233547.3233672]
17. Wang Y, Fu S, Shen F, Henry S, Uzuner O, Liu H. The 2019 n2c2/OHNLP Track on Clinical Semantic Textual Similarity: Overview. JMIR Med Inform 2020 Nov 27;8(11):e23375 [FREE Full text] [doi: 10.2196/23375] [Medline: 33245291]
18. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019 Presented at: Human Language Technology: Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL); 2019; Minneapolis, MN p. 4171-4186. [doi: 10.18653/v1/n19-1423]
19. Lipscomb C. Medical Subject Headings (MeSH). Bull Med Libr Assoc 2000 Jul;88(3):265-266 [FREE Full text] [Medline: 10928714]
20. Mueller J, Thyagarajan A. Siamese Recurrent Architectures for Learning Sentence Similarity. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press; 2016 Presented at: Thirtieth AAAI Conference on Artificial Intelligence; February 12-17, 2016; Phoenix, AZ p. 2786-2792.
21. Neculoiu P, Versteegh M, Rotaru M. Learning text similarity with siamese recurrent networks. 2016 Aug 11 Presented at: 5th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2020; August 11, 2016; Berlin, Germany p. 148-157. [doi: 10.18653/v1/w16-1617]

22. Hu B, Lu Z, Li H, Chen Q. Convolutional Neural Network Architectures for Matching Natural Language Sentences. 2014 Dec 8 Presented at: Neural Information Processing Systems (NeurIPS); December 8-13, 2014; Montreal, Quebec, Canada p. 2042-2050 URL: http://papers.nips.cc/paper/5550-convolutional-neural-network-architectures-for-matching-natural-language-sentences.pdf
23. Wang K, Yang B, Xu G, He X. Medical Question Retrieval Based on Siamese Neural Network Transfer Learning Method. In: Database Systems for Advanced Applications. Cham, Switzerland: Springer International Publishing; Apr 24, 2019:49-64.
24. Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A. Supervised learning of universal sentence representations from natural language inference data. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.: Copenhagen, Denmark; 2017 Sep Presented at: 2017 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics; Stroudsburg, PA. [doi: 10.18653/v1/d17-1070]
25. He H, Gimpel K, Lin J. Multi-Perspective Sentence Similarity Modeling with Convolutional Neural Networks. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015. Stroudsburg, PA: Association for Computational Linguistics; 2015 Sep 17 Presented at: Conference on Empirical Methods in Natural Language Processing (EMNLP 2015); September 17-21, 2015; Lisbon, Portugal p. 1576-1586. [doi: 10.18653/v1/d15-1181]
26. Kim S, Kang I, Kwak N. Semantic Sentence Matching with Densely-Connected Recurrent and Co-Attentive Information. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2019 Jul 17 Presented at: AAAI Conference on Artificial Intelligence; January 27 to February 1, 2019; Honolulu, HI p. 6586-6593. [doi: 10.1609/aaai.v33i01.33016586]
27. Yin W, Schütze H, Xiang B, Zhou B. ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs. TACL 2016 Dec;4:259-272. [doi: 10.1162/tacl_a_00097]
28. Wang Z, Hamza W, Florian R. Bilateral Multi-Perspective Matching for Natural Language Sentences. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI). 2017 Aug 19 Presented at: Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI); August 19-25, 2017; Melbourne, Australia p. 4144-4150. [doi: 10.24963/ijcai.2017/579]
29. He H, Lin J. Pairwise Word Interaction Modeling with Deep Neural Networks for Semantic Similarity Measurement. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics; 2016 Jun 12 Presented at: Human Language Technology: Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL); June 12-17, 2016; San Diego CA p. 937-948. [doi: 10.18653/v1/n16-1108]
30. Tan C, Wei F, Wang W, Lv W, Zhou M. Multiway Attention Networks for Modeling Sentence Pairs. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. 2018 Jul 13 Presented at: Twenty-Seventh International Joint Conference on Artificial Intelligence; July 13-19, 2018; Stockholm, Sweden p. 4411-4417. [doi: 10.24963/ijcai.2018/613]
31. Gong Y, Luo H, Zhang J. Natural Language Inference over Interaction Space. 2018 Apr 30 Presented at: 6th International Conference on Learning Representations, ICLR 2018; May 3, 2018; Vancouver, BC, Canada URL: https://openreview.net/forum?id=r1dHXnH6-
32. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le Q. XLNet: Generalized Autoregressive Pretraining for Language Understanding. 2019 Dec 8 Presented at: Neural Information Processing Systems (NeurIPS), 2019; December 8-14, 2019; Vancouver, BC, Canada p. 5754-5764 URL: http://papers.nips.cc/paper/8812-xlnet-generalized-autoregressive-pretraining-for-language-understanding.pdf
33. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019 Jul 26. URL: https://arxiv.org/abs/1907.11692 [accessed 2019-07-26]
34. Liu Z, Chen Y, Tang B, Wang X, Chen Q, Li H, et al. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. J Biomed Inform 2015 Dec;58 Suppl:S47-S52 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.009] [Medline: 26122526]
35. Xiong Y, Shen Y, Huang Y, Chen S, Tang B, Wang X, et al. A Deep Learning-Based System for PharmaCoNER. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, BioNLP-OST@EMNLP-IJNCLP 2019. Stroudsburg, PA: Association for Computational Linguistics; 2019 Dec 4 Presented at: 5th Workshop on BioNLP Open Shared Tasks, BioNLP-OST@EMNLP-IJNCLP 2019; November 4, 2019; Hong Kong, China p. 33-37. [doi: 10.18653/v1/d19-5706]
36. Dong C, Zhang J, Zong C, Hattori M, Di H. Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition. Cham, Switzerland: Springer International Publishing; 2016 Dec 2 Presented at: Natural Language Understanding and Intelligent Applications - 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016; December 2-6, 2016; Kunming, China p. 239-250. [doi: 10.1007/978-3-319-50496-4_20]
37. Niu J, Yang Y, Zhang S, Sun Z, Zhang W. Multi-task Character-Level Attentional Networks for Medical Concept Normalization. Neural Process Lett 2018 Jun 18;49(3):1239-1256. [doi: 10.1007/s11063-018-9873-x]
38. Clark K, Manning C. Improving Coreference Resolution by Learning Entity-Level Distributed Representations. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016. Stroudsburg, PA: The Association for Computer Linguistics; 2016 Aug 7 Presented at: 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016; August 7-12, 2016; Berlin, Germany. [doi: 10.18653/v1/p16-1061]

39. Wu T, Wang Y, Wang Y, Zhao E, Yuan Y, Yang Z. Representation Learning of EHR Data via Graph-Based Medical Entity Embedding. arXiv. 2019 Oct 7. URL: https://arxiv.org/abs/1910.02574 [accessed 2019-10-07]
40. Yang J, Liang S, Zhang Y. Design Challenges and Misconceptions in Neural Sequence Labeling. In: Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018. 2018 Aug 20 Presented at: 27th International Conference on Computational Linguistics, COLING 2018; August 20-26, 2018; Santa Fe, NM p. 3879-3889 URL: https://www.aclweb.org/anthology/C18-1327/
41. 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]
42. Apache cTAKESTM - clinical Text Analysis Knowledge Extraction System. URL: https://ctakes.apache.org/ [accessed 2020-03-22]
43. Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating Embeddings for Modeling Multi-relational Data. In: Advances in Neural Information Processing Systems. 2013 Dec 5 Presented at: 27th Annual Conference on Neural Information Processing Systems 2013; December 5-8, 2013; Lake Tahoe, NV p. 2787-2795 URL: http://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data.pdf
44. Google Research: BERT. 2020. URL: https://github.com/google-research/bert [accessed 2020-08-06]
45. Peng Y, Yan S, Lu Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In: Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019. 2019 Aug 1 Presented at: 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019; August 1, 2019; Florence, Italy p. 58-65. [doi: 10.18653/v1/w19-5006]
46. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
47. Malheiros Y. senticnet: Access SenticNet data using Python Internet. URL: https://github.com/yurimalheiros/senticnetapi [accessed 2020-12-16]
48. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. Journal of the American Medical Informatics Association 2010 Jan 01;17(1):19-24. [doi: 10.1197/jamia.m3378]

Abbreviations

BERT: bidirectional encoder representation from transformers
ClinicalSTS: clinical semantic textual similarity
CNN: convolutional neural network
EHR: electronic health record
KG: knowledge graph
MLP: multilayer perceptron
NLP: natural language processing
OHNLP: Open Health Natural Language Processing
OOV: out of vocabulary
PCC: Pearson correlation coefficient
SemEval: Semantic Evaluation
STS: semantic textual similarity

Edited by Y Wang; submitted 10.08.20; peer-reviewed by X Yang, M Manzanares, M Memon; comments to author 22.09.20; revised version received 10.11.20; accepted 16.11.20; published 29.12.20.

Please cite as:
Xiong Y, Chen S, Chen Q, Yan J, Tang B
Using Character-Level and Entity-Level Representations to Enhance Bidirectional Encoder Representation From Transformers-Based Clinical Semantic Textual Similarity Model: ClinicalSTS Modeling Study
JMIR Med Inform 2020;8(12):e23357
URL: http://medinform.jmir.org/2020/12/e23357/
doi:10.2196/23357
PMID:33372664

Original Paper

Extraction of Family History Information From Clinical Notes: Deep Learning and Heuristics Approach

João Figueira Silva[1*], MSc; João Rafael Almeida[1,2*], MSc; Sérgio Matos[1], PhD

[1]Department of Electronics, Telecommunications and Informatics, Institute of Electronics and Informatics Engineering of Aveiro, University of Aveiro, Aveiro, Portugal

[2]Department of Information and Communications Technologies, University of A Coruña, A Coruña, Spain

[*]these authors contributed equally

Corresponding Author:
João Figueira Silva, MSc
Department of Electronics, Telecommunications and Informatics
Institute of Electronics and Informatics Engineering of Aveiro
University of Aveiro
IEETA - Universidade de Aveiro
Campus Universitário do Santiago
Aveiro, 3810-193
Portugal
Phone: 351 234 370 500
Email: joaofsilva@ua.pt

Abstract

Background: Electronic health records store large amounts of patient clinical data. Despite efforts to structure patient data, clinical notes containing rich patient information remain stored as free text, greatly limiting its exploitation. This includes family history, which is highly relevant for applications such as diagnosis and prognosis.

Objective: This study aims to develop automatic strategies for annotating family history information in clinical notes, focusing not only on the extraction of relevant entities such as family members and disease mentions but also on the extraction of relations between the identified entities.

Methods: This study extends a previous contribution for the 2019 track on family history extraction from national natural language processing clinical challenges by improving a previously developed rule-based engine, using deep learning (DL) approaches for the extraction of entities from clinical notes, and combining both approaches in a hybrid end-to-end system capable of successfully extracting family member and observation entities and the relations between those entities. Furthermore, this study analyzes the impact of factors such as the use of external resources and different types of embeddings in the performance of DL models.

Results: The approaches developed were evaluated in a first task regarding entity extraction and in a second task concerning relation extraction. The proposed DL approach improved observation extraction, obtaining F_1 scores of 0.8688 and 0.7907 in the training and test sets, respectively. However, DL approaches have limitations in the extraction of family members. The rule-based engine was adjusted to have higher generalizing capability and achieved family member extraction F_1 scores of 0.8823 and 0.8092 in the training and test sets, respectively. The resulting hybrid system obtained F_1 scores of 0.8743 and 0.7979 in the training and test sets, respectively. For the second task, the original evaluator was adjusted to perform a more exact evaluation than the original one, and the hybrid system obtained F_1 scores of 0.6480 and 0.5082 in the training and test sets, respectively.

Conclusions: We evaluated the impact of several factors on the performance of DL models, and we present an end-to-end system for extracting family history information from clinical notes, which can help in the structuring and reuse of this type of information. The final hybrid solution is provided in a publicly available code repository.

(JMIR Med Inform 2020;8(12):e22898) doi:10.2196/22898

KEYWORDS

natural language processing; rule-based; deep learning; contextual embeddings; word embeddings; family medical history; information extraction; clinical notes; electronic health record

Introduction

Background

For many years, the rapid progress in technology has continually pushed the field of medicine forward, striving for the improvement of health care quality. Novel tools provide new possibilities, such as access to new types of information (eg, medical imaging and genome sequencing) and larger amounts of data, along with associated challenges such as how to store and organize the resulting vast amounts of multimodal medical information. The electronic health record (EHR) solves this by providing an electronic infrastructure for storing structured and unstructured information generated throughout time [1], thus maintaining the patient trajectories by maintaining a longitudinal view over the medical history of patients. Such data can then be explored for applications such as cohort selection [2] or to provide medical entities with clinical decision support [3-5].

Despite being harder to explore, unstructured data can contain relevant information that is not obtainable elsewhere [6], which is particularly evident in clinical notes, where medical narratives allow for more accurate and complete descriptions of medical situations [7]. As there is significant interest in exploring and reusing information from clinical notes, a possible approach is to process free text and extract relevant information that can be stored as structured data [7]. This process has historically been manual, consisting of having clinical experts review clinical notes in search for relevant information. However, heavy reliance on a manual component greatly limits the potential and usability of this process as it cannot scale with the increasing volumes of information [5].

Another possible solution for these cost and scalability issues is the development of automatic systems capable of annotating and extracting relevant content from clinical notes, which has led to greater research efforts in the field of clinical natural language processing (NLP) in the past years. These efforts have led to the creation of international challenges that provide appropriate data sets and enable performance benchmarking of new methods and solutions. The importance of these challenges is widely acknowledged because of the current lack of adequate resources [8], which impedes the development of more advanced solutions [5]. As such, despite the acknowledged interest and value of automated solutions, their development is very complex as it must cope with the challenging nature of working with clinical free text and with the lack of publicly available resources.

Owing to the flexible nature of clinical notes, developed solutions can target the extraction of different types of information from clinical narratives. This process of extracting information is usually split in named entity recognition (NER), named entity normalization (NEN), and relation extraction (RE). NER has the objective of detecting entities of interest in the text, such as diseases or family relatives, whereas NEN is responsible for mapping these entities to normalized concepts in coding standards, such as systematized nomenclature of medicine clinical terms [9] or RxNorm [10] in the case of medical text. RE is focused on detecting relationships between the entities (eg, detecting connections between drugs and adverse drug events) and is very important as it allows the leap from concept extraction to concept understanding [5].

This study focuses on the extraction of the family history component from clinical notes, which can provide insight into disease susceptibility and is important for the prevention, diagnosis, and treatment of specific diseases [11,12]. A demonstration example is the work by Wang et al [13] in which they used a text corpus containing 3 million clinical notes to analyze the patient family history, focusing on family members, medical problems, and their associations, and discovered (1) considerable compliance between positive and negative medical issues mentioned in the reports considering the diagnosis and family history and (2) the existence of medical problems a decade before the diagnosis dates of the determined problem. This study extends a previous contribution [14] by exploring deep learning (DL) approaches for the detection of family history entities in clinical notes and integrating this component in an improved version of the previously developed solution, creating a hybrid system for extracting entities and relations from family history information. The final hybrid solution is provided in a publicly available code repository [15].

The main contributions of this study are as follows:

- This study proposes a strategy to automatically annotate large amounts of EHRs, allowing quick detection of comorbidities with family relations.
- We evaluate the impact of using different DL architectures and embeddings in clinical information extraction.
- We improved the family history information extraction pipeline by combining automatic concept annotations with DL and rule-based architectures to discover entities and relations in the clinical notes.

Related Work

This study is focused on performing NER on clinical notes to extract family history information, namely, family members and observations such as disease mentions, and on detecting associations between detected entities. Correctly detecting family relatives in clinical notes is far from a straightforward task as the following situations must be considered: (1) notes frequently have cascaded information regarding family relatives (eg, “The patient’s grandmother had cancer in her late 60s [she had a cousin who died from cancer] but his grandfather has no history of cancer.”); (2) notes can mention family members with no blood relations, such as the partners of the patients and their relatives; or (3) the relationship of the family member may not be directly expressed. The existence of such situations where the relationship is complex to understand because of the numerous kinship degrees can eventually lead computational systems to lose context, failing to correctly determine the relationship between the detected entity and the patient. In contrast, disease observations can also be troublesome to detect, as, for instance, they can be mentioned as a sequence of several complex terms or even by disjoint mentions.

Existing solutions typically follow rule-based or machine learning-based approaches; however, it is also possible to combine both approaches in hybrid systems. Furthermore, owing to the reckoned potential of DL approaches in the medical field

[16], recent years have shown the emergence of DL-based solutions [5].

For many years, rule-based models were the preferred architecture when developing solutions for extracting family history information, supported by the rationale that, in theory, a good set of rules can manage good concept coverage, thus producing excellent results. Goryachev et al [17] proposed a rule-based algorithm and demonstrated the success of this kind of architecture, whereas Friedlin et al [18] used a rule-based model to extract and code clinical data from clinical reports.

With the growing interest in the development of NLP solutions, generic frameworks such as unstructured information management application [19] and general architecture for text engineering [20] were created to provide support in the development of information extraction systems, from which popular solutions such as clinical text analysis and knowledge extraction system were derived [21]. Despite aiming to offer modular flexible processing workflows that can be reused, these frameworks have the drawback of requiring a deep understanding of the tools given their high-level abstractions.

In contrast with the previous frameworks, toolkits were developed with the goal of providing a set of stand-alone tools that can be easily combined in a processing pipeline. Examples of popular toolkits are the Natural Language Toolkit (NLTK) [22], Apache OpenNLP [23], Stanford CoreNLP [24], and Clinical Language Annotation, Modelling and Processing [25]. Despite the interest in these toolkits, they were developed considering general text instead of biomedical or clinical text, which commonly require specialized tools. Neji was developed to tackle this limitation, providing a modular architecture that integrates specialized modules for biomedical NLP. Thus, it combines the benefits of general frameworks and toolkits with those of specialized tools [26]. These modules can apply different methodologies, such as rule-based models, dictionary matching, and machine learning models. Moreover, Neji provides configurable web services that enable easy integration of its annotation capabilities in external tools [27].

More recently, with the success of DL approaches in text processing problems, DL is being adopted in solutions designed for biomedical and clinical text. One of the key areas where DL has impacted is representation learning, for instance, with the creation of dense representations such as word embeddings. These can be fine-tuned to specific domains and can be easily integrated in other learning algorithms, helping them achieve improved performances in NLP tasks [28]. BioWordVec is an example of publicly available biomedical and clinical word embeddings [29]. However, these embeddings still have the limitation of not considering context, which results in the same word having the same representation when used in completely different contexts (eg, *suits* in *your offer suits our needs* and he *always wears suits*). This was addressed by the development of contextual embeddings such as Embeddings from Language Models [30] and bidirectional encoder representations from transformers (BERT) [31]. These embeddings can also be fine-tuned to specific domains, resulting in the creation of variations such as BioBERT [32] and clinicalBERT [33].

Embeddings are widely used in DL solutions because the resulting dense representations can be easily explored by various DL model architectures. One particular architecture that achieves state-of-the-art results in biomedical and clinical text problems such as NER is the bidirectional long short-term memory (BiLSTM) network coupled with conditional random fields (CRF). Dai et al [34] compared the use of word embeddings (word2vec) and BERT for NER in clinical notes, with a BiLSTM-CRF model, and demonstrated better performance when using BERT to represent clinical text. Li et al [35] used character embeddings, medical dictionaries, and part-of-speech features in a BiLSTM-Att-CRF model, which consists of a BiLSTM with an attention layer bridging the BiLSTM and CRF. This architecture was used to perform clinical NER in EHR notes, and it obtained interesting results, demonstrating the potential of attention mechanisms [35]. More recently, Shi et al [36] used a deep joint learning architecture based on BiLSTMs with word and part-of-speech embeddings for extracting family history information, such as entities and relations from clinical text. Although the demonstrated success of DL approaches at extracting entities and relations from clinical notes, particularly when using BiLSTM-CRF derived architectures, has led to a rapid growth in such solutions, these frequently fail to provide system implementations that hinder their adoption and reproducibility.

Methods

Data Set

This work was originally developed under the scope of the 2019 national NLP clinical challenges (n2c2)/open health NLP track on family history extraction, which had the objective of extracting family history information from EHR clinical notes [37]. This challenge track was split into 2 subtasks: the first one being oriented to named entities and the second one focusing on extracting relations between those entities. More detailed descriptions of each subtask are provided in this section. The second subtask directly depended on the first one, as the challenge had the objective of evaluating developed systems as end-to-end family history summarization solutions.

Training and test data sets were provided by challenge organizers. The training data set consisted of 99 unannotated clinical notes, manual annotations of entities and relations for each clinical note, and a gold standard file with eligible entities and relations for the full training set; the test data set consisted of 117 unannotated clinical notes (a gold standard file with eligible entities and relations for the full test set was only provided after the challenge terminated). Both gold standard files contained the annotations for each document without providing any additional information (eg, annotation span or respective line in document). More detailed statistics of data sets are provided in Table 1.

Table 1. Detailed data set statistics.

Type	Training	Test	Total
Clinical notes, n (%)	99 (45.8)	117 (54.2)	216 (100)
Annotated entities, n (%)			
Family member	667 (53.4)	583 (46.6)	1250 (100)
Observation	930 (50.7)	906 (49.4)	1836 (100)
Annotated relations, n (%)			
Family member: living status	376 (51.9)	349 (48.1)	725 (100)
Family member: observation	740 (49.50)	755 (50.50)	1495 (100)

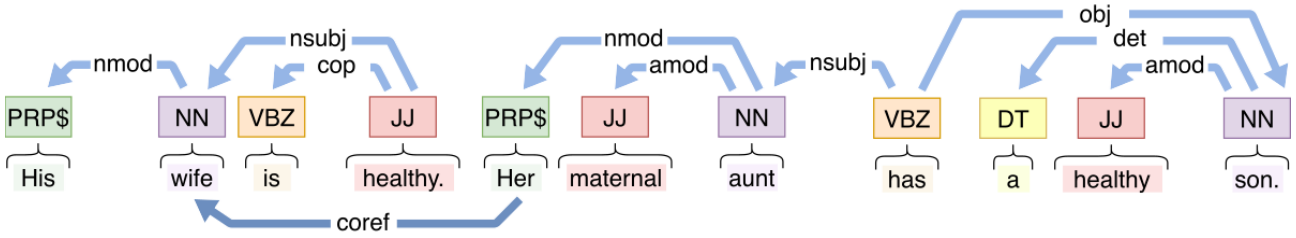
The first subtask had the objective of identifying family member entities and disease mentions in the clinical notes. When extracting family member entities, it was required to extract both the family relationship (eg, son, father, or uncle) and the family side (eg, maternal). The list of relationships considered was provided by organizers and comprised the following: father, mother, parent, brother, sister, son, daughter, child, grandfather, grandmother, grandparent, cousin, sibling, uncle, and aunt. Any relationship outside the provided list (eg, nephew or great grandparent) should be considered invalid. Moreover, clinical notes could contain family member mentions related to the patient and to the patient's partner. As the challenge was focused on the patient, all partner-associated family relationships should be discarded.

The second subtask focused on extracting relations between the previously extracted entities and considered 2 types of relations. The first type involved detecting living status mentions, which should be used to assign a living status score to the respective family member entity. This living status score was computed by multiplying the properties of being alive and healthy, where each property could have a value from 0 to 2 (0: no, 1: not applicable, and 2: yes). The second type of relations involved assigning relations between detected disease mentions and the corresponding family members, taking into consideration if the observation was negated or not (eg, nonnegated: *the patient has diabetes* and negated: *there are no reports of cancer*).

Shortest Dependency Path and Coreference Resolution

The first approach, which was originally used in the challenge submission, combined handcrafted rules and dictionary matching with dependency parsing and coreference resolution. First, a preprocessing step based on Stanford CoreNLP dependency parsing and coreference resolution annotators was applied to all documents. Figure 1 illustrates the result of applying these annotators to an example text fragment.

Figure 1. Illustrative example of dependency parsing and coreference resolution from Stanford CoreNLP. amod: adjectival modifier; cop: copula; coref: coreference; det: determiner; DT: determiner; JJ: adjective; nmod: nominal modifier; NN: noun; nsubj: nominal subject, obj: object; PRP$: possessive pronoun; VBZ: verb third person singular present.



For the first subtask, the process of entity extraction was divided into 2 subproblems targeting family members and disease mention extraction separately. To extract family member entities, a lexicon was compiled that included all family relationships considered for the challenge, expanded with lexical variants and plural forms, along with others identified by examining an extended family tree, such as partner, great grandmother, nephew, and half-uncle. Although the latter family members should not be considered in the final evaluation, their inclusion was necessary at this stage to avoid erroneous associations with other family members during the following step.

The next step consisted of coreference resolution, for which a coreference graph was created to add the corresponding family member annotations to coreferencing pronouns. Considering the example presented in Figure 1, the family member annotation assigned to the mention *wife* is carried over to the pronoun *her* based on the coreference relation. In the example, this also means that the *maternal aunt* mention gets associated to the *wife* family member. In addition, a process of family relationship resolution was performed by applying a set of rules to map extracted mentions to the corresponding family link, with the resulting family link inheriting the family side if it had been extracted. In the same example sentence, the aunt's son is mapped to *cousin*, and this carries over the family side mention, leading to the final annotation of (wife's) *maternal cousin*. Finally, the resulting list of extracted family members was filtered to remove family links other than those targeted in the challenge.

The process of extracting disease mentions consisted of a simpler pipeline, in which a dictionary was first compiled from the unified medical language system Metathesaurus [38]. This

dictionary consisted of a filtered version of the Metathesaurus, containing entries only from the Anatomy and Disorders semantic groups, and was used to configure a Neji annotation service. Once the service was set up, all documents were annotated through the web service and a list of extracted mentions per document was created. As this annotation mechanism could introduce many irrelevant entries (false positives) resulting in a lower precision, a false positive list was created by automatically annotating the corpus provided in the SemEval task on Analysis of Clinical Text [39] and identifying false positives against the gold standard annotation. The resulting false positive list was then used to filter the disease mentions extracted in the n2c2 subtask.

For the second subtask, the objective was to extract 2 types of relations for the previously obtained entities. First, a small lexicon regarding living status was extracted from the training corpus, resulting in the following list: *alive, alive and well, dead, deceased, died, doing well, generally healthy, good general health, good health, healthy, living, living and well, otherwise healthy, passed away, stillborn, well, and without problems*. This lexicon was used to extract living status mentions from the documents, which were then mapped into an integer value using the scale previously described in the data set subsection. Finally, the dependency graph created in the first subtask was used to extract the shortest dependency path that associated each disease mention/living status with a family member. This approach disregarded the negation component in observations; therefore, all disease-family member relations were considered nonnegated.

Rule-Based Engine

The second approach used in the official submissions for the n2c2 challenge track followed a different strategy and consisted of a rule-based engine. This solution involved the creation of rules for family member recognition and dictionaries for observation extraction and processed both subtasks as an end-to-end system outputting the required submission files for both subtasks. After the challenge contribution, this approach was adapted and improved as described further in this section.

The engine processed each sentence in a document sequentially, aiming to link sentences when one of the system processing flows did not detect family members in a sentence. Therefore, using this approach, we created a system that tried to answer the following 3 questions:

1. Who is the subject of the sentence?
2. Which observations are in the sentence?
3. Is the subject alive?

Although answering these 3 questions does not entirely solve the proposed problem, managing to correctly answer them simplifies the process of establishing relations between extracted concepts. The first step in the processing flow splits the document into sentences and removes a considerable set of words. This set was composed of the most common English verbs and the most common conjugations, several adjectives, and names. This procedure preserved relevant words and reduced the distance between words that allowed the correct identification of family members and their respective family side. For instance, for a rule-based system, it is easier to find the family member *cousin* in the cleaned sentence *patient's uncle son* than in the original sentence *the patients' uncle has one son*. In this example, this could be erroneously processed as a sentence where the primary subject is the patient's uncle, instead of the cousin.

After cleaning the sentences, the system applied rules that enable the identification of the subject in the most trivial cases, using exact matching. When no subject was identified, the system processed this using another component, with more complex rules. In this case, rules have more properties such as a set of words that should exist before and after the detected family member, and if this should be discarded or not. These properties enable the generation of very precise rules, which, if used, can increase the potential of the system for the specifications of the challenge at the cost of reducing its reuse in other scenarios (ie, trade-off specificity-generalizing capability).

When no family member was detected with the previous rules, the system executed another component that tried to identify if the sentence currently being processed was related to the previous sentence. If the sentence being processed was the first sentence in the document, the system considered by default that it was related to the patient. Finally, the system ran a last component, which was always executed, to discover whether the sentence was related to the patient or the patient's partner. If the sentence was associated with the partner, the system discarded the family member entity as required by the challenge guidelines.

Observation extraction consisted of a simpler process than that of family member detection. However, it followed the same principles and used the initial preprocessing for cleaning a set of words. For the challenge, we created a vocabulary based on the observations annotated in the training set and used it in the test set. Simultaneously, the system applied rules to map the detected observation to the identified subject in the sentence. When it was not possible to identify a relation in a sentence, the system did not discard the extracted observations as they were still important for the first subtask.

Living status identification was performed using 2 sets of rules: one targeting deceased subjects and the other targeting healthy and alive subjects. Owing to time constraints, we did not try to identify cases where subjects were alive but not healthy because based on a statistical analysis, mentions for this group of entries represented only 12.2% (46/376) of the living status entries in the gold standard of the training set.

The rule-based engine pipeline processes documents individually and sentence by sentence following a sequential flow. In this pipeline, the detected words have different levels of importance. For instance, terms like *partner* and *patient* coexisting in the same sentences are weighted differently. These weights were considered by the complementary rules during subject identification in a sentence. Disambiguation was performed using a set of verbs and specific words in situations where it was not clear whether the sentence was related to the patient, the patient's relatives, the patient's partner, or the partner's relatives. Figure 2 shows an excerpt of a clinical note that

illustrates clearly how the system processes original sentences and what is the result of this processing.

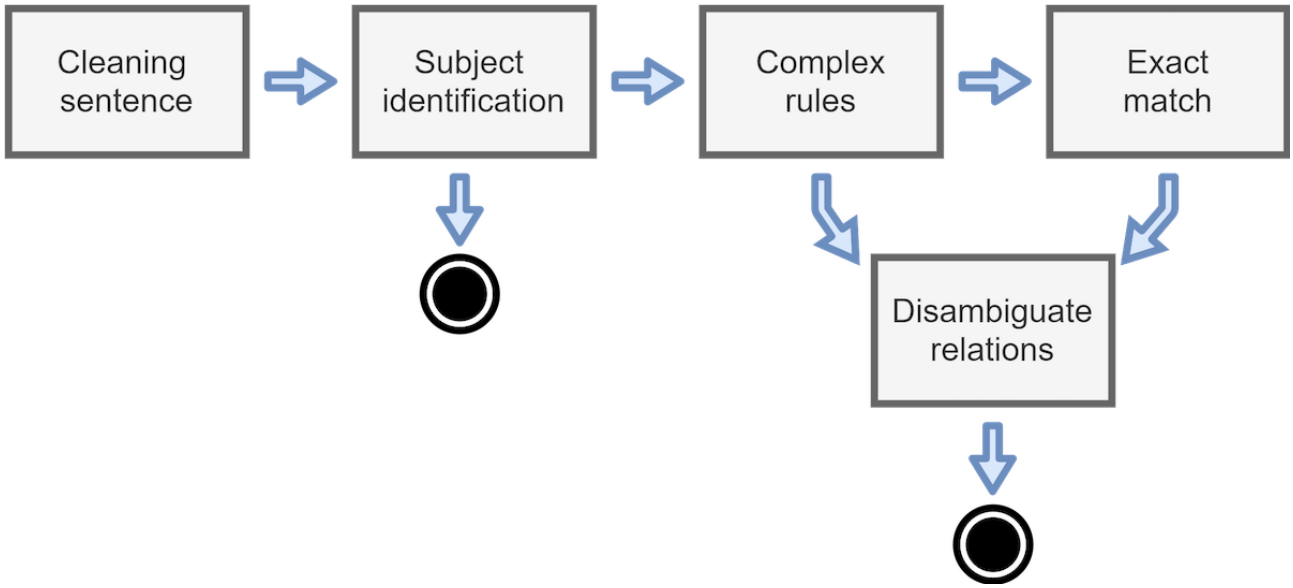
Figure 2. The 3 left concepts represent the main points that the system tries to identify in the text on the right. Highlighted on the right are relevant words for the system to be able to make decisions. Auxiliary words that help identify the subject are represented in green. The words used to identify if the relatives are related to the patient or the partner are highlighted in purple. Blue represents annotated family members, and yellow is used for diseases. Red is used to highlight words concerning subject living status.

Subject	Observation	Living status	
Patient	-	-	Family history information was obtained from the patient and her partner this morning.
Patient	-	-	Details from the family histories are on file in the Department of Medical Genetics.
Patient	-	-	Pertinent information is as follows:
[Sister, Brother]	-	-	Ms. Benjamin has one sister, age 32, and two brothers, ages 34 and 17, who are all reportedly healthy.
Nephew	Dubowitz syndrome	-	One of her brothers has a son diagnosed with Dubowitz syndrome.
Parents	-	Alive	Her parents, ages 53 and 50, are alive and well.
...			...

This engine managed good results in the annotation of the family members of the patient. However, the methodology used to extract observations was not the best, regardless of possible improvements to produce more accurate results. Therefore, in a postchallenge contribution, we removed the components for detecting observations and improved components responsible for extracting the family members of the patient and their living status. The living status component was reused with small adjustments to be more generic and compatible with different data sets, yet maintaining the same philosophy of trying only to identify whether the patient is healthy and alive or dead.

The family members annotator was rebuilt following the initial principles but without specific sets of rules that were generated from the training set of the challenge (ie, to reduce overfitting). The system pipeline is presented in a scheme (Figure 3) representing the system pipeline and how components are interconnected. This flow starts by trying to identify if the subject in the sentence is the patient. If not identified, the previously described complex rules are executed. The third component performs exact matching over a clean sentence for trivial annotations, and the output of these components is filtered to disambiguate relations between family members and to remove any relations that should be discarded (eg, to adhere to challenge evaluation guidelines).

Figure 3. Overview of the processing workflow responsible for family members detection, for the rule-based engine.



In the complex rules component, rules follow a 6-part structure where it is defined the keyword that triggers the rule (eg, *father or grandparent*), and a list of terms that must appear before or after this keyword are defined. Next, this structure contains a flag that indicates whether the annotated relative must be considered or discarded and indicates which is the detected relative. As an example, if the keyword *grandparents* is detected in the clean text, a rule can identify it as a paternal grandparent if there exists the set of words *patients* and *paternal*, in this order, preceding the keyword.

Regarding the disambiguation component, the system contains a set of rules composed of 4 elements. These rules have 2 relatives and a mapping to the real relation of this subject to the patient. As an example, if the component annotates and processes the relatives *father* and *brother*, the system will map them to paternal uncle and return the corrected annotation. Besides the above-mentioned examples, the rule-based system contains a more extensive list of rules that were used for the processes of partial and exact match search.

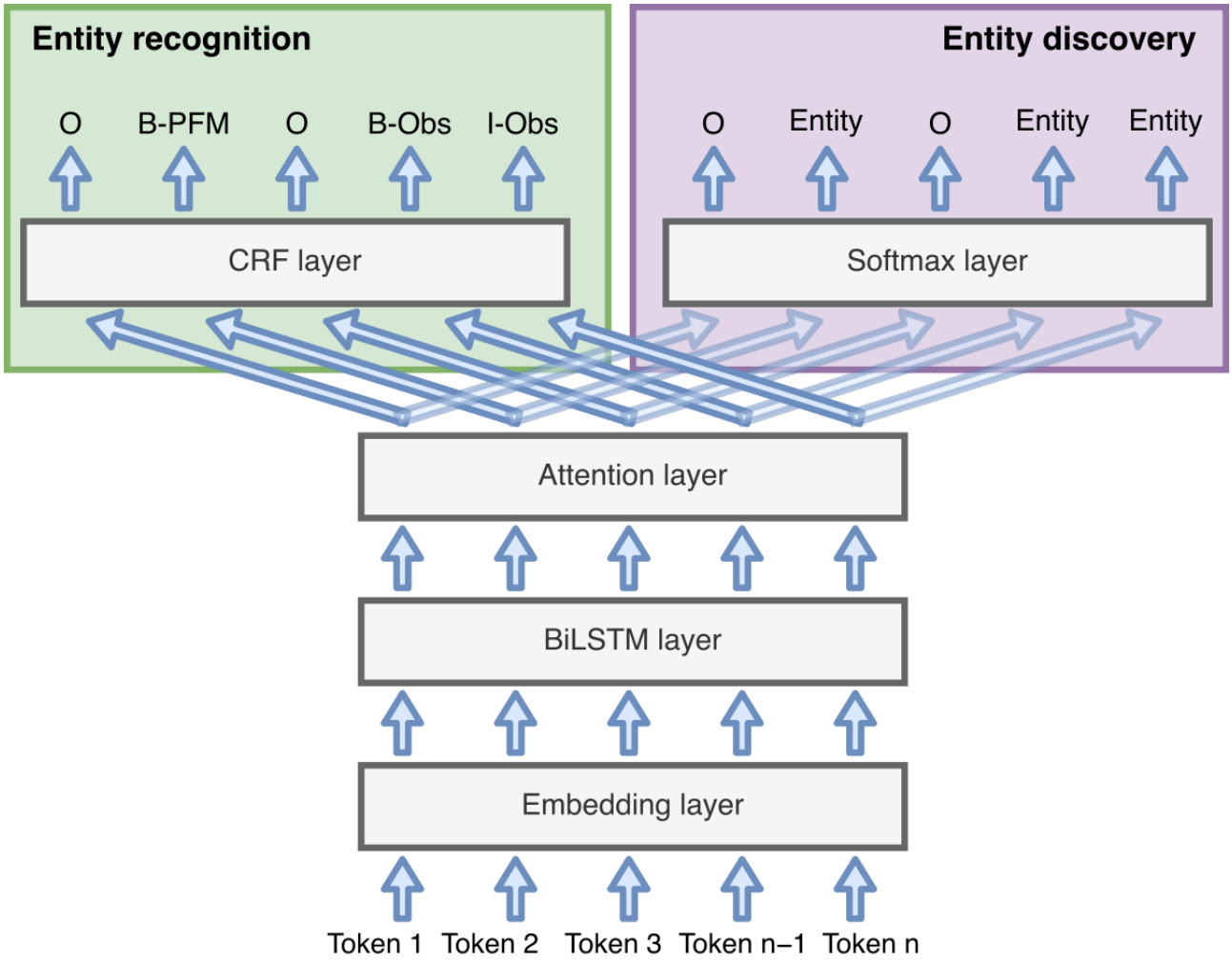
DL for Entity Extraction

Owing to the acknowledged potential and success of recent DL solutions in clinical text problems, we extended the original contribution with a novel approach based on DL. The implementation of this solution considered several aspects, namely:

- Following the trend in state-of-the-art solutions, we explored the widely used attention-based BiLSTM-CRF with the attention mechanism placed between the BiLSTM and CRF layers [35] and compared it with a simple linear classifier (with softmax) to evaluate the impact of model architecture in downstream tasks.
- Similar to the approach presented by Yang et al [40], an additional task regarding named entity discovery was integrated with the objective of improving model perception of unknown entities. This downstream task was set as optional; thus, it is possible to train models for NER and for NER and discovery.
- Different types of embeddings were explored for clinical text representation to assess their impact on model performance. BioWordVec word embeddings and clinicalBERT contextual embeddings were used.
- To evaluate the impact of using external resources in model development, Neji annotations were integrated into the input to the model.

A schematized view of the model architecture used in this study (attention-based BiLSTM-CRF) is presented in Figure 4.

Figure 4. Schematic diagram of the general deep learning model architecture used in this study, showing the 2 possible downstream tasks. The entity recognition task is always executed, whereas the entity discovery task was added as optional to enable model development with and without it. BiLSTM: bidirectional long short term memory; B-Obs: beginning observation; B-PFM: beginning patient family member; CRF: conditional random field; I-Obs: inside observation; n: number of tokens in tokenized sentence; O: outside.



The named entity discovery downstream task consists of a binary classification problem where the system classifies whether an input token is part of an entity or not, disregarding the respective class (ie, if it is an observation or family member mention). This optional task was integrated with the objective of making the model consider the trade-off between discovering more entities and correctly identifying them. When enabled, it is reflected in model training during backpropagation, with the total loss resulting from a linear combination between the losses of both downstream tasks.

Before training any model, it was necessary to preprocess the data set. Text preprocessing began by splitting each document in sentences using the sentence splitter from NLTK, followed by tokenization. However, 2 different tokenization methods had to be used because word and contextual embeddings take different tokenizing approaches: the NLTK word tokenizer was

used for word embeddings, and the BERT tokenizer was used for contextual embeddings. The resulting tokenized sentences were tagged using the BIO (beginning, inside, and outside) tagging scheme. Finally, to assess the impact of using external resources, all documents were annotated using Neji, which uses standard vocabularies to detect entity mentions in the input text. Neji annotations, consisting of text spans and entity classes, were then mapped to the tokens in the corresponding sentence, with each token being assigned an integer value similar to the BIO scheme: 0 for tokens not annotated by Neji, 1 for the first token in an annotation, and 2 for the following tokens. The resulting lists of classes were normalized and concatenated to the embedding representations and then forwarded through the BiLSTM layer.

Model training and evaluation were performed using 5-fold cross validation. The Adam optimizer was used, and models were trained with early stopping (the patience parameter can be adjusted). Each training epoch consisted of 100 iterations, during which the training partition was randomly sampled. A detailed list of hyperparameters is provided in Table 2.

Table 2. List of hyperparameters used for deep learning model training.

Hyperparameters	Value
Dimension of BioWordVec embeddings	200
Dimension of clinicalBERT[a] embeddings	768
BiLSTM[b] hidden size	256
Number of attention heads	2
Epochs	100
Patience	5
Iterations per epoch	100
Dropout rate	0.5
Learning rate	0.001
Batch size	32
Epochs for training BioWordVec embeddings	2

[a]clinicalBERT: clinical bidirectional encoder representations from transformers.

[b]BiLSTM: bidirectional long short-term memory.

In addition, because contextual embeddings provide additional information when compared with word embeddings, we enabled the training of word embeddings for a number of epochs at the beginning of model training, after which the embedding layer was frozen. Finally, as contextual embeddings can partition words in various smaller tokens (eg, *carcinoma* is split in *car,##cin,* and *##oma)*, the model could classify only parts of a word as entities (eg, *##cin* and *##oma* classified as entities and *car* as nonentity), resulting in incomplete entities and poor results. Therefore, we added a reconstruction mechanism where the full word is considered when only a part of it is classified as an entity.

The DL approach obtained interesting results in observation extraction but poor performance in family member detection, which goes in contrast with the rule-based approach. As such, we created a final hybrid solution that integrates the DL approach as an observation extraction module in the rule-based engine.

Results

The original contribution consisted of the development of 2 different approaches for entity and RE: one using shortest dependency paths combined with coreference resolution and another using a rule-based engine. These approaches were validated in the n2c2 challenge on family history extraction. Results obtained in the test data set (Table 3) showed that overall, the first approach performed better in the entity extraction subtask, whereas the rule-based approach performed better in the RE subtask.

Table 3. Original overall test results for the 2 national natural language processing clinical challenges subtasks; approach 1: shortest dependency path and coreference resolution and approach 2: rule-based engine.

Subtasks and approach	Precision	Recall	F_1 score
Subtask 1			
Approach 1	0.6501	0.8892	0.7510
Approach 2	0.8507	0.6211	0.7180
Subtask 2			
Approach 1	0.5406	0.5005	0.5198
Approach 2	0.6468	0.5992	0.6221

As the results obtained during the challenge had margins for improvement, and DL-based approaches dominated system submissions in the challenge, we opted to experiment with DL to improve the previous contribution. For the sake of simplicity, tables presenting DL-related results only contain F_1 score values. However, more detailed results (including precision and recall metrics) are presented in Multimedia Appendix 1.

For the DL-based approach, we started by testing a simple model configuration composed of a linear layer and a softmax function, using contextual embeddings for clinical text representation (Table 4). This simple model served as a reference point to assess the potential of using contextual embeddings to represent clinical text.

Table 4. Cross validation results on the training data set (5-fold cross validation) for subtask 1 using a deep learning model composed of clinical bidirectional encoder representations from transformers embeddings, a linear layer, and softmax function, with and without token reconstruction. For simplicity purposes, only F1 scores are presented.

Reconstruction approach and model configuration	Family member	Observations	Overall
No reconstruction			
Baseline	0.3071	0.6620	0.5647
Baseline+ED[a]	0.1764	0.6397	0.5204
Baseline+Neji	0.3088	0.7019	0.5924
Baseline+ED+Neji	0.1840	0.6841	0.5523
Reconstruction			
Baseline	0.3071	0.7444	0.6241
Baseline+ED	0.1764	0.7142	0.5753
Baseline+Neji	0.3088	0.7712	0.6418
Baseline+ED+Neji	0.1840	0.7593	0.6070

[a]ED: entity discovery.

After testing with a simple architecture and evaluating the impact of adding an entity discovery downstream task and external resources to the model, we proceeded to the more complex architecture of the attention-based BiLSTM-CRF, which has been widely explored in the state of the art. This architecture was first tested using contextual embeddings for text representation to assess the impact of model capacity on the resulting model performance (Table 5). After observing the improvements resulting from the change in model architecture, we then evaluated the influence of the embeddings used in the final system results by training the same architecture with word embeddings (Table 5). As word embeddings capture less information than their contextual counterpart, we integrated the possibility of fine-tuning word embeddings for a number of epochs at the beginning of the training process, freezing the embeddings after that point.

Table 5. Cross validation results on the training data set (5-fold cross validation) for subtask 1 using the attention-based bidirectional long short-term memory network coupled with conditional random fields with different types of embeddings. When using word embeddings, some configurations enabled embedding fine-tuning for 2 epochs. For simplicity purposes, only F1 scores are presented.

Embeddings type and model configuration	Family member	Observations	Overall
clinicalBERT[a]			
Baseline	0.4103	0.8596	0.7194
Baseline+ED[b]	0.3788	0.8481	0.7023
Baseline+Neji	0.3545	0.8478	0.6908
Baseline+ED+Neji	0.3485	0.8688	0.7081
BioWordVec			
Baseline	0.5921	0.8140	0.7317
Baseline+ED	0.6553	0.8276	0.7627
Baseline+ET[c]	0.6166	0.8285	0.7513
Baseline+ED+ET	0.6219	0.8367	0.7579
Baseline+ED+Neji	0.7222	0.8529	0.8036
Baseline+ED+ET+Neji	0.7266	0.8587	0.8092

[a]clinicalBERT: clinical bidirectional encoder representations from transformers.

[b]ED: entity discovery.

[c]ET: embeddings training.

Although the use of a more complex model architecture provided promising results, there was a common trend among all used models, which was the fact that these approaches performed much better at extracting observations than family members.

Considering the fact that the rule-based engine struggled in observation extraction while obtaining good performance in family member extraction [14] and that it performed better in the RE subtask than the shortest dependency path approach, we created a hybrid system that complements the rule-based engine by adding a DL module responsible for extracting disease mentions. Table 6 presents the results obtained with the hybrid solution in the test data set.

Table 6. Test results for both subtasks using the final hybrid solution: rule-based engine combined with deep learning module for observation extraction.

Subtask and annotation type	Precision	Recall	F_1 score
Subtask 1			
Family members	0.7887	0.8307	0.8092
Observations	0.7523	0.8332	0.7907
Overall	0.7662	0.8322	0.7979
Subtask 2			
Living status	0.5964	0.6462	0.6248
Observations	0.4635	0.4371	0.4499
Overall	0.5100	0.5063	0.5082

Discussion

Principal Findings

DL for Entity Extraction

Word embeddings have been the go-to method for text representation in the past years. However, contextual embeddings have made a big impact in recent years as they consider positional information and context in the resulting representation, which provides them with higher disambiguation capability than that of word embeddings. As such, our initial tests were performed using publicly available contextual embeddings fine-tuned on biomedical and clinical corpora.

First, we analyzed the impact of reconstructing annotated tokens on the resulting performance. Tests with a simple model (Table 4) showed improved performance in every model configuration when using token reconstruction. However, it is noticeable that only observation extraction benefited from this process, with family member extraction maintaining its F_1 scores. This is explained by the fact that disease mentions can be very specific and more complex when compared with family members, for instance, the word *mother* is tokenized by the contextual embedding tokenizers as *mother*, whereas *carcinoma* is

tokenized as *car, ##cin, and ##oma*. Owing to this different word decomposition, the DL model can classify only parts of the word as an entity, resulting in incomplete entities. The reconstruction procedure solved this issue by adding the missing parts to these entities. Tests with the simple model also demonstrated that the use of external resources such as Neji annotations can help improve entity extraction, whereas adding an additional downstream task regarding entity discovery led to worse results with this model. Finally, it was clear that the model managed to extract disease mentions from clinical notes but failed in the detection of family members, leading to lower overall F_1 scores.

After performing the initial tests with a simple model and verifying the importance of token reconstruction when using contextual embeddings, we moved to the more complex architecture of the attention-based BiLSTM-CRF (Table 5). To be able to compare it with the previous model, we began by testing the new model with contextual embeddings. Starting with baseline models, it is possible to see that changing to the higher capacity model increased F_1 scores by approximately 0.1 across all categories. Next, it is possible to observe that complementing the baseline model with the entity discovery task and Neji resources resulted in worse overall F_1 scores; nonetheless, their combination led to an increase in the F_1 score for observation extraction (0.8596 to 0.8688).

Finally, to evaluate the influence of using different types of embeddings to represent clinical text, we tested the same model architecture with publicly available word embeddings fine-tuned on biomedical and clinical corpora. Comparing baseline models, word embeddings led to a higher overall performance (0.7317 vs 0.7194), lowering the observation extraction F_1 score but improving that of family member extraction. Adding extra mechanisms such as external annotations and entity discovery progressively increased model performance, with the final model showing a much higher overall F_1 score compared with the best contextual embedding configuration (0.8092 vs 0.7194). This higher overall performance was caused by a significant increase in the family member F_1 score (0.4103 to 0.7266), although observation extraction decreased from 0.8688 to 0.8587 F_1 score.

The previous results demonstrated that despite the increasing focus on contextual embeddings, word embeddings can obtain good results when using state-of-the-art model architectures. In spite of its much better performance in family member extraction, the word embedding model still obtained subpar performance when compared with the rule-based engine in the same task (0.7266 vs 0.8823). As the objective was to integrate the best approach for observation extraction in the rule-based engine, and contextual embeddings obtained the upper hand in that aspect (0.8688 to 0.8587), we integrated the attention-based BiLSTM-CRF with clinicalBERT embeddings in the hybrid system.

Hybrid System

The original rule-based system was developed focusing on the n2c2 challenge and contained sets of rules that were adjusted to the training set. These rules were removed after the challenge, whereas other existing rules were carefully adjusted to create a better system that retained its generalizing capabilities.

With the objective of exploring the best developed approaches for each component of the subtasks, we based the final system on the improved rule-based engine and substituted its weaker component (observation extraction) by a DL-based module. The result was a hybrid system capable of extracting family members and observations along with their respective relations.

As experienced in the original contribution, the results obtained in the test set showed a decrease in performance (Table 6), presenting an overall F_1 score of 0.7979 in subtask 1 and an overall F_1 score of 0.5082 in subtask 2. For the first subtask, the hybrid system showed an improvement from the previous best result of 0.7510 overall F_1 to 0.7979 (a 4.69 percentage point increase). Regarding the RE subtask, although the overall F_1 score decreased from 0.6221 to 0.5082, there are 2 aspects that should be considered. The first aspect is that adjustments were made to the rule-based engine, which reduced the specificity of its rules and impacted the challenge performance. The second one is that results presented for subtask 2 were obtained using a modified version of the evaluator. The adjusted evaluator performs a more exact analysis of the system output, resulting in lower performance values compared with the original counterpart. A more detailed explanation of this last aspect is provided in the following subsection of *Evaluation and Error Analysis*.

Evaluation and Error Analysis

The annotations resulting from the approaches described were evaluated using precision, recall, and F_1 score metrics. The items considered in subtask 1 evaluation were the patient family members combined with their family side and the observations in each document. Regarding family members, if the system does not properly extract relatives' family side, the results are considered a false positive and a false negative. However, in the case of observations, the evaluator was more flexible. More specifically, if observations were partially annotated (eg, for the observation *diabetes type 2*, the system extracted only *diabetes*), the evaluator considered a true positive. This evaluator was provided by the n2c2 organizers, and we maintained its principles.

The evaluation process for the RE subtask considered (1) the attribution of living status to family members, with correct family side, and (2) the association of observations to family members, including the indication of whether the observation was negated or not. The original evaluator considers each family member, observation, and negation status triple correctly identified by a system. However, the evaluator considers it as a true positive if only the observation or only the negation status were correctly extracted for a given relative. This formulation produces additional true positives, even for annotations that are not completely correct. Therefore, we changed the behavior of this evaluator to consider as true positive only when the system correctly extracted the family member, the respective family side, the (possibly partial) observation, and the observation logical status, as we believe that the extraction is more useful if it is completely correct. As an effect of this change, the F_1

scores of our challenge submission reduced approximately 10 percentage points when compared with the official results. For instance, when using the new evaluator, approach 1 reduced its F_1 score from 0.5198 to 0.4431, whereas approach 2 decreased its F_1 score from 0.6221 to 0.4818.

To understand what affects our results, we randomly selected some false positives and performed a manual analysis on the training set. This analysis led to the detection of inconsistencies in the gold standard annotations, which adversely affected the performance of our system. For instance, in the same clinical notes, 2 identical sentences regarding different family members were annotated with different living statuses. Another example was that at least 14 relatives without living status were annotated when this was present in the gold standard raw data. This raw data consists of the XML files supplied along with the clinical notes in the training set, which were the base of the submission gold standard file. In some of the clinical notes, we detected observations that were present in the text but not annotated in the gold standard and observations that were detected and present in subtask 1 gold standard but not attributed to any subject (despite having the family member also annotated in the gold standard). Although we were not able to perform an in-depth analysis and assess how much this affected our scores, the identified inconsistencies had some impact on performance.

Limitations and Future Work

The resulting system was built to be more generic than the previous version, which was used in the n2c2 challenge. Despite the improvements made to the system, there are still some limitations. Textbox 1 presents some sentences extracted from the clinical notes that are representative examples of the system limitations.

Textbox 1. Analyses of some of the false positives and false negatives classified by the proposed system. Family member annotations are emphasized in the sentence using italics.

Child not applicable (N/A)

"Mr. Smith's father suffers from cancer. He has several *children* through several other women..."

Daughter N/A

"The maternal/paternal great-aunt that has diabetes had several children. One of these individuals had a cancer of an unknown type and is deceased. The second *daughter* is the individual with diabetes type 2..."

Parent N/A

"John's *parents* are both reportedly healthy at age 63, but they have not seen a physician in approximately 30 years. John's mother had one second trimester miscarriage..."

Sibling N/A

"Saul's father is a 39-year-old man who is a college graduate and who has a total of 5 *siblings*..."

Grandparents N/A

"While living in Texas, they lived with extended family, including Peter's *grandparents*..."

The first example of these limitations concerns the establishment of incorrect sentence connections in certain situations. Depending on the scenario, in the first sentence in Textbox 1, it could be annotated *child* or *sibling*, as it is influenced by the order in which rules are applied during family members detection. However, in this example, the pronoun *he* refers to the patient's father. Thus, the mentioned children are patient's half-siblings, a relative that should not be considered according to the guidelines.

The problem in the second example is also related to sentence linking. The system detects a daughter because it loses the sentence context. In addition, the existence of *maternal/paternal* before a relative led to inconsistencies in the detection because there are no rules for these situations. Despite all those problems, the relative annotated as daughter is in fact a third-degree cousin, a relationship that should not be considered. The third and fourth examples show other cases where there was an incorrect family member annotation because of the system losing context within the sentences.

The final example is a special case because the annotation was correctly performed but was not considered in the gold standard annotations, as the clinical notes did not provide any clinical information about the relative. Moreover, the clinical information regarded as necessary for annotating a relative mention is not exclusively composed of observations and may comprehend other types of information such as medication intake or medical procedures, which invalidates the possibility of filtering such situations based on observation associations alone.

Although these might not be the only problems, the limitations presented were those that stood out the most. This led us to analyze possible future work for this contribution, which we could split in different topics. First, we need to test this system in another data set, with a more solid gold standard. This will help us understand the performance of the system as well as its versatility in detail. Another task is the extension of the clinical information extracted. The current version has models designed to extract observations. However, we intend to build other models to extract drugs and procedures, among other medical categories that were not required in the challenge. This extension would lead to a reformulation of the detection of patient's relatives and allow filtering mentions with no medical information, such as the last example in Textbox 1). Finally, there is also the possibility of exploring machine learning and

DL for the process of establishing relations between extracted entities.

Conclusions

We present an extension to a previous work that focused on extracting family history information from clinical notes. Specifically, we developed a more generic system and improved the previous F_1 score in the entity extraction subtask by approximately 5 percentage points by combining different approaches. Although the rule-based engine succeeded in extracting patient relatives because of the range of possibilities in the text, this approach failed in the detection of observations. However, the use of DL models helped rectify this gap, with the hybrid system taking advantage of the best characteristics of these 2 methodologies. The hybrid solution is provided in a publicly available code repository.

This study promotes new strategies to easily annotate large amounts of clinical reports currently available in EHR systems. If these reports were annotated and indexed, it would be simpler for a clinician to search for reports mentioning specific concepts. In addition, with data in a structured format, this information can be reused in other scenarios, such as predicting the patient's susceptibility or predisposition to diseases.

Acknowledgments

This work was supported by the European Union/European Federation of Pharmaceutical Industries and Associations Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 806968 and by the NEw Targets in DIAstolic heart failure: from coMOrbidites to persoNalizeD medicine (NETDIAMOND) project (POCI-01-0145-FEDER-016385), cofunded by Centro 2020 program, Portugal 2020, European Union. JS and JA are supported by Foundation for Science and Technology (national funds), under the grant numbers PD/BD/142878/2018 and SFRH/BD/147837/2019, respectively.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed results for all deep learning model configurations tested in this work. Precision, recall, and F1-scores are provided separately for family member and observation extraction along with overall results.
[PDF File (Adobe PDF File), 72 KB - medinform_v8i12e22898_app1.pdf]

References

1. Katehakis DG, Tsiknakis M. Electronic health record. In: Wiley Encyclopedia of Biomedical Engineering. New Jersey, United States: John Wiley & Sons; 2006.
2. Antunes R, Silva JF, Pereira A, Matos S. Rule-based and Machine Learning Hybrid System for Patient Cohort Selection. In: Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5. 2019 Presented at: HEALTHINF'19; February 22-24, 2019; Prague, Czech p. 59-67. [doi: 10.5220/0007349300590067]
3. Kukafka R, Ancker JS, Chan C, Chelico J, Khan S, Mortoti S, et al. Redesigning electronic health record systems to support public health. J Biomed Inform 2007 Aug;40(4):398-409 [FREE Full text] [doi: 10.1016/j.jbi.2007.07.001] [Medline: 17632039]
4. Almeida JR, Oliveira JL. GenericCDSS - A Generic Clinical Decision Support System. In: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS). 2019 Presented at: CBMS'19; June 5-7, 2019; Córdoba, Spain p. 186-191. [doi: 10.1109/cbms.2019.00046]
5. 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]
6. Jensen K, Soguero-Ruiz C, Mikalsen KO, Lindsetmo R, Kouskoumvekaki I, Girolami M, et al. Analysis of free text in electronic health records for identification of cancer patient trajectories. Sci Rep 2017 Apr 7;7:46226 [FREE Full text] [doi: 10.1038/srep46226] [Medline: 28387314]
7. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc 2011;18(2):181-186 [FREE Full text] [doi: 10.1136/jamia.2010.007237] [Medline: 21233086]
8. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform 2020 Mar 31;8(3):e17984 [FREE Full text] [doi: 10.2196/17984] [Medline: 32229465]
9. Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. Proc AMIA Symp 2001:662-666 [FREE Full text] [Medline: 11825268]
10. Liu S, Ma W, 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]

11. Guttmacher AE, Collins FS, Carmona RH. The family history-more important than ever. N Engl J Med 2004 Nov 25;351(22):2333-2336. [doi: 10.1056/NEJMsb042979] [Medline: 15564550]
12. Dick RS, Steen EB, Detmer DE. The Computer-Based Patient Record: An Essential Technology for Health Care, Revised Edition. Washington, DC: The National Academies Press; 1997.
13. Wang Y, Wang L, Rastegar-Mojarad M, Liu S, Shen F, Liu H. Systematic analysis of free-text family history in electronic health record. AMIA Jt Summits Transl Sci Proc 2017;2017:104-113 [FREE Full text] [Medline: 28815117]
14. Almeida JR, Matos S. Rule-based Extraction of Family History Information From Clinical Notes. In: Proceedings of the 35th Annual ACM Symposium on Applied Computing. 2020 Mar Presented at: SAC'20; March 30-April 3, 2020; Online. [doi: 10.1145/3341105.3374000]
15. PatientFM: An end-to-end system for extracting family history information from clinical notes. GitHub. 2020. URL: https://github.com/bioinformatics-ua/PatientFM [accessed 2020-12-18]
16. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 2018 Apr;15(141) [FREE Full text] [doi: 10.1098/rsif.2017.0387] [Medline: 29618526]
17. Goryachev S, Kim H, Zeng-Treitler Q. Identification and extraction of family history information from clinical reports. AMIA Annu Symp Proc 2008 Nov 6:247-251 [FREE Full text] [Medline: 18999129]
18. Friedlin J, McDonald CJ. Using a natural language processing system to extract and code family history data from admission reports. AMIA Annu Symp Proc 2006:- epub ahead of print [FREE Full text] [Medline: 17238544]
19. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Nat Lang Eng 1999;10(3-4):327-348. [doi: 10.1017/s1351324904003523]
20. Cunningham H. GATE, a general architecture for text engineering. Comput Hum 2002;36(2):223-254. [doi: 10.3115/974281.974299]
21. 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]
22. Bird S, Klein E, Loper E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit 1st Edition. California, United States: O'Reilly Media; 2009.
23. Welcome to Apache OpenNLP The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text. Apache OpenNLP. URL: http://opennlp.apache.org/ [accessed 2020-06-01]
24. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2014 Presented at: ACL'14; June 23-24, 2014; Baltimore, Maryland p. 55-60. [doi: 10.3115/v1/p14-5010]
25. 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 1;25(3):331-336 [FREE Full text] [doi: 10.1093/jamia/ocx132] [Medline: 29186491]
26. Campos D, Matos S, Oliveira JL. A modular framework for biomedical concept recognition. BMC Bioinformatics 2013 Sep 24;14:281 [FREE Full text] [doi: 10.1186/1471-2105-14-281] [Medline: 24063607]
27. Matos S. Configurable web-services for biomedical document annotation. J Cheminform 2018 Dec 21;10(1):68 [FREE Full text] [doi: 10.1186/s13321-018-0317-4] [Medline: 30578450]
28. 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: NIPS'13; December 5-10, 2013; Lake Tahoe p. 3111-3119.
29. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci Data 2019 May 10;6(1):52 [FREE Full text] [doi: 10.1038/s41597-019-0055-0] [Medline: 31076572]
30. Peters M, 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: NAACL'18; June 1-6, 2018; New Orleans, Louisiana. [doi: 10.18653/v1/n18-1202]
31. 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 Presented at: NAACL'19; June 2-7, 2019; Minneapolis, Minnesota. [doi: 10.3115/1614108]
32. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
33. Alsentzer E, Murphy J, Boag W. Publicly Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019 Presented at: ClinicalNLP'19; June 7, 2019; Minneapolis, Minnesota, USA p. 72-78. [doi: 10.18653/v1/w19-1909]

34. Dai Z, Wang X, Ni P, Li Y, Li G, Bai X. Named Entity Recognition Using BERT BiLSTM CRF for Chinese Electronic Health Records. In: 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). 2019 Presented at: CISP-BMEI'19; October 19-21, 2019; Suzhou, China p. 1-5 URL: https://ieeexplore.ieee.org/abstract/document/8965823 [doi: 10.1109/cisp-bmei48845.2019.8965823]
35. Li L, Zhao J, Hou L, Zhai Y, Shi J, Cui F. An attention-based deep learning model for clinical named entity recognition of Chinese electronic medical records. BMC Med Inform Decis Mak 2019 Dec 5;19(Suppl 5):235 [FREE Full text] [doi: 10.1186/s12911-019-0933-6] [Medline: 31801540]
36. Shi X, Jiang D, Huang Y, Wang X, Chen Q, Yan J, et al. Family history information extraction via deep joint learning. BMC Med Inform Decis Mak 2019 Dec 27;19(Suppl 10):277 [FREE Full text] [doi: 10.1186/s12911-019-0995-5] [Medline: 31881967]
37. 2019 n2c2 Shared-task and Workshop, Track 2: n2c2/ohnlp Track on Family History Extraction. Harvard Medical School. URL: https://n2c2.dbmi.hms.harvard.edu/track2 [accessed 2020-06-01]
38. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
39. Pradhan S, Elhadad N, Chapman W, Manandhar S, Savova G. Semeval-2014 Task 7: Analysis of Clinical Text. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2015 Presented at: SemEval'14; August 23-24, 2014; Dublin, Ireland p. a. [doi: 10.3115/v1/s14-2007]
40. Yang J, Liu Y, Qian M, Guan C, Yuan X. Information extraction from electronic medical records using multitask recurrent neural network with contextual word embedding. Appl Sci 2019 Sep 4;9(18):3658. [doi: 10.3390/app9183658]

Abbreviations

BERT: bidirectional encoder representations from transformers
BiLSTM: bidirectional long short-term memory
BIO: beginning, inside, and outside
CRF: conditional random fields
DL: deep learning
EHR: electronic health record
n2c2: national NLP clinical challenges
NEN: named entity normalization
NER: named entity recognition
NLP: natural language processing
NLTK: Natural Language Toolkit
RE: relation extraction

Edited by Y Wang; submitted 07.08.20; peer-reviewed by JA Benítez-Andrades, S Liu, R Kate; comments to author 25.09.20; revised version received 20.10.20; accepted 03.11.20; published 29.12.20.

Please cite as:
Silva JF, Almeida JR, Matos S
Extraction of Family History Information From Clinical Notes: Deep Learning and Heuristics Approach
JMIR Med Inform 2020;8(12):e22898
URL: http://medinform.jmir.org/2020/12/e22898/
doi:10.2196/22898
PMID:33372893

Original Paper

Growth of Ambulatory Virtual Visits and Differential Use by Patient Sociodemographics at One Urban Academic Medical Center During the COVID-19 Pandemic: Retrospective Analysis

Sarah F Gilson[1,2], MD; Craig A Umscheid[1,2], MD, MS; Neda Laiteerapong[1,2], MD, MS; Graeme Ossey[2,3], MBA; Kenneth J Nunes[4], MD; Sachin D Shah[1,2,5], MD

[1]Department of Medicine, University of Chicago, Chicago, IL, United States

[2]Center for Healthcare Delivery Science and Innovation, University of Chicago Medicine, Chicago, IL, United States

[3]Digital Health, University of Chicago Medicine, Chicago, IL, United States

[4]Department of Obstetrics & Gynecology, University of Chicago, Chicago, IL, United States

[5]Department of Pediatrics, University of Chicago, Chicago, IL, United States

Corresponding Author:
Sachin D Shah, MD
Department of Medicine
University of Chicago
5841 S Maryland Ave, MC 3051
Chicago, IL, 60637
United States
Phone: 1 773 834 8455
Email: sdshah@uchicago.edu

Abstract

Background: Despite widespread interest in the use of virtual (ie, telephone and video) visits for ambulatory patient care during the COVID-19 pandemic, studies examining their adoption during the pandemic by race, sex, age, or insurance are lacking. Moreover, there have been limited evaluations to date of the impact of these sociodemographic factors on the use of telephone versus video visits. Such assessments are crucial to identify, understand, and address differences in care delivery across patient populations, particularly those that could affect access to or quality of care.

Objective: The aim of this study was to examine changes in ambulatory visit volume and type (ie, in-person vs virtual and telephone vs video visits) by patient sociodemographics during the COVID-19 pandemic at one urban academic medical center.

Methods: We compared volumes and patient sociodemographics (age, sex, race, insurance) for visits during the first 11 weeks following the COVID-19 national emergency declaration (March 15 to May 31, 2020) to visits in the corresponding weeks in 2019. Additionally, for visits during the COVID-19 study period, we examined differences in visit type (ie, in-person versus virtual, and telephone versus video visits) by sociodemographics using multivariate logistic regression.

Results: Total visit volumes in the COVID-19 study period comprised 51.4% of the corresponding weeks in 2019 (n=80,081 vs n=155,884 visits). Although patient sociodemographics between the COVID-19 study period in 2020 and the corresponding weeks in 2019 were similar, 60.5% (n=48,475) of the visits were virtual, compared to 0% in 2019. Of the virtual visits, 61.2% (n=29,661) were video based, and 38.8% (n=18,814) were telephone based. In the COVID-19 study period, virtual (vs in-person) visits were more likely among patients with race categorized as other (vs White) and patients with Medicare (vs commercial) insurance and less likely for men, patients aged 0-17 years, 65-74 years, or ≥75 years (compared to patients aged 18-45 years), and patients with Medicaid insurance or insurance categorized as other. Among virtual visits, compared to telephone visits, video visits were more likely to be adopted by patients aged 0-17 years (vs 18-45 years), but less likely for all other age groups, men, Black (vs White) patients, and patients with Medicare or Medicaid (vs commercial) insurance.

Conclusions: Virtual visits comprised the majority of ambulatory visits during the COVID-19 study period, of which a majority were by video. Sociodemographic differences existed in the use of virtual versus in-person and video versus telephone visits. To ensure equitable care delivery, we present five policy recommendations to inform the further development of virtual visit programs and their reimbursement.

(JMIR Med Inform 2020;8(12):e24544) doi:10.2196/24544

KEYWORDS

telemedicine; telehealth; video visit; telephone visit; virtual visit; COVID-19; age; sex; race; insurance; demographic; retrospective

Introduction

The COVID-19 pandemic has significantly altered the landscape of health care delivery. One of the major changes resulting from the pandemic has been the rapid adoption of virtual (ie, telephone and video) visits and other telemedicine programs that facilitate health care services via health care information technologies to accommodate necessary reductions in in-person care [1,2]. A major driver for this adoption was the Centers for Medicare & Medicaid Services (CMS) expansion of virtual visit reimbursement on March 17, 2020, under the 1135 waiver authority. This allowed for Medicare reimbursement of multiple visit types performed virtually, including outpatient clinic visits, retroactively starting March 6, 2020, and continuing for the duration of the public health emergency [3]. This shift to reimburse virtual visits helped clinicians continue caring for patients despite widespread shelter-in-place orders and may represent the beginning of a new era for ambulatory medicine.

Unfortunately, access to virtual visits may not be equitable in the United States. Differential access to the internet and devices and differences in health literacy may leave patients without the ability to attend video visits. Thus, those patients may only be able to participate in telephone visits if they are unable to attend in-person visits. Surveys by the Pew Research Center in 2019 found lower rates of internet usage and smartphone ownership among people ages ≥65 years compared to younger adults [4,5]. When examining access to internet and internet technology by race, Black adults had lower rates of access to the internet and lower rates of desktop or laptop computer ownership than White adults [4,6]. A recent study of Medicare beneficiaries found that digital access was lowest among patients who were ≥85 years, Black, or received Medicaid [7]. Additionally, adults who are older, men, and Black have been shown to have lower health literacy levels than those who are younger, women, and White; and low health literacy is associated with a greater likelihood of needing help performing online tasks [8-10]. These disparities in access to the internet and devices and lower health literacy levels may lead to corresponding disparities in health care delivery and quality, particularly if the quality of health care visits and visit satisfaction are greater with video visits compared to telephone visits [11-13]. Furthermore, patients who opted out of virtual visits entirely and continued to attend in-person visits during the pandemic may have increased their risk of exposure to COVID-19 or experienced decreased appointment availability due to the decrease in in-person capacity required to maintain COVID-19 social distancing. Thus, though virtual visits have been considered an integral part of delivery of health care during the pandemic, access to those visits (especially video visits) may have been affected by underlying differences in access to technology and health literacy.

There is already existing evidence that other recent innovations in health care technology may exacerbate differences in health care access. For example, patient portal use, which has the potential to improve the quality and efficiency of health care delivery, differs with respect to race, insurance, and neighborhood broadband internet access [14]. One study found that patient portal use was lower among Black (vs White) patients; Medicare, Medicaid, and uninsured (vs commercially insured) patients; and patients with decreased neighborhood broadband internet access [14]. Other studies using data prior to the COVID-19 pandemic have additionally suggested that telemedicine and patient-facing health information technology utilization is lower among men, patients over 65 years, non-White patients, patients without commercial insurance, and patients living in neighborhoods with low internet access; this lack of internet access and technology proficiency continues to impede wider adoption of health information technology among racial minorities and those without commercial insurance [15-18]. Given prior research on the benefits of telemedicine interventions on clinical outcomes, such as improvement in glycemic control in medically underserved patients with diabetes, these disparities in the use of and access to digital health may directly translate into disparities in health care quality [19].

Despite widespread interest in the use of virtual visits for ambulatory patient care during the COVID-19 pandemic, few studies to date have evaluated the adoption of ambulatory virtual visits during the pandemic by age, race, sex, or insurance [20]. The studies that have been published recently show that patients using virtual visits during the pandemic were more likely to be younger adults as compared to older adults, female, non-White, and not commercially-insured [2,21-23]. This may be due in part to the lack of patient readiness for virtual visits, which one study found was more prevalent in patients who were older, male, or Black, and affected video visits more than telephone visits [24]. However, most of the studies published on data from the pandemic did not evaluate the impact of these sociodemographic factors on the use of telephone versus video virtual visits. Such assessments are crucial to identify, understand, and address differences in care delivery across patient populations, and inform policy decisions, particularly those like reimbursement rules, which could affect access to or quality of care.

In this study, we aimed to (1) assess changes in visit volume, type, and patient sociodemographics from the start of the COVID-19 national emergency to the end of May 2020, compared to the same weeks in 2019; and (2) elucidate differences in the use of ambulatory virtual visits (as compared to in-person visits) and, for those using virtual visits, the use of video visits (compared to telephone visits) by age, sex, race, and insurance. We hypothesize that (1) total visit volumes decreased and virtual visits increased during the COVID-19 pandemic, while patient sociodemographics remained similar between the two time periods; and (2) patients who utilized in-person visits during the COVID-19 study period were more

likely to be younger than patients who utilized virtual visits, and of those using virtual visits, patients utilizing video visits were more likely to be younger, White, and have commercial insurance than patients utilizing telephone visits [2,21-23].

Methods

Setting

The University of Chicago Medical Center (UCMC) is the flagship institution of University of Chicago Medicine, and includes 5 multispecialty faculty ambulatory practice sites in Chicago, IL, and the surrounding area, with over 600,000 encounters per year. UCMC began offering virtual visits in March 2020 in response to the widespread shelter-in-place orders at the city, state, and regional level due to the COVID-19 pandemic. Telephone visits began during the week of March 15, 2020. Video visits began with a pilot program in the hematology/oncology, pediatrics, psychiatry, gastroenterology, and obstetrics/gynecology practices on March 26, 2020, followed by a broad roll-out to all ambulatory faculty clinics on April 6, 2020. All practices used a HIPAA (Health Insurance Portability and Accountability Act)-compliant Zoom platform to enable video visits, which was not integrated into the institution's electronic health record system (Epic) during the evaluated time period.

Immediately after the City of Chicago and State of Illinois shelter-in-place orders were enacted, patients with previously scheduled in-person office visits were contacted and given the option to either reschedule or convert their appointment to a virtual visit. If a patient agreed to a virtual visit, a video visit was encouraged. Patients scheduled for video visits were sent the following through the patient portal or email: a Zoom link for the video visit; a brief prevideo visit checklist followed by more detailed instructions describing the technical requirements to participate in the video visit; and a link to a video highlighting methods to best prepare for the video visit and a demonstration of what to expect. If the patient was unable or unwilling to participate in a video visit, a telephone visit was scheduled, and they were told to expect a call from their provider at the scheduled appointment time. Patients reaching out to schedule new virtual visits were also preferentially offered video visits but were given the opportunity to schedule a telephone visit as well in accordance with their preferences. The availability of virtual visits was marketed widely to our patient population through our patient portal, marketing emails, and our health system's internet home page. Beginning on May 1, 2020, patients were given the option to begin self-scheduling video visits (but not telephone visits) through the patient portal.

Study Population and Measures

All adult and pediatric outpatient clinic visits occurring in UCMC faculty practice locations from March 15 to May 31, 2019, and March 15 to May 31, 2020, were included. The type of outpatient clinic visit was classified as in-person or virtual, and virtual visits were further classified as telephone or video, based on the scheduled visit type for all completed visits. Patient sociodemographic data were examined for each visit, including age, sex, race, and insurance. Age was categorized into 5 groups: 0-17 years, 18-45 years, 46-64 years, 65-74 years, and ≥75 years. Patients were grouped into 3 racial categories: White, Black, and other (which included Asian/Mideast Indian, American Indian or Alaska Native, Native Hawaiian/other Pacific Islander, more than one race, patient declined, and unknown). Insurance was categorized as Medicare (including Medicare-Medicaid Alignment Initiative), Medicaid, commercial, or other. The data were extracted from the institution's electronic health record data warehouse. This project received a formal determination of Quality Improvement according to institutional policy. As such, this initiative was not reviewed by the Institutional Review Board.

Statistical Analysis

First, we used descriptive statistics to examine weekly and overall visit volumes during the study period, which were the 11 weeks following the COVID-19 national emergency declaration (March 15 to May 31, 2020), compared to visit volumes in the corresponding weeks of the 2019 calendar year. Next, we examined visit type (in-person, video, telephone) and patient sociodemographics (age, sex, race, insurance) associated with the visit and compared these characteristics to those visits occurring during the same date range in 2019. Last, we examined differences in ambulatory visit type (in-person vs virtual; and for those with virtual visits, video vs telephone) by patient sociodemographics (age, sex, race, insurance) for visits occurring during the COVID-19 study period.

Data were summarized with chi-square tests where appropriate. Because of the large sample size, statistical significance was set at $P \leq .001$. To estimate the association between patient sociodemographics and visit type (in-person vs virtual, and video vs phone for those with virtual visits), we performed logistic regression. Results were similar between unadjusted and adjusted analyses; only adjusted analyses are presented. Data were analyzed using RStudio, version 3.6.3 (RStudio, PBC).

Results

Visit Volumes and Visit Types

In the week of March 15-21, 2020, the ambulatory visit volume dropped to 34% of visit volumes when compared to the same week in 2019 (n=4877 vs n=14,343 visits) and reached a nadir of 20.8% of 2019 levels (n=2476 vs n=11,930 visits) in the following week. By the week of May 24-30, 2020, the ambulatory visit volume had rebounded to 81.8% of the volume of the same week in 2019 (n=9451 vs n=11,554 visits). Total visit volumes from March 15 to May 31, 2020, were 51.4% of 2019 volumes (n=80,081 vs n=155,884 visits).

Virtual ambulatory visits increased from 0 to 48,475 visits between March 15 to May 31, 2020, and comprised 60.5% of total ambulatory visit volume, with the remaining 39.5% (n=31,606) conducted in person (Table 1 and Figure 1). Among virtual visits performed during the study period, 61.2% (n=29,661) were by video and 38.8% (n=18,814) were by telephone. For comparison, in 2019, there were no virtual visits for the same time period. Patient sociodemographics were similar for those with ambulatory visits between March 15 to May 31, 2020, and the corresponding weeks in 2019 (Table 1).

Table 1. Associations between patient sociodemographics and ambulatory visit type from March 15 to May 31 in 2019 and 2020.

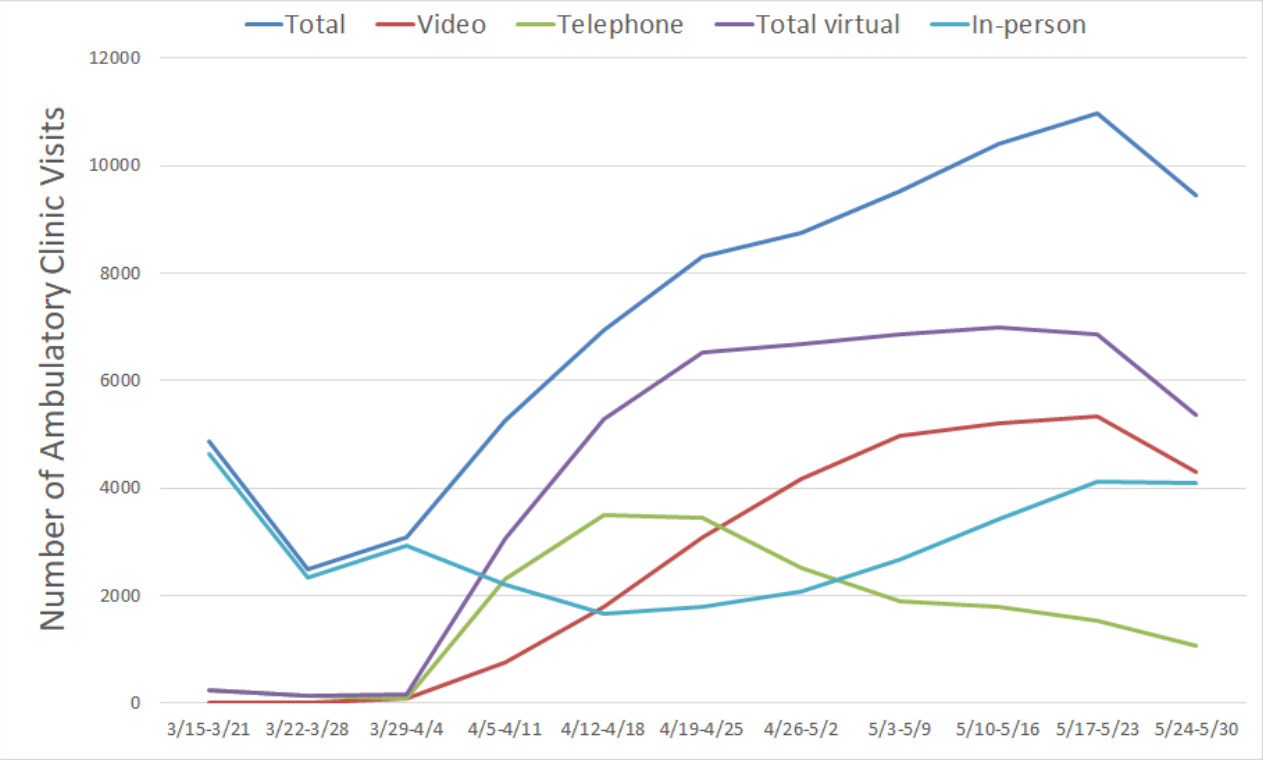
Characteristic	Total visits in 2019 (n=155,884), n (%)	Total visits in 2020 (n=80,081)				
		Overall (n=80,081), n (%)	In-person visits (n=31,606), n (%)	Virtual visits (n=48,475), n (%)	Virtual vs in-person	
					aOR[a] (95% CI)	*P* value[b]
Age (years)						<.001
0-17	20,513 (13.2)	10,085 (12.6)	4937 (15.6)	5148 (10.6)	0.71 (0.68-0.75)	
18-45	39,879 (25.6)	21,386 (26.7)	8192 (25.9)	13,194 (27.2)	Reference	
46-64	43,546 (27.9)	22,283 (27.8)	8455 (26.8)	13,828 (28.5)	1.01 (0.97-1.05)	
65-74	29,132 (18.7)	15,140 (18.9)	5957 (18.8)	9183 (19.0)	0.80 (0.76-0.84)	
≥75	22,814 (14.6)	11,187 (14.0)	4065 (12.9)	7122 (14.7)	0.86 (0.80-0.91)	
Sex						<.001
Female	95,032 (61.0)	48,571 (60.7)	18,429 (58.3)	30,142 (62.2)	Reference	
Male	—[c]	—	—	—	0.88 (0.85-0.90)	
Race						<.001
White	72,618 (46.6)	36,007 (45.0)	14,112 (44.7)	21,895 (45.2)	Reference	
Black	65,645 (42.1)	34,852 (43.5)	14,141 (44.7)	20,711 (42.7)	0.98 (0.95-1.01)	
Other	17,621 (11.3)	9222 (11.5)	3353 (10.6)	5869 (12.1)	1.22 (1.16-1.28)	
Insurance						<.001
Commercial	53,470 (34.3)	27,642 (34.5)	9817 (31.1)	17,825 (36.8)	Reference	
Medicare	23,663 (15.2)	11,620 (14.5)	5575 (17.6)	6045 (12.5)	1.27 (1.21-1.34)	
Medicaid	75,100 (48.2)	39,424 (49.2)	15,169 (48.0)	24,255 (50.0)	0.74 (0.70-0.77)	
Other	3651 (2.3)	1395 (1.8)	1045 (3.3)	350 (0.7)	0.21 (0.19-0.24)	

[a]aOR: adjusted odds ratio.

[b]Chi-square test.

[c]Not applicable.

Figure 1. Ambulatory visit volumes and types from March 15 to May 31, 2020. Note: all visit volumes decreased during the final week of May due to Memorial Day clinic closures.



Association Between Ambulatory Visit Type (In-Person vs Virtual) and Patient Sociodemographics

In unadjusted analyses, there were statistically significant differences between those who received in-person and virtual visits for all sociodemographics examined (Table 1). In adjusted analyses, virtual visits were less likely than in-person visits for patients aged 0-17 years (odds ratio [OR] 0.71, 95% CI 0.68-0.75), 65-74 years (OR 0.80, 95% CI 0.76-0.84), and ≥75 years (OR 0.86, 95% CI 0.80-0.91), compared to patients aged 18-45 years (Table 1). Men were less likely (OR 0.88, 95% CI 0.85-0.90) to attend a virtual visit than women. There was no difference in the odds of virtual visit attendance between White and Black patients; however, patients with race categorized as other were more likely to attend a virtual visit (OR 1.22, 95% CI 1.16-1.28) compared to White patients. Medicare patients were more likely (OR 1.27, 95% CI 1.21-1.34) than patients with commercial insurance to attend virtual visits (vs in-person visits), whereas patients with Medicaid insurance were less likely (OR 0.74, 95% CI 0.70-0.77) than patients with commercial insurance to have virtual visits. Patients with insurance categorized as other were also less likely to have a virtual visit (OR 0.21, 95% CI 0.19-0.24) than patients with commercial insurance.

Association Between Virtual Visit Type (Telephone vs Video) and Patient Sociodemographics for Those With Virtual Visits

In unadjusted analyses, there were statistically significant differences across all sociodemographics examined except sex between those using telephone versus video visits (Table 2). In adjusted analyses, results were similar, except there were differences by sex as well. Video visits were more likely than telephone visits for patients aged 0-17 years (OR 3.32, 95% CI 3.01-3.67), while video visits were less likely than telephone visits for patients aged 46-64 years (OR 0.56, 95% CI 0.54-0.60), 65-74 years (OR 0.47, 95% CI 0.44-0.50), and ≥75 years (OR 0.30, 95% CI 0.27-0.32), compared to patients aged 18-45 years. Men were less likely to attend a video visit (OR 0.94, 95% CI 0.90-0.97) than women. Black patients were less likely to attend a video visit (OR 0.55, 95% CI 0.52-0.57) compared to White patients. Video visits were less likely than telephone visits for Medicare patients (OR 0.69, 95% CI 0.65-0.74) and Medicaid patients (OR 0.72, 95% CI 0.67-0.77) compared to patients with commercial insurance.

Table 2. Associations between patient sociodemographics and type of virtual visit from March 15 to May 31, 2020.

Characteristic	Total virtual visits (n=48,475), n (%)	Virtual visits (n=48,475)			
		Telephone visits (n=18,814), n (%)	Video visits (n=29,661), n (%)	Video vs telephone	
				aOR[a] (95% CI)	*P* value[b]
Age (years)					<.001
0-17	5148 (10.6)	554 (2.9)	4594 (15.5)	3.32 (3.01-3.67)	
18-45	13,194 (27.2)	3507 (18.6)	9687 (32.7)	Reference	
46-64	13,828 (28.5)	5677 (30.2)	8151 (27.5)	0.56 (0.54-0.60)	
65-74	9183 (19)	4587 (24.4)	4596 (15.5)	0.47 (0.44-0.50)	
≥75	7122 (14.7)	4489 (23.9)	2633 (8.8)	0.30 (0.27-0.32)	
Sex					.17
Female	30,142 (62.2)	11,771 (62.6)	18,371 (61.9)	Reference	
Male	—[c]	—	—	0.94 (0.90-0.97)	
Race					<.001
White	21,895 (45.2)	7084 (37.7)	14,811 (49.9)	Reference	
Black	20,711 (42.7)	10,064 (53.4)	10,647 (35.9)	0.55 (0.52-0.57)	
Other	5869 (12.1)	1666 (8.9)	4203 (14.2)	0.95 (0.89-1.01)	
Insurance					<.001
Commercial	17,825 (36.8)	9846 (52.4)	7979 (26.9)	Reference	
Medicare	6045 (12.5)	2127 (11.3)	3918 (13.2)	0.69 (0.65-0.74)	
Medicaid	24,255 (50.0)	6741 (35.8)	17,514 (59.1)	0.72 (0.67-0.77)	
Other	350 (0.7)	100 (0.5)	250 (0.8)	1.03 (0.81-1.31)	

[a]aOR: adjusted odds ratio.

[b]Chi-square test.

[c]Not applicable.

Discussion

Principal Findings

Total visit volumes in the COVID-19 study period were approximately half of that in 2019, although patient sociodemographics were similar. Recovery of clinic volumes after the escalation of the pandemic was largely driven by virtual ambulatory care, which comprised over 60% (n=48,475) of total ambulatory clinic volumes from March 15 through May 31, 2020, a majority of which were video visits. Children, adults ≥65 years, men, and patients with Medicaid coverage were less likely to have virtual visits, whereas patients with Medicare coverage were more likely to have virtual visits compared to patients with commercial insurance coverage. For those who attended virtual visits, children were more likely to have video visits, while adults ≥46 years, men, Black patients, and patients with Medicare or Medicaid coverage were less likely to have video visits.

The sociodemographic differences in virtual visits we identified are in line with prior research. For example, prior research found that women were more likely than men to shelter in place due to concerns about the risk of COVID-19 infection for themselves and their family; this would make virtual visits a more appealing visit type for women [25]. Additionally, studies prior to the pandemic demonstrated that women used virtual visits more often than men [11]. Similarly, patients with Medicare insurance may have been more concerned about acquiring COVID-19 infection and prefer to shelter in place, leading to their increased likelihood of attending a virtual visit. In contrast, pediatric well visits (and well visits for most non-Medicare beneficiaries) must still be performed in person to be reimbursed; therefore, many pediatric patients continued to attend in-person visits even during the COVID-19 pandemic.

The sociodemographic differences in virtual (vs in-person) visits and video (vs telephone) visits illustrate the digital divide [26]. The patient populations with lower levels of access to internet and smart devices and lower digital literacy were the same sociodemographic groups found in our study to have a lower likelihood of completing virtual or video visits, including older adults, Black patients, and patients without commercial insurance [4-9]. Our results also match prior studies on virtual visit use during the pandemic, which found that patients using virtual visits during the pandemic were more likely to be

younger adults as compared to older adults, White, and commercially insured [21-23]. Requirements for a video visit include internet, a capable device, and a basic level of digital literacy, so patients who do not have all three (or do not have a readily available family member to assist) are unable to attend video visits. One study performed during the pandemic found higher prevalence of "unreadiness" to attend video visits in those sociodemographic groups found to be less likely to attend video visits, including patients who were older, Black, and men, similar to our findings [24]. These findings raise concerns about the role video visits may play in exacerbating existing health inequities, particularly since the quality of health care visits and visit satisfaction are greater with video visits compared to telephone visits [11-13]. Moreover, these health disparities may be significantly worsened if the current reimbursement parity between telephone and video visits is discontinued, and especially if telephone visits are no longer reimbursed altogether following the public health emergency.

The shift in the delivery of ambulatory care through virtual visits was incentivized by the new virtual reimbursement policies from CMS and private insurance companies. The significant contribution of virtual visits to overall ambulatory visit volumes is likely to continue once the COVID-19 pandemic has ended. The volume of virtual ambulatory visits at UCMC has continued to grow even after the end of the study period, indicating sustained interest in virtual visits likely due to continued safety concerns related to the pandemic, ongoing reimbursement for these services, and physician and patient satisfaction with this new option for care delivery [27,28]. Given the interest in and development of virtual visits prior to the pandemic and the proliferation of virtual visits during the pandemic, virtual visits for ambulatory care are likely to remain popular among both patients and providers even after the COVID-19 pandemic [1,2]. University of Chicago Medicine's 2025 Strategic Vision (developed prior to the pandemic) includes an "aim to build a digitally enabled organization for patients" and a goal to expand access to care, both of which are aided by the expansion of virtual visit services [29]. However, if reimbursement for virtual visits is discontinued or significantly reduced after the pandemic or public health emergency ends, many medical centers are likely to stop making significant investments in the continued development of their telemedicine programs and the availability of virtual visits for patients would be expected to decline.

Recommendations

The results of this study and our review of the virtual visit landscape has prompted us to offer five recommendations (Textbox 1). First, given the differences in virtual visit use by certain sociodemographic groups demonstrated in this study and the lower effective reimbursement rates for telephone visits compared to video visits, medical institutions like UCMC with high proportions of older, Black, and/or Medicare/Medicaid patients may experience lower reimbursement rates because of the barriers these groups face to completing video visits. For a video visit, providers can bill for all time spent on patient care on the encounter date, including documentation; for a telephone visit, they can only bill for time spent in direct communication (on the telephone call) with a patient on the encounter date. To avoid effectively penalizing medical institutions providing care to vulnerable populations, government and commercial insurers should help address these disparities by *maintaining reimbursement parity between video and telephone visits*. Second, given the rapid growth and early success of virtual visits, and the role they will likely play in blended models of care, *legislation that makes virtual visit reimbursement permanent* is essential to allow for the long-term investment by health care systems and providers needed to improve the virtual visit infrastructure and experience. Third, government insurers and specialty societies should collaborate to *establish guidance to help distinguish ambulatory care best suited for virtual versus in-person care*. Fourth, quality improvement initiatives should be undertaken at medical institutions to *support and improve access to and usability of video visits* in populations encountering the greatest barriers to its use. Last, *advocacy for policy changes and more universal broadband access* are essential to help close the digital divide experienced by our most vulnerable patient populations, which would help address the differential access to virtual visits described in this study.

Textbox 1. Recommendations to improve access to and use of virtual visits.

1. Maintain reimbursement parity between video and telephone visits
2. Pass legislation making virtual visit reimbursement permanent
3. Establish guidance to distinguish ambulatory care best suited for virtual versus in-person care
4. Perform quality improvement initiatives to improve access to and usability of video visits in vulnerable populations
5. Advocate for policy changes and universal broadband access to close the digital divide

Limitations

Our study has limitations. First, this study only examined a single medical center and was a retrospective analysis; despite this, the diversity of the patient population examined in our study enabled our analysis of ambulatory virtual visit use. Second, our study only examined a limited set of variables, which were used as surrogates for the social determinants of health described in this paper, such as access to broadband internet, health literacy, tech literacy, education, and income, and did not examine virtual and video visit use by ethnicity due to limited data availability. Third, this area of clinical practice and study is rapidly changing and will likely continue to change rapidly over the next few months to years. Further studies at other medical institutions should be conducted to confirm our findings and examine additional sociodemographic variables. Future analyses of ambulatory virtual visits should also investigate patient satisfaction and outcomes by patient visit

type (eg, new, return, consult), given the differences in reimbursement by visit type category, and whether ambulatory virtual visits increase the geographic area served by academic medical centers or medical institutions with subspecialty care, as already suggested by limited data [30].

Conclusion

The COVID-19 pandemic has drastically changed the health care delivery landscape largely due to the growth of ambulatory virtual visits, which have rapidly become a vital component of health care delivery. Given the differential use of these technologies by age, sex, race, and insurance, these changes also risk perpetuating and even exacerbating existing disparities in health care access and quality, especially if reimbursement policies do not sufficiently account for these differences and the digital divide remains unaddressed.

Acknowledgments

The authors would like to thank Megan Huisingh-Scheetz, Michael Cui, and Omar Jamil for their input regarding this study; Whitney Westphal and Mary Kate Springman for their assistance in obtaining and cleaning data for this study; and Mengqi Zhu for her review of analytic methods.

Conflicts of Interest

None declared.

References

1. Mehrotra A, Ray K, Brockmeyer DM, Barnett ML, Bender JA. Rapidly Converting to “Virtual Practices”: Outpatient Care in the Era of Covid-19. NEJM Catalyst. 2020 Apr 1. URL: https://catalyst.nejm.org/doi/full/10.1056/CAT.20.0091 [accessed 2020-05-25]
2. Mann DM, Chen J, Chunara R, Testa PA, Nov O. COVID-19 transforms health care through telemedicine: Evidence from the field. J Am Med Inform Assoc 2020 Jul 01;27(7):1132-1135 [FREE Full text] [doi: 10.1093/jamia/ocaa072] [Medline: 32324855]
3. Medicare Telemedicine Health Care Provider Fact Sheet. Centers for Medicare & Medicaid Services. URL: https://www.cms.gov/newsroom/fact-sheets/medicare-telemedicine-health-care-provider-fact-sheet [accessed 2020-05-31]
4. Anderson M, Perrin A, Jiang J, Kumar M. 10% of Americans don't use the internet. Who are they? Pew Research Center. 2019 Apr 22. URL: https://www.pewresearch.org/fact-tank/2019/04/22/some-americans-dont-use-the-internet-who-are-they/ [accessed 2020-05-25]
5. Anderson M, Perrin A. Technology use among seniors. Pew Research Center. 2017 May 17. URL: https://www.pewresearch.org/internet/2017/05/17/technology-use-among-seniors/ [accessed 2020-06-02]
6. Perrin A, Turner E. Smartphones help blacks, Hispanics bridge some – but not all – digital gaps with whites. Pew Research Center. 2020 Aug 20. URL: https://www.pewresearch.org/fact-tank/2019/08/20/smartphones-help-blacks-hispanics-bridge-some-but-not-all-digital-gaps-with-whites/ [accessed 2020-05-26]
7. Roberts ET, Mehrotra A. Assessment of Disparities in Digital Access Among Medicare Beneficiaries and Implications for Telemedicine. JAMA Intern Med 2020 Oct 01;180(10):1386-1389. [doi: 10.1001/jamainternmed.2020.2666] [Medline: 32744601]
8. Cajita MI, Cajita TR, Han H. Health Literacy and Heart Failure: A Systematic Review. J Cardiovasc Nurs 2016;31(2):121-130 [FREE Full text] [doi: 10.1097/JCN.0000000000000229] [Medline: 25569150]
9. Choi NG, Dinitto DM. The digital divide among low-income homebound older adults: Internet use patterns, eHealth literacy, and attitudes toward computer/Internet use. J Med Internet Res 2013 May 02;15(5):e93 [FREE Full text] [doi: 10.2196/jmir.2645] [Medline: 23639979]
10. Vollbrecht H, Arora V, Otero S, Carey K, Meltzer D, Press VG. Evaluating the Need to Address Digital Literacy Among Hospitalized Patients: Cross-Sectional Observational Study. J Med Internet Res 2020 Jun 04;22(6):e17519 [FREE Full text] [doi: 10.2196/17519] [Medline: 32496196]
11. Handschu R, Scibor M, Willaczek B, Nückel M, Heckmann JG, Asshoff D, STENO Project. Telemedicine in acute stroke: remote video-examination compared to simple telephone consultation. J Neurol 2008 Nov;255(11):1792-1797. [doi: 10.1007/s00415-008-0066-9] [Medline: 19156491]
12. Richter KP, Shireman TI, Ellerbeck EF, Cupertino AP, Catley D, Cox LS, et al. Comparative and cost effectiveness of telemedicine versus telephone counseling for smoking cessation. J Med Internet Res 2015 May 08;17(5):e113 [FREE Full text] [doi: 10.2196/jmir.3975] [Medline: 25956257]
13. Hammersley V, Donaghy E, Parker R, McNeilly H, Atherton H, Bikker A, et al. Comparing the content and quality of video, telephone, and face-to-face consultations: a non-randomised, quasi-experimental, exploratory study in UK primary care. Br J Gen Pract 2019 Sep;69(686):e595-e604 [FREE Full text] [doi: 10.3399/bjgp19X704573] [Medline: 31262846]
14. Perzynski AT, Roach MJ, Shick S, Callahan B, Gunzler D, Cebul R, et al. Patient portals and broadband internet inequality. J Am Med Inform Assoc 2017 Sep 01;24(5):927-932 [FREE Full text] [doi: 10.1093/jamia/ocx020] [Medline: 28371853]

15. Ganguli I, Orav EJ, Lupo C, Metlay JP, Sequist TD. Patient and Visit Characteristics Associated With Use of Direct Scheduling in Primary Care Practices. JAMA Netw Open 2020 Aug 03;3(8):e209637 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.9637] [Medline: 32852551]
16. Reed ME, Huang J, Graetz I, Lee C, Muelly E, Kennedy C, et al. Patient Characteristics Associated With Choosing a Telemedicine Visit vs Office Visit With the Same Primary Care Clinicians. JAMA Netw Open 2020 Jun 01;3(6):e205873 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.5873] [Medline: 32585018]
17. Walker DM, Hefner JL, Fareed N, Huerta TR, McAlearney AS. Exploring the Digital Divide: Age and Race Disparities in Use of an Inpatient Portal. Telemed J E Health 2020 May;26(5):603-613. [doi: 10.1089/tmj.2019.0065] [Medline: 31313977]
18. Ackerman SL, Sarkar U, Tieu L, Handley MA, Schillinger D, Hahn K, et al. Meaningful use in the safety net: a rapid ethnography of patient portal implementation at five community health centers in California. J Am Med Inform Assoc 2017 Sep 01;24(5):903-912 [FREE Full text] [doi: 10.1093/jamia/ocx015] [Medline: 28340229]
19. Heitkemper EM, Mamykina L, Travers J, Smaldone A. Do health information technology self-management interventions improve glycemic control in medically underserved adults with diabetes? A systematic review and meta-analysis. J Am Med Inform Assoc 2017 Sep 01;24(5):1024-1035 [FREE Full text] [doi: 10.1093/jamia/ocx025] [Medline: 28379397]
20. Hong Y, Lawrence J, Williams D, Mainous I. Population-Level Interest and Telehealth Capacity of US Hospitals in Response to COVID-19: Cross-Sectional Analysis of Google Search and National Hospital Survey Data. JMIR Public Health Surveill 2020 Apr 07;6(2):e18961 [FREE Full text] [doi: 10.2196/18961] [Medline: 32250963]
21. Jaffe DH, Lee L, Huynh S, Haskell TP. Health Inequalities in the Use of Telehealth in the United States in the Lens of COVID-19. Popul Health Manag 2020 Oct 01;23(5):368-377. [doi: 10.1089/pop.2020.0186] [Medline: 32816644]
22. Nouri S, Khoong EC, Lyles CR, Karliner L. Addressing Equity in Telemedicine for Chronic Disease Management During the Covid-19 Pandemic. NEJM Catalyst. 2020 May 4. URL: https://catalyst.nejm.org/doi/full/10.1056/CAT.20.0123 [accessed 2020-11-03]
23. Thronson LR, Jackson SL, Chew LD. The Pandemic of Health Care Inequity. JAMA Netw Open 2020 Oct 01;3(10):e2021767 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.21767] [Medline: 33006616]
24. Lam K, Lu AD, Shi Y, Covinsky KE. Assessing Telemedicine Unreadiness Among Older Adults in the United States During the COVID-19 Pandemic. JAMA Intern Med 2020 Oct 01;180(10):1389-1391. [doi: 10.1001/jamainternmed.2020.2671] [Medline: 32744593]
25. Frederiksen B, Gomez I, Salganicoff A, Ranji U. Coronavirus: A Look at Gender Differences in Awareness and Actions. Kaiser Family Foundation. 2020 Mar 20. URL: https://www.kff.org/coronavirus-covid-19/issue-brief/coronavirus-a-look-at-gender-differences-in-awareness-and-actions/ [accessed 2020-07-27]
26. Ramsetty A, Adams C. Impact of the digital divide in the age of COVID-19. J Am Med Inform Assoc 2020 Jul 01;27(7):1147-1148 [FREE Full text] [doi: 10.1093/jamia/ocaa078] [Medline: 32343813]
27. Polinski JM, Barker T, Gagliano N, Sussman A, Brennan TA, Shrank WH. Patients' Satisfaction with and Preference for Telehealth Visits. J Gen Intern Med 2016 Mar;31(3):269-275 [FREE Full text] [doi: 10.1007/s11606-015-3489-x] [Medline: 26269131]
28. Donelan K, Barreto E, Sossong S, Michael C, Estrada JJ, Cohen AB, et al. Patient and clinician experiences with telehealth for patient follow-up care. Am J Manag Care 2019 Jan;25(1):40-44. [Medline: 30667610]
29. Vision 2025: Advancing the Forefront. UChicago Medicine. URL: https://tinyurl.com/y2z5wpw6 [accessed 2020-06-02]
30. Ray KN, Mehrotra A, Yabes JG, Kahn JM. Telemedicine and Outpatient Subspecialty Visits Among Pediatric Medicaid Beneficiaries. Acad Pediatr 2020 Jul 08;20(5):642-651 [FREE Full text] [doi: 10.1016/j.acap.2020.03.014] [Medline: 32278078]

Abbreviations

CMS: Centers for Medicare & Medicaid Services
HIPAA: Health Insurance Portability and Accountability Act
OR: odds ratio
UCMC: University of Chicago Medical Center

Edited by C Lovis; submitted 24.09.20; peer-reviewed by J Hefner, D Kaelber; comments to author 24.10.20; revised version received 11.11.20; accepted 15.11.20; published 04.12.20.

Please cite as:
Gilson SF, Umscheid CA, Laiteerapong N, Ossey G, Nunes KJ, Shah SD
Growth of Ambulatory Virtual Visits and Differential Use by Patient Sociodemographics at One Urban Academic Medical Center During the COVID-19 Pandemic: Retrospective Analysis
JMIR Med Inform 2020;8(12):e24544
URL: https://medinform.jmir.org/2020/12/e24544
doi:10.2196/24544
PMID:33191247

Original Paper

Global Infectious Disease Surveillance and Case Tracking System for COVID-19: Development Study

Hsiu-An Lee[1,2,3,4,5], MS; Hsin-Hua Kung[2,3,4,5], MS; Yuarn-Jang Lee[6], PhD; Jane C-J Chao[7,8], PhD; Jai Ganesh Udayasankaran[3,4,9], MSc, MBA; Hueng-Chuen Fan[10,11,12], PhD; Kwok-Keung Ng[3,13], BS; Yu-Kang Chang[10,11], PhD; Boonchai Kijsanayotin[3,4,14], MSc, MD, PhD; Alvin B Marcelo[3,4,15], MD; Chien-Yeh Hsu[2,3,4,5,8], PhD

[1]Department of Computer Science and Information Engineering, Tamkang University, New Taipei, Taiwan
[2]Taiwan e-Health Association, Taipei, Taiwan
[3]Asia eHealth Information Network, Hong Kong, Hong Kong
[4]Standards and Interoperability Lab, Smart Healthcare Center of Excellence, Taipei, Taiwan
[5]Department of Information Management, National Taipei University of Nursing and Health Sciences, Taipei, Taiwan
[6]Division of Infection Diseases, Department of Internal Medicine, Taipei Medical University Hospital, Taipei, Taiwan
[7]Nutrition Research Center, Taipei Medical University Hospital, Taipei, Taiwan
[8]Master Program in Global Health and Development, College of Public Health, Taipei Medical University, Taipei, Taiwan
[9]Sri Sathya Sai Central Trust, Prasanthi Nilayam, Puttaparthi, India
[10]Department of Rehabilitation, Jen-Teh Junior College of Medicine, Nursing and Management, Miaoli, Taiwan
[11]Department of Medical Research, Tung's Taichung Metroharbor Hospital, Taichung, Taiwan
[12]Department of Pediatrics, Tung's Taichung Metroharbor Hospital, Taichung, Taiwan
[13]eHealth Research Institute, Hong Kong, Hong Kong
[14]Thai Health Information Standards Development Center, Health System Research Institute, Ministry of Public Health, Bangkok, Thailand
[15]University of the Philippines, Manila, Philippines

Corresponding Author:
Chien-Yeh Hsu, PhD
Department of Information Management
National Taipei University of Nursing and Health Sciences
365 Ming-te Road
Peitou District
Taipei, 11219
Taiwan
Phone: 886 28227101
Email: cyhsu@ntunhs.edu.tw

Abstract

Background: COVID-19 has affected more than 180 countries and is the first known pandemic to be caused by a new virus. COVID-19's emergence and rapid spread is a global public health and economic crisis. However, investigations into the disease, patient-tracking mechanisms, and case report transmissions are both labor-intensive and slow.

Objective: The pandemic has overwhelmed health care systems, forcing hospitals and medical facilities to find effective ways to share data. This study aims to design a global infectious disease surveillance and case tracking system that can facilitate the detection and control of COVID-19.

Methods: The International Patient Summary (IPS; an electronic health record that contains essential health care information about a patient) was used. The IPS was designed to support the used case scenario for unplanned cross-border care. The design, scope, utility, and potential for reuse of the IPS for unplanned cross-border care make it suitable for situations like COVID-19. The Fast Healthcare Interoperability Resources confirmed that IPS data, which includes symptoms, therapies, medications, and laboratory data, can be efficiently transferred and exchanged on the system for easy access by physicians. To protect privacy, patient data are deidentified. All systems are protected by blockchain architecture, including data encryption, validation, and exchange of records.

Results: To achieve worldwide COVID-19 surveillance, a global infectious disease information exchange must be enacted. The COVID-19 surveillance system was designed based on blockchain architecture. The IPS was used to exchange case study information among physicians. After being verified, physicians can upload IPS files and receive IPS data from other global cases. The system includes a daily IPS uploading and enhancement plan, which covers real-time uploading through the interoperation of the clinic system, with the module based on the Open Application Programming Interface architecture. Through the treatment of different cases, drug treatments, and the exchange of treatment results, the disease spread can be controlled, and treatment methods can be funded. In the Infectious Disease Case Tracking module, we can track the moving paths of infectious disease cases. The location information recorded in the blockchain is used to check the locations of different cases. The Case Tracking module was established for the Centers for Disease Control and Prevention to track cases and prevent disease spread.

Conclusions: We created the IPS of infectious diseases for physicians treating patients with COVID-19. Our system can help health authorities respond quickly to the transmission and spread of unknown diseases, and provides a system for information retrieval on disease transmission. In addition, this system can help researchers form trials and analyze data from different countries. A common forum to facilitate the mutual sharing of experiences, best practices, therapies, useful medications, and clinical intervention outcomes from research in various countries could help control an unknown virus. This system could be an effective tool for global collaboration in evidence-based efforts to fight COVID-19.

(JMIR Med Inform 2020;8(12):e20567) doi:10.2196/20567

KEYWORDS

blockchain; infectious disease surveillance; international collaboration; HL7 FHIR; COVID-19 defense; COVID-19

Introduction

COVID-19, which presumably originated in bats and was transmitted to humans by means of unknown mechanisms in Wuhan, Hubei Province, China in December 2019, has affected more than 180 countries and territories around the world. On March 11, 2020, the World Health Organization (WHO) characterized the COVID-19 outbreak as a pandemic. This is the first pandemic known to be caused by a new virus. Although the complete clinical picture with regard to COVID-19 is not fully known, based on currently available information, older adults and people with serious underlying medical conditions might be at a higher risk for the severe illness caused by COVID-19.

Since a total of 41 cases with an unknown etiology of pneumonia were confirmed in Wuhan City, Hubei Province, China in December 2019 [1], COVID-19 has spread rapidly across that country and around the world [2-8]. Thus far, it has affected more than 12,723,798 people in 188 countries and regions (data obtained through July 12, 2020) [9]. COVID-19 is now the most serious infectious disease event after severe acute respiratory syndrome (SARS) in 2003, and no effective vaccine, drug, or treatment has been found.

Many different infectious diseases still exist in the world, such as the Ebola hemorrhagic fever, the highly pathogenic avian influenza, SARS, Middle East respiratory syndrome (MERS)–related coronavirus, and seasonal influenza. When an infectious disease event occurs suddenly, it is crucial to find a quick treatment and control method. Normal patient treatment needs to be based on the medical history and symptoms of the different cases.

The rise of COVID-19 was sudden and marked by the global information flow not being fast enough and the case reports being transmitted slowly, which has led to a sluggish treatment progress, patients not being cured in an efficient manner, and the infectious disease still not being effectively controlled. In today's age of information, our global connectivity gives us a strong advantage in the fight against infectious diseases. We can analyze large amounts of data to identify outbreaks across different parts of the world, and we can use advanced machine learning models to predict their future movement across different geographical territories. The challenge is that collating relevant data and standardizing it on a global level is a complicated task. In many parts of the world, data does not flow easily from hospitals into the public realm or across borders. Global data standards have yet to be developed, and this creates gaps in the data sets and delays in how the data can be used to shape global health efforts. One way of improving the speed that data is standardized could be to encourage better interconnectivity across national data systems by using more homogenous data standards. This would require a great deal of collaboration between the various stakeholders, and it could be challenging to promote it across borders [10].

The challenge of a slow and insufficient global information flow could be tackled by a good framework such as the Asia eHealth Information Network's Governance, Architecture, Program Management, Standards and Interoperability framework as well as a good collaboration model.

According to different research case reports in China [2,5,6], of the patients who are in the 18 years and older group, 61.9% (n=172) were male, and in another report, 2 of 13 patients with COVID-19 were children, who ranged between 2 and 15 years old [11]. Conclusions of the symptoms and disease history of patients with COVID-19 were found in these studies. Hypertension and cardiovascular disease were the two most common diseases in the adult patient group, followed by diabetes mellitus. With regard to the symptoms, fever was the most common (n=28, 92.8%), followed by a cough (n=194, 69.8%), dyspnea (n=96, 34.5%), myalgia (n=77, 27.7%), a headache (n=20, 7.2%), diarrhea (n=17, 6.1%), a sore throat (5.1% [6]), and pharyngeal (17.4% [2]). Wang et al [2] showed that the intensive care rate was significant in older patients. Other research noted that patients who needed intensive care

had a greater percentage of dyspnea than those not needing intensive care [2,5]. From a report presented by a Beijing research team, among 13 patients with COVID-19, 12 (92.3%) had a fever, with a mean of 1.6 days before the patient went to a hospital, and they had a cough (46.3%), myalgia (23.1%), upper airway congestion (61.5%), and a headache (23.1%) [11].

Although there are many reports and studies on COVID-19, the details of disease control and treatment are still being broadcast slowly, which may cause the disease spread to be out of control and make it difficult to share the experiences of successful case treatments. According to the control status and experience of COVID-19, all cases should be uploaded to the WHO website by different governments, but the route of transmission is still difficult to track, and treatment experiences in different countries cannot be effectively shared. A literature review of infectious disease surveillance, presented by Jajosky and Groseclose [12], and an analysis of the timeliness of reporting by the National Notifiable Diseases Surveillance System showed that longer reporting lags and the variability among the states limit its usefulness. Some systems have the function of being a static continuous spatial map of infectious disease risk, while others have the function of continuously updating the reporting of infectious diseases, but there is still no system that combines these two functions [13].

After the rise of COVID-19, the problem has developed into the pathogenic spread across, and among, nations by means of international travel, which has unfortunately enabled the pathogens to invade new countries and adapt to new environments and hosts faster [14,15]. In many countries where the public health infrastructure is poor or where there is an insufficient budget to develop it, the ability of electronic disease surveillance, including data collection and an analysis capability, should be improved [16,17]. Furthermore, the data exchange of international infectious disease reports and information has certain constraints, not only out of fear for the repercussions on trade and tourism but also because of the delays in data transfer through the multiple levels of governments or organizations [18]. After experiencing epidemic infectious diseases caused by mutant viruses such as SARS and MERS, we have found that, when facing treatment for unknown diseases, related health organizations and authorities should conduct comprehensive tests, using different drugs and treatment methods, and they should then present the differences between each case and the analyzed treatment results to find the best treatment. However, this process is tedious and dangerous, and it creates uncertainties regarding patient treatment. In the face of new infectious diseases, the exchange of treatment results and case experiences is critical.

When facing a new type of infectious disease, it is important not only to treat the disease but also to prevent its contagion. For example, hundreds of COVID-19 cases in South Korea were found to have occurred at the same church. Hundreds of cases in Japan were found to have originated on a cruise ship. In Hong Kong, several cases were found to have been infected through a hot pot meal. Iran's speedy and large-scale infection may be due to specific types of religious behavior. In Italy, the outbreak may have been caused by the Italian culture, where hugs and kisses are a common way of greeting someone. During the SARS outbreak in 2003, it was found that infections were caused by the drainage designs of high-rise buildings [19]. Information on the correlation between the context of the event, living, transportation or environmental design, religion, and cultural behavior is critical for studying COVID-19 transmission.

To understand the epidemiology and trends of COVID-19, the WHO has provided a template for a case-based reporting form and a data dictionary for that case-based reporting form, and it has requested member countries to report probable and confirmed cases of COVID-19 infection within 48 hours of their identification [20]. These reports are sent through the National Focal Point and the Regional Contact Point for International Health Regulations at the appropriate WHO regional office. The WHO has asked the countries to provide aggregated data for surveillance when it is not feasible to report case-based data.

However, to the best of our knowledge, there has thus far been no functional collaborative global case exchange model that can cocreate case data on COVID-19 and facilitate care coordination across countries. The aim of this study is to design an infectious disease surveillance module for the global exchange of infectious cases and the sharing of treatment experience. Information on the movement and path tracking of cases, including the linkage and correlation between each case, can also be included in infectious disease control in various countries. Therefore, when an infectious disease outbreak occurs, it can be quickly controlled.

In the initial stages of the COVID-19 outbreak, little research was available on the data format of the disease, and no one knew what the best data format was; there had only been some discussions on the importance of clinical data exchange regarding the disease.

Currently, several places have created a Fast Healthcare Interoperability Resources (FHIR)–based COVID-19 data structure. A good example is provided by the National Coordinator for Health Information Technology in its Interoperability Standards Advisory section of Interoperability for the COVID-19 Novel Coronavirus Pandemic [21,22], namely, the Logica COVID-19 (FHIR v4.0.1) Implementation Guide CI Build. The Logica used Health Level 7 (HL7) FHIR profiles for COVID-19 to create an implementation guide for a collection or library of data elements that relate to COVID-19. This can be used in many different situations where COVID-19 data are shared to support patient care, billing, research, or public reporting.

Another example can be found in the Dedalus COVID-19 Solution [23]. In their "COVID-19 Simplifier Project," they used FHIR resources in the Dedalus COVID-19 Solution software. The data elements cover a patient self-assessment, a remote clinical assessment, and telemedicine and self-monitoring. They claimed that their first activations will be in Italy and France. Our study uses a similar method that started from COVID-19–related clinical data, and we used the International Patient Summary (IPS) as a basis for the data structure. The IPS document is an electronic health record (EHR) extract that contains essential health care information for the necessary care of patients. Due to the rapid outbreak of the disease in the early weeks, no format had been designed for

the exchange of COVID-19 data. Therefore, we designed a version of the IPS that can be used for COVID-19.

An IPS document is an EHR extract that contains essential health care information about a patient [24]. It is designed to support the used case scenario for *unplanned, cross-border care*, but it is not limited to that. It is intended to be international (ie, to provide generic solutions for global application beyond a particular region or country), and the IPS data set is minimal and nonexhaustive, specialty agnostic, and condition independent yet still clinically relevant. The design, global scope, and utility of IPS toward unplanned cross-border care, and its potential for reuse, make it suitable for a situation like COVID-19. The FHIR confirmed that IPS, including the symptoms, therapies, medications, and laboratory data, can be efficiently transferred and exchanged on the system for easy access by physicians. Patient data are deidentified to protect their privacy. In addition, the blockchain-based architecture can be used to ensure the security and immutability of the case data.

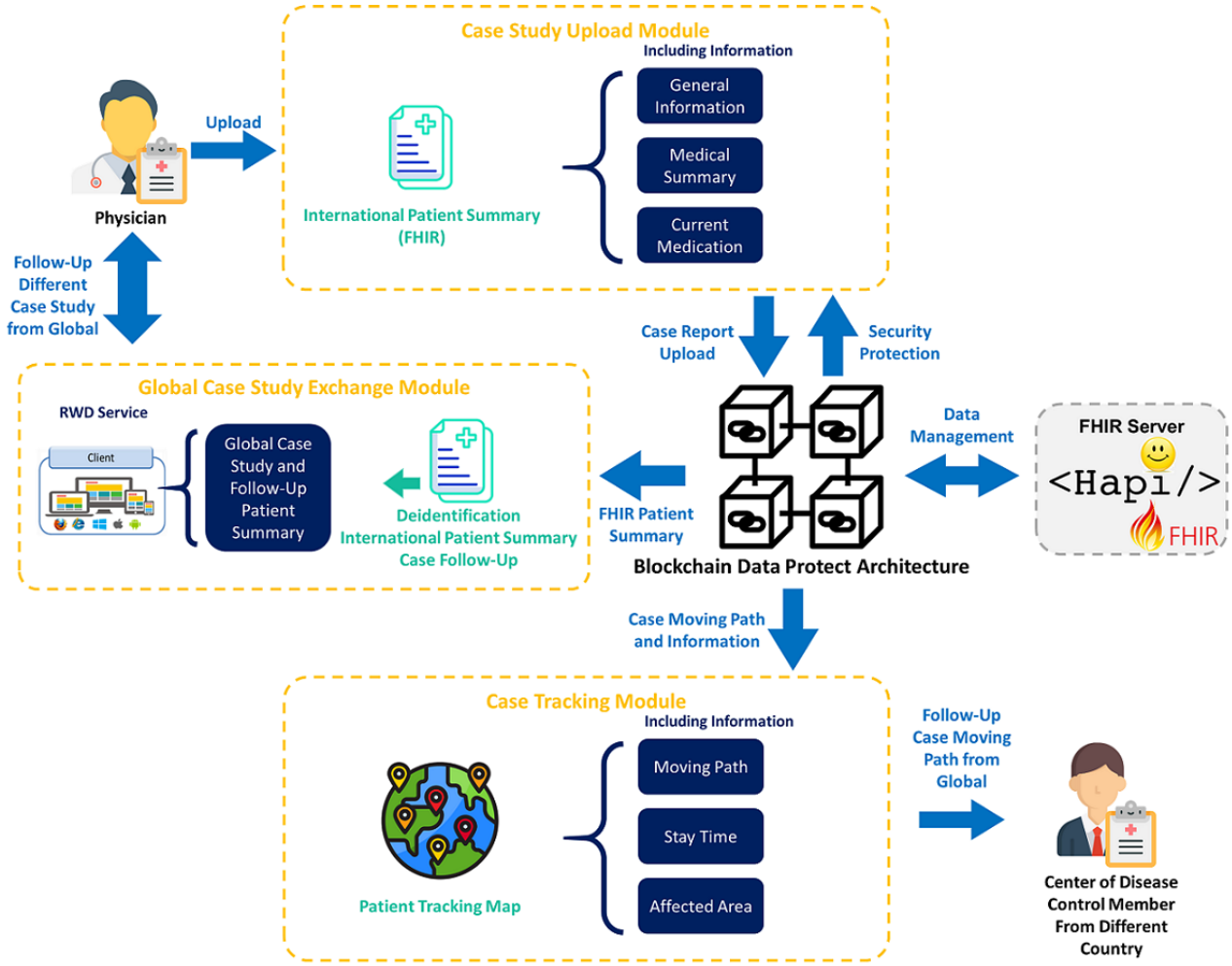
Our goal is to provide an immediate reference for people to use in the current crisis, so the design is not focused on a single use case, and the IPS therefore has a more general data structure that focuses on the clinical data needed for COVID-19. We understand that the data structure will not be perfect or comprehensive, but it can be modified in the future after more and more institutions use the data structure to exchange records. According to the research of Holmgren et al [25], the inability of hospitals to receive electronic data is an obstacle for the effective monitoring of patient symptoms. Therefore, the aim of our study is to create a COVID-19 data structure and a system that can share the data among health care institutions. It is expected that the proposed system can contribute to the control of the COVID-19 situation.

Methods

Architecture for the Global Infectious Disease Surveillance and Case Tracking Model

This study designs a global infectious disease surveillance and case-tracking model, and it includes a "Case Study Upload Module," a "Global Case Study Exchange Module," and a "Case Tracking Module." Each module has different goals. The architecture of the global infectious disease surveillance and case-tracking model is shown in Figure 1.

Figure 1. Architecture for the Global Infectious Disease Surveillance and Case Tracking module. FHIR: Fast Healthcare Interoperability Resources; RWD: Responsive Web Design.



The main goal of the "Case Study Upload Module" is to allow physicians worldwide to continuously upload the IPS documents and to include detailed information about the treatment of patients with COVID-19. Through sharing experience and

patient summaries with other physicians, they can find better essential treatment methods. This module has the ability to identify and verify the identity of physicians in different countries or regions. The "Global Case Study Exchange Module" allows physicians to brainstorm together on different patient summaries and to learn about, and find, possible potential treatments. A large amount of open and complete information is required for currently unsolved disease treatment issues. Under the condition of privacy protection and the provision of correct information with regard to the different case symptoms, treatment methods, drugs, etc, it may be possible to find the best antidote to solve the infectious disease crisis the world is facing. The "Case Tracking Module" allows Centers for Disease Control and Prevention (CDC) members to track a patient's movement path before a diagnosis is made. The tracking map is shown in the module. According to different patients' statements about their own moving paths, a moving map can be established that contains international paths. CDC members will be able to carry out risk control and track high-risk groups according to this map, thereby effectively controlling the scope of disease infection and completing it as soon as possible.

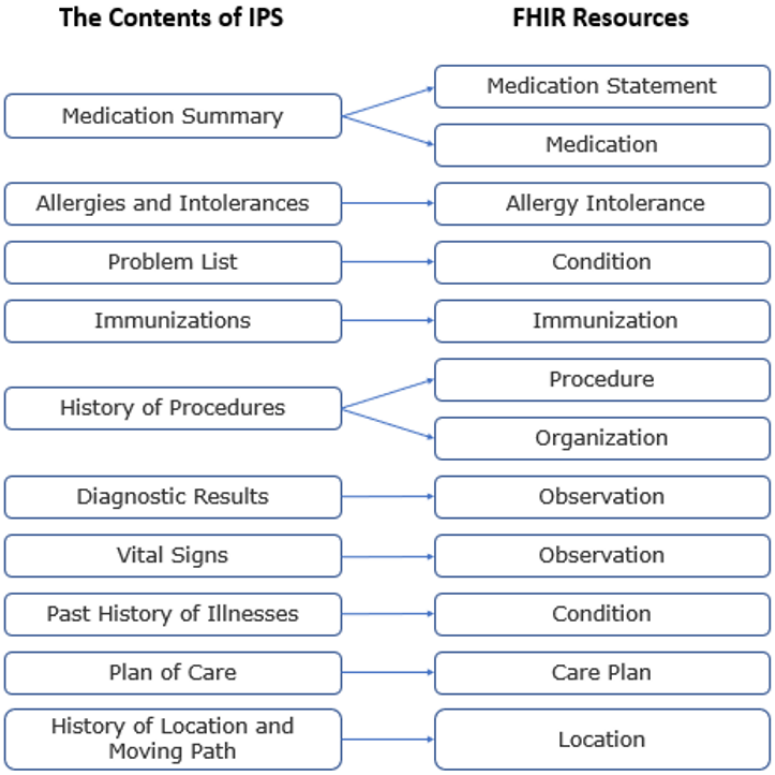
The security and correctness of the IPS are protected by blockchain architecture. When IPS data are uploaded, the details of the data will be deidentified, the block will store the data update log, and the IPS hash value is calculated by the Secure Hash Algorithm (SHA)-256. The IPS data are stored in the HAPI FHIR database, which is open source and an implementation of the interoperability of HL7 FHIR for health care systems in Java. It was developed as an open community by a global team [26]. The IPS continuity of each patient will be connected through the information of the blockchain. User identities are divided into two types, namely, physicians and CDC members. Physicians need to be authenticated through their medical ID certificate in their countries, and CDC members are registered and managed by the CDC units in various countries.

The IPS Tailored for COVID-19 Case Data

An "International Patient Summary Implementation Guide" has been published by HL7 FHIR. The goal is to provide a universal international solution for global health care service applications. This study uses the IPS (Standard for Trial Use 1-FHIR R4, launched on August 6, 2019) as a case study, as it provides treatment and health care information records for global cases of unknown infectious diseases. IPS is a minimal and nonexhaustive patient summary, which means that it is not intended to copy the full content of an EHR. The IPS is usable by clinicians for the unscheduled cross-border care of a patient and focuses on a patient's current condition, instead of anything specific to a particular condition. Furthermore, the IPS is applied on a global scale to address the international feasibility of use as much as possible.

To provide a reference for global cases, the IPS is designed to include information on the following: "Medication Summary," "Allergies and Intolerances," "Problem List," "Immunizations," "History of Procedures," "Diagnostic Results," "Vital Signs," "Past History of Illness," "Plan of Care," "History of Location and Moving Path before Diagnosis," and "Location." The structure of the IPS is shown in Figure 2.

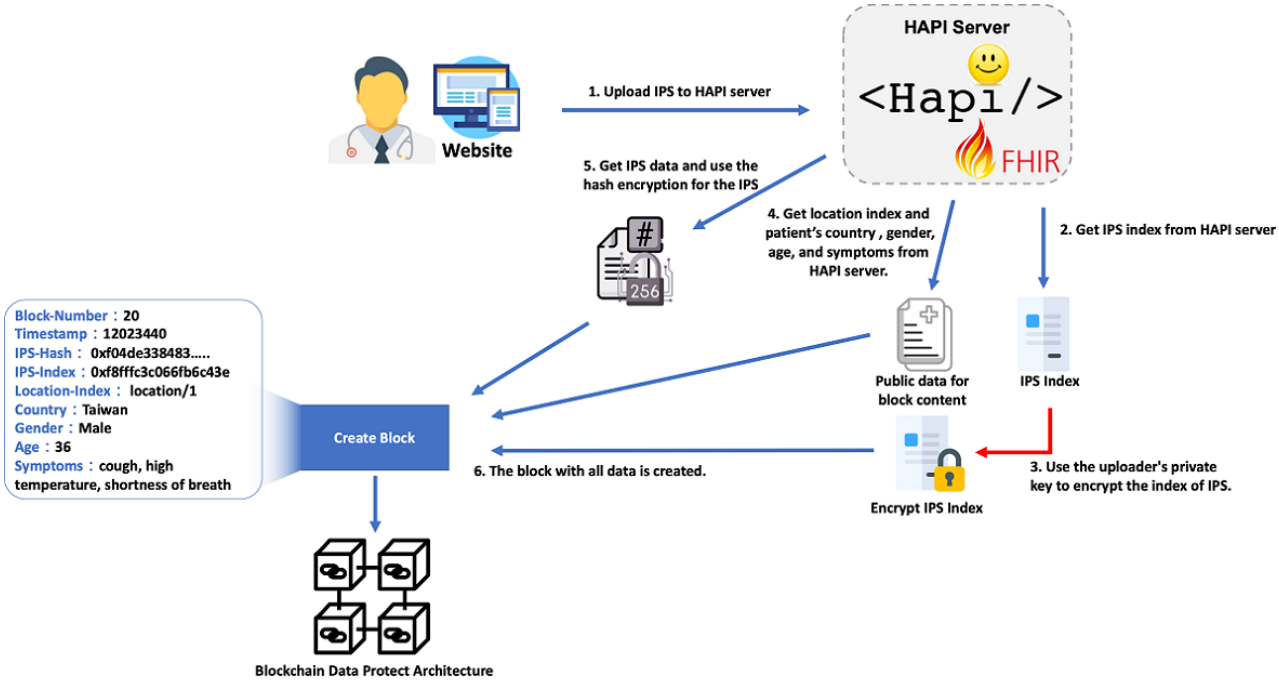
Figure 2. IPS contents mapped to the structures of FHIR resources. FHIR: Fast Healthcare Interoperability Resources; IPS: International Patient Summary.



Case Study Upload Module for IPS Protection and Validation

The physician uploads the patient's IPS document to the system's HAPI server, and the HAPI server corresponds to the IPS index with the blockchain architecture. The IPS index information was designed to connect the data from the HAPI server, including the IPS hash value and the encrypted IPS index value. The deidentified and simplified case data include the gender, age, symptoms, country, and location index value of the HAPI server. After the physician has been authenticated, they have permission to upload the IPS document and view its study cases. The encryption and decryption for the data upload process and architecture is shown in Figure 3.

Figure 3. The encryption and decryption for the data upload process and architecture. FHIR: Fast Healthcare Interoperability Resources; IPS: International Patient Summary.



The steps of this process are as follows:

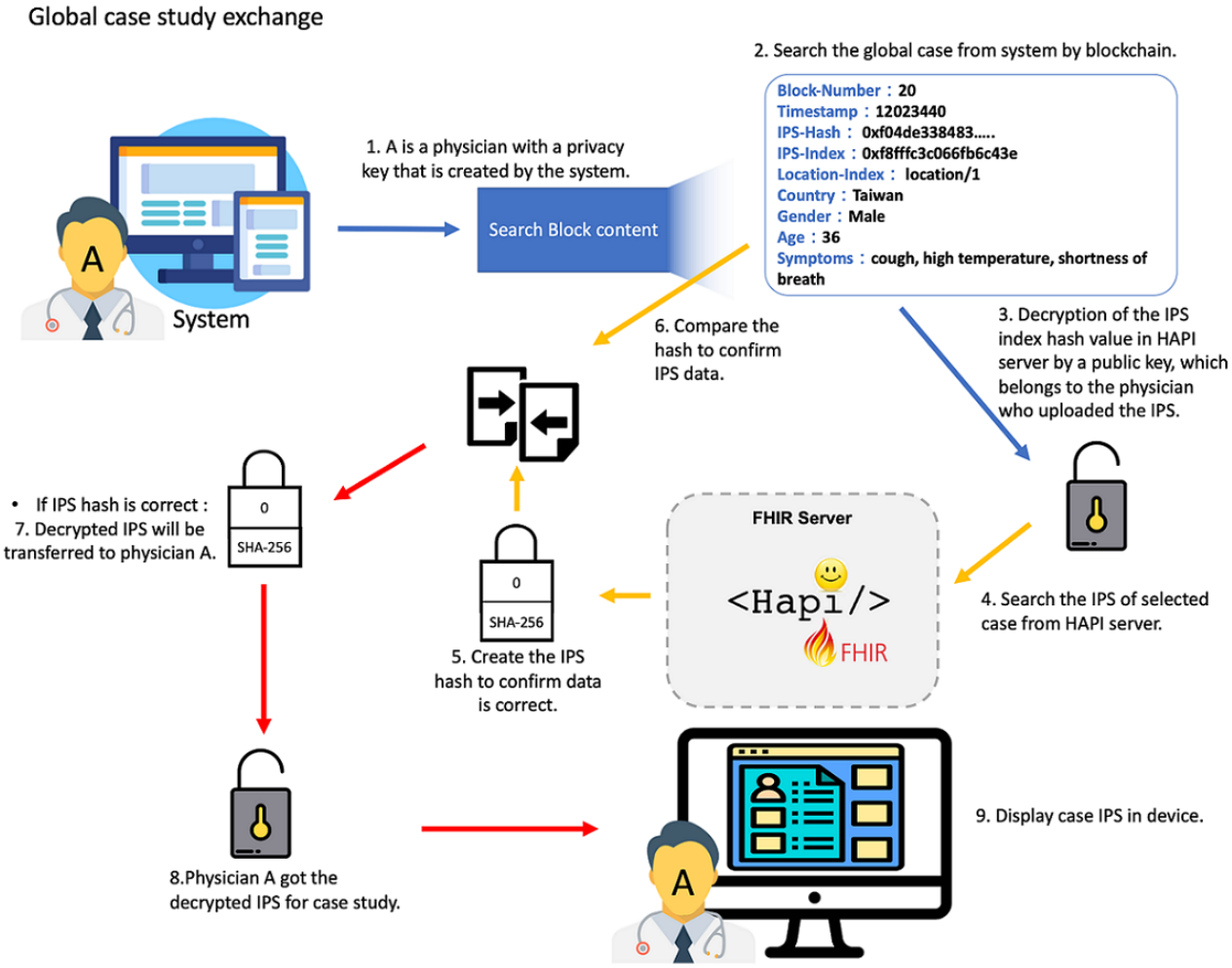
- Step 1: The certified physician uploads the patient's IPS file to the system, and the IPS file will be stored in the HAPI server. Patient identification will be replaced by a globally unique identifier (GUID), which is an 128-bit number that is used to identify the information in the system.
- Step 2: The data index position of the IPS is obtained from the HAPI server.
- Step 3: The private key of the uploaded physician is used to encrypt the IPS index of the data, which is stored in the HAPI server.
- Step 4: The anonymous IPS public information is obtained from the HAPI server, including the mobile path index position, gender, age, country, and symptoms.
- Step 5: The hash value of the IPS file is calculated by the SHA-256 encryption function.
- Step 6: The content of this block is transferred to the blockchain architecture, and a new block is established by the blockchain architecture.

Global Case Study Exchange Module

In a state of globalization, new diseases or clinical pathways that are not treated correctly are likely to rage around the world. COVID-19 has spread worldwide, and therapeutic vaccines and drugs have not yet been developed to treat it. This study constructed a global patient summary exchange model and shared the global research progress through case analyses so that physicians in different regions of the world can refer to the results of acquisition and test cases while at the same time obtaining and learning more about the unknown disease and finding the best treatment process.

Our study is designed for IPS sharing, which can help clinical physicians to find successful treatments and clinical pathways to improve the patients' survival and reduce sequelae. We have designed the model so that physicians need to register first and provide proof of their identity. The system provides each physician with a privacy key for IPS decryption. This system allows physicians to view the summary of the patient cases that have been uploaded all over the world, and it provides a filter function of the cases. Specific cases can be tracked by using this module. The process of how physicians get the IPS files of global study cases is shown in Figure 4.

Figure 4. Process of physicians getting the IPS files of global case studies. FHIR: Fast Healthcare Interoperability Resources; IPS: International Patient Summary; SHA: Secure Hash Algorithm.



We designed a nine-step process for completing the Systems Engineering Initiative for Patient Safety (SEIPS) access to international cases, which includes a data search, decryption, verification, and transmission.

- Step 1: The system verifies the identity of the user, confirming that the user is a physician with registration data.
- Step 2: A list of global patients and simple case information is provided to the physicians, including the patient's region, country, age, and gender.
- Step 3: The index of the selected SEIPS is decrypted by the privacy key of the physician who is uploading the IPS file.
- Step 4: The selected patient IPS file is retrieved from the decentralized database.
- Step 5: The decrypted IPS data are hashed again by SHA-256.
- Step 6: The hash value that is decrypted in step 5 is compared to the hash value in the blockchain.
- Step 7: If the two hash values are equal, it means the data are correct, and the decrypted data are transmitted to the physician.
- Step 8: The system confirms that the physician has obtained the decrypted case study data.
- Step 9: All the IPS files of the selected cases are presented on the physician's display.

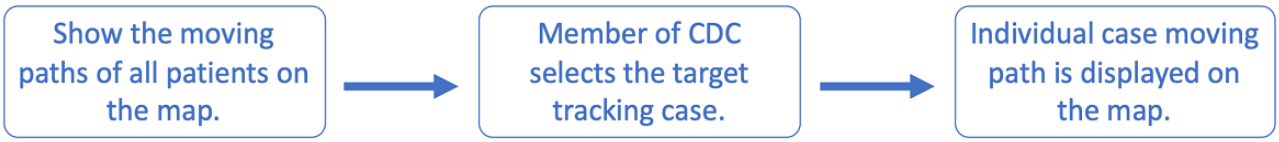
The module is designed as a web-based application, and it includes the Open Application Programming Interface (API) architecture. The module provides various APIs to let the public and private physicians' clinic management system operate easily with the module and to conduct the case exchange.

Case Tracking Module for Infectious Disease Prevention

The prevalence of international tourism and the rapid movement of populations, in an era of globalization, have increased the spread of COVID-19. In just 3 months, it has spread from a limited area (one city in Asia) to becoming a source of infection throughout the world, and the number of infected people continues to increase.

To effectively control the scope of infection and prevent continued expansion, the movement path of patients who are infected needs to be tracked. The FHIR "Location" resource is included in the patient's IPS file, and it helps CDC members effectively track the patients and prevent the continued spread of the disease, based on the record of moving paths and time stamps. The workflow of case tracking is shown in Figure 5.

Figure 5. Workflow of case tracking. CDC: Centers for Disease Control and Prevention.



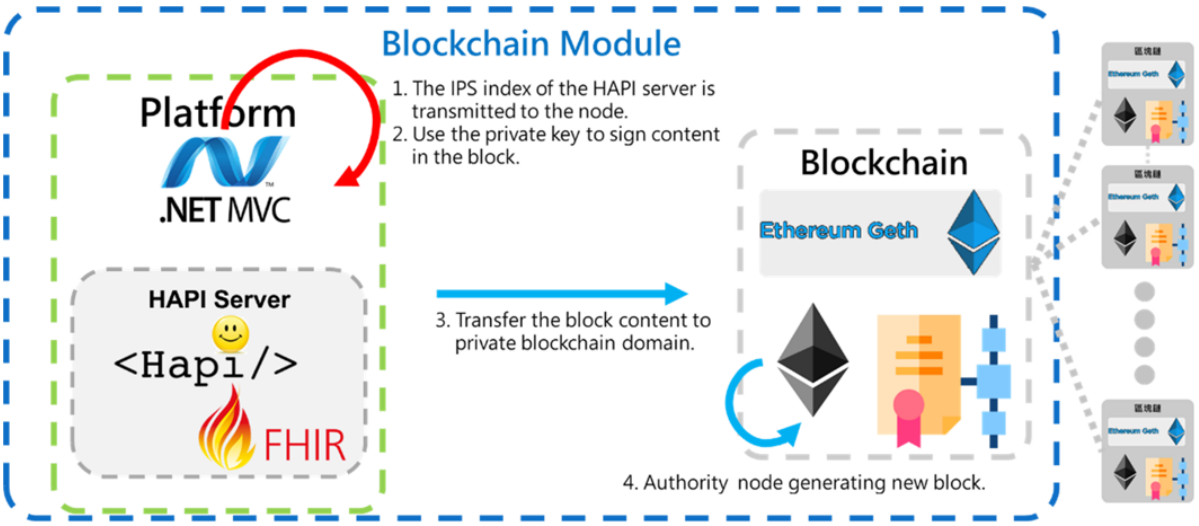
Structure of Blockchain Security

The blockchain architecture was established as the security protection mechanism of IPS data, and the HAPI server was used as a data server for the FHIR IPS. The block in the blockchain is public data for all users and includes the IPS index information and deidentified simple case data, which includes gender, age, symptoms, country, and the HAPI server data index.

Blockchains have many different authority mechanisms. In this study, considering the privacy of a patient's medical data and the need to process a large amount of medical information, the blockchain was built in a private chain, and a Proof of Authority (PoA), with a fast transaction speed and high privacy, was adopted as the consensus on the blockchain. In 2015, PoA was proposed by the Ethereum cofounder, Gavin Wood [27]. This consensus algorithm is used to set up trusted nodes as block validators. It is a centralized consensus mechanism that ensures data security and data verification through authorization mechanisms. The blocks on the chain are generated by trusted nodes, which can improve the efficiency of the generating blocks and ensure consistent data. At the mean times, the system runs well. The ownership of the nodes depends on the policy of the health care authority in different areas. For example, it can be a hospital center or the CDC of a nation.

The process of generating a new block includes four steps, as shown in Figure 6.

Figure 6. The process of generating a new block. FHIR: Fast Healthcare Interoperability Resources; IPS: International Patient Summary; MVC: model–view–controller.



- Step 1: The IPS index of the HAPI server is transmitted to the node.
- Step 2: The private key is used to sign the content in the block.
- Step 3: The block content is transferred to the private blockchain domain.
- Step 4: The authority node generates a new block.

Blockchain architecture will automatically copy the new block data to other nodes to complete the goal of blockchain decentralization.

Results

Global Infectious Disease Surveillance of the IPS for Case Studies

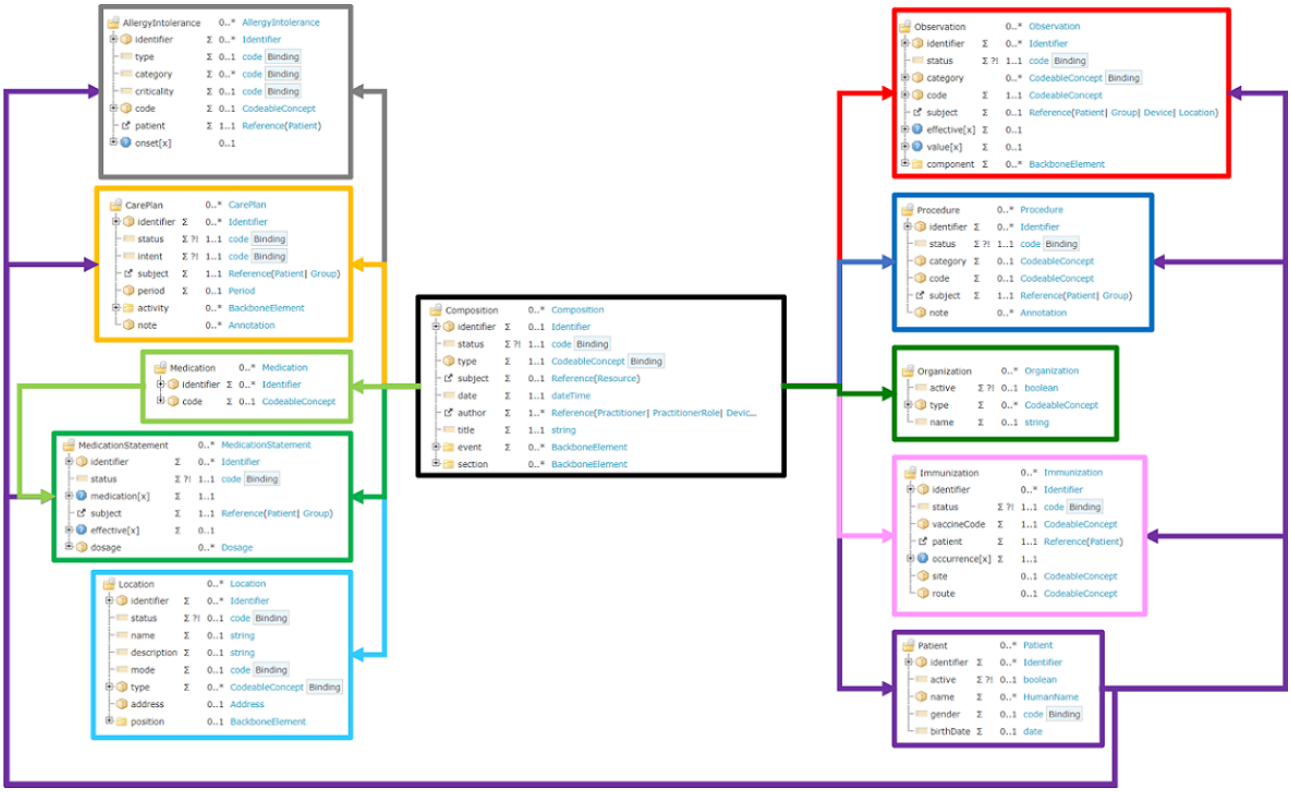
When facing the spread of an unknown disease around the world, such as COVID-19, global case studies must be shared and exchanged quickly. Clinical data must be allowed to be transmitted efficiently and safely to jointly find the most appropriate control and treatment methods through international cooperation. Because different patients have different disease histories, family disease histories, and life environments, their symptoms and disease progression will be different.

An example of this is the SARS outbreak in 2003. After the outbreak, Hong Kong found numerous problems in the surveillance systems of communicable diseases, and the 2003 contact tracking system was inadequate for dealing with the

scale of the SARS epidemic. The public health surveillance systems were not well developed in the private sector and in community clinics, there was no comprehensive laboratory surveillance system, and the hospital authority's laboratory database was not linked to the department of health in the early stages of the epidemic.

This study thus designs an IPS that complies with infectious disease surveillance and clinically meaningful data, according to the IPS HL7 FHIR guidelines. The IPS that we designed includes the following: "Medication Summary," "Allergies and Intolerances," "Problem List," "Immunizations," "History of Procedures," "Diagnostic Results," "Vital Signs," "Past History of Illness," "Plan of Care," and "History of Location and Moving Path before Diagnosis." The IPS content with the structures of FHIR resources is shown in Figure 7.

Figure 7. International Patient Summary contents are mapped to the structures of the Fast Healthcare Interoperability Resources.



The Medication Summary section includes a description of the current and past medications that a patient takes. The Allergies or Intolerances section of a patient includes a description of the kind of reaction, the agents that caused it, as well as the criticality and the certainty of the allergy. The Problem List section includes clinical problems and the conditions of the patient that are currently being monitored. The Immunizations section includes a patient's current immunization status and pertinent immunization history. The History of Procedures section includes a description of the patient procedures that are within the scope of the IPS. The Diagnostic Results section includes the relevant observations and in vitro biological specimens that are collected from the patient. In this section, the laboratory, imaging, and pathology reports may be included. The Vital Signs section includes the data collected when the patient received a medical service or was under surveillance in the hospital, such as the body temperature, blood pressure, heart rate, respiratory rate, height, weight, and BMI. The History of Illnesses section includes the patient's disease history. This section can help physicians to make clinical decisions and get more information from the data. The Plan of Care section includes a description of the clinical care, such as a plan of the proposals, goals, monitoring, tracking, and ordering of requirements to improve the patient's condition. The History of Location and Moving Path section includes where the patient has moved from and to during the incubation period of the infectious disease, as well as the location where the patient was infected (eg, a hospital, hotel, restaurant, bus, plane, or cruise ship). This section is important for controlling the spread of the disease, identifying potential patients, and completing prevention.

After the data of the FHIR IPS is uploaded, the system accepts the input by using the JavaScript Object Notation format. The FHIR IPS integrates each different resource into the same file as a "bundle" resource, and finally, it is uploaded into the HAPI server.

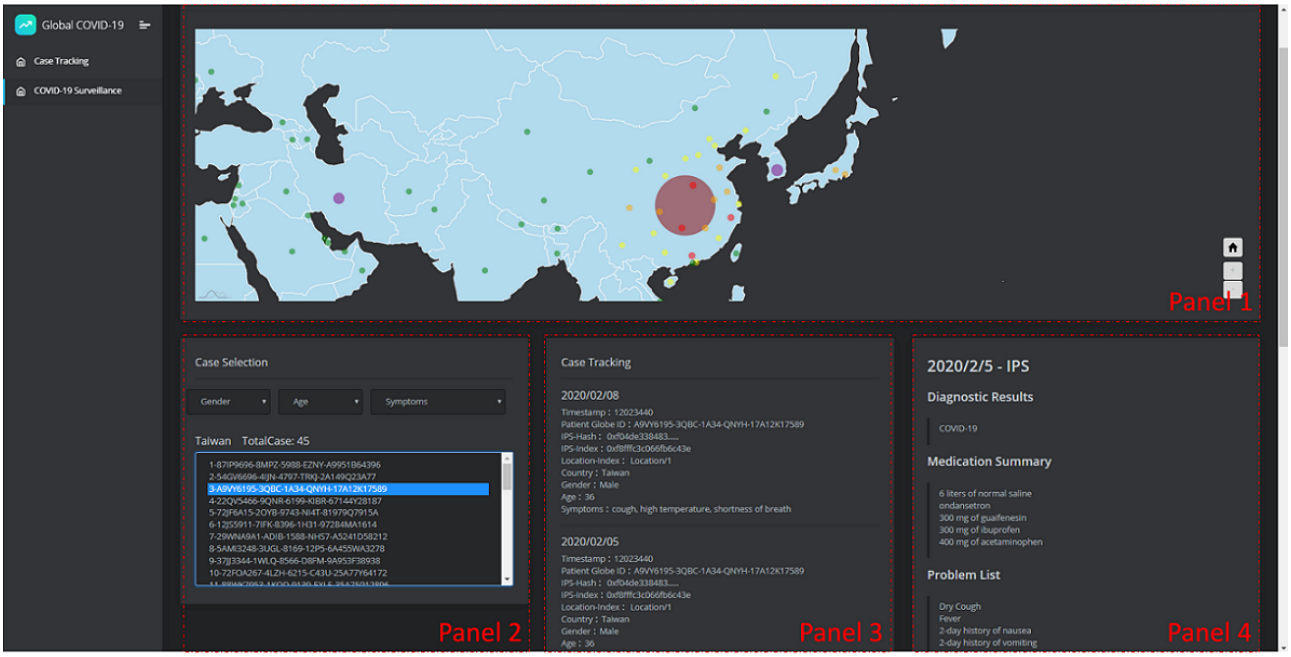
Global COVID-19 Surveillance System for Case Studies

To achieve the purpose of global COVID-19 surveillance and to enhance health resilience, the exchange of global infectious disease information must be enacted. The COVID-19 surveillance system was built and designed based on the blockchain architecture. The IPS is used to exchange case study information among physicians. When physicians pass the system verification, they can upload the case IPS file and get the IPS data of other global cases from the system. The IPS file should

be uploaded daily by the physician. The system includes daily IPS uploading and an enhancement plan, which covers real-time uploading through the interoperation of the clinic system with the module, based on the Open API architecture.

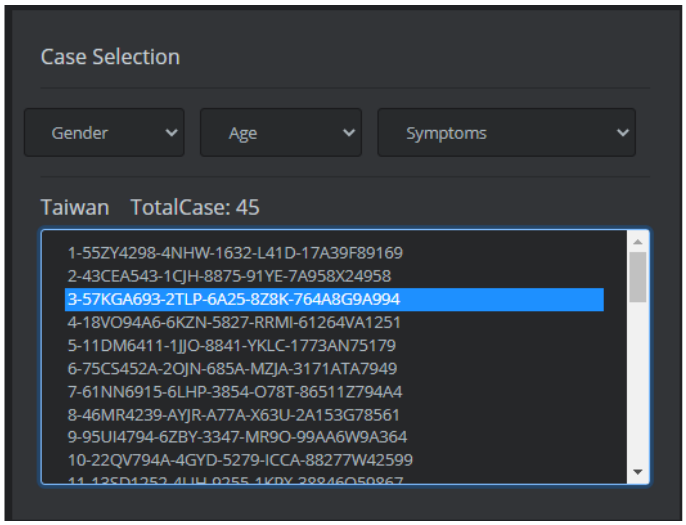
All physician users have access to the case IPS files in the case study system to support clinical decision making. The system's user interface (UI) is shown in Figure 8, and it is divided into four panels that achieve different functions. The case diagram is displayed in Panel 1, where users can obtain the number of cases and international case distribution information. Cases from different places can be selected in Panel 2, as well as in the system UI, as shown in Figure 9. The screening conditions are gender, age, and symptoms, which are used to screen-reference the cases that are similar to their own case. The case IPS information can be viewed in Panel 3, which includes all the uploaded IPS files, the basic information of the patient summary, and the IPS information on the blockchain. The detailed IPS content is viewed in Panel 4. The authenticated physician can use this system to share and exchange the patient IPS files to provide international references. Through the treatment of different cases, the drug treatments, and the exchange of the patient treatment results, the spread of the disease can be controlled, and treatment methods can be funded.

Figure 8. The COVID-19 surveillance system. IPS: International Patient Summary.



In our design, the user selects the country to track the case in Panel 1, and the country circle represents the number of cases. After selecting the country, Panel 2 will display the total number of case data that have been uploaded, as well as the GUID that each case represents in the system. Panel 2 gives the option to filter cases. After selecting a case, Panels 3 and 4 will display the IPS information of the selected case. The case selection (Panel 2) is shown in Figure 9. It is a Taiwanese example, and the patient GUID is represented as "5AIF63A5-9KWE-1653-AR1I-49682N29A22."

Figure 9. User interface of Case Selection (Panel 2).



Blockchain Information of IPS Data

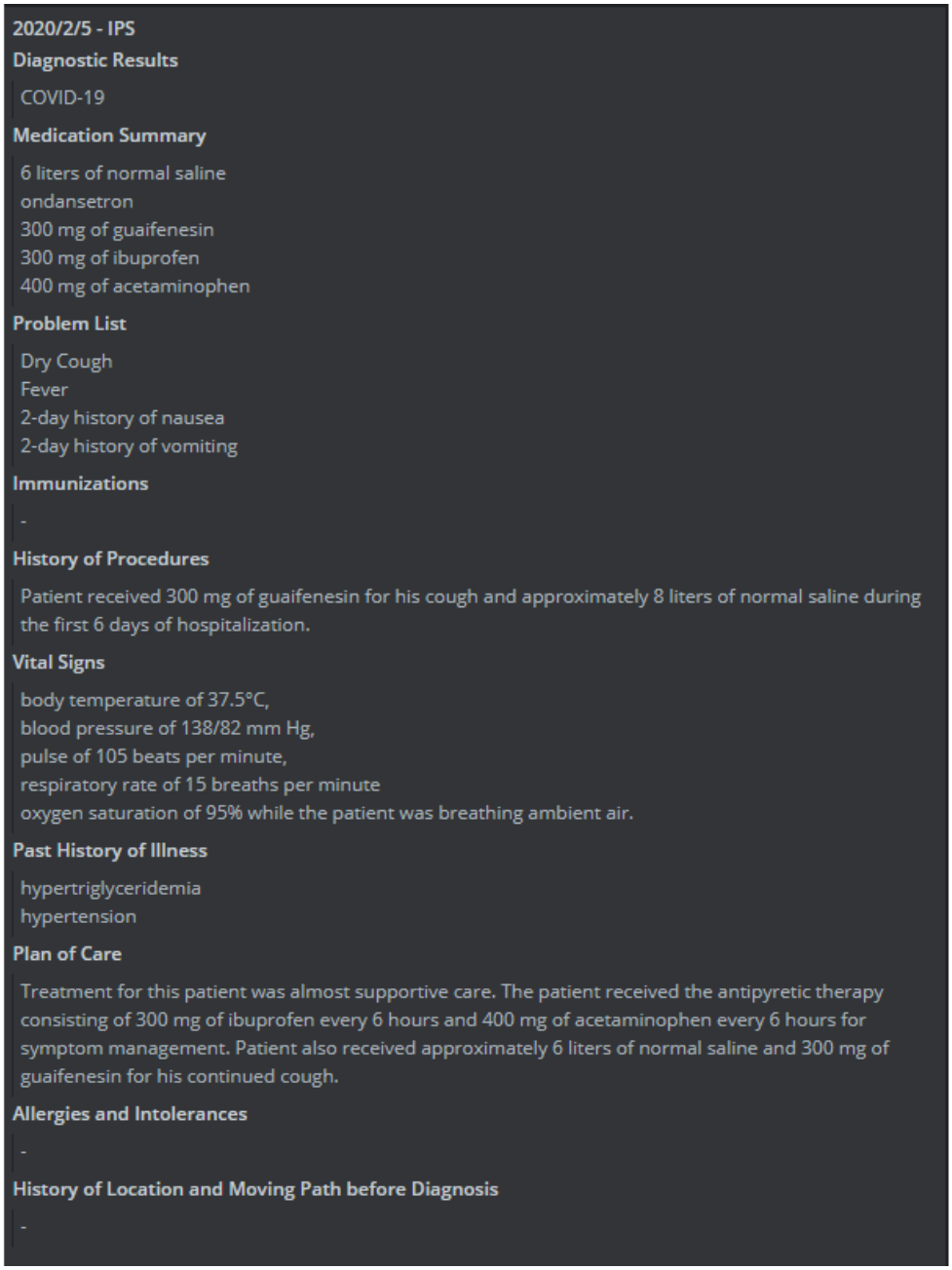
In this study, all uploaded IPS information will be verified and stored in the uploading record by using the blockchain. Panel 3 is mainly the block information of the selected case. Figure 10 shows the block information of a patient whose GUID is "57KGA693-2TLP-6A25-8Z8K-764A8G9A994." In the example, two blocks mean that the case has two uploaded IPS files, and the block information includes a time stamp, the GUID, and the IPS-hash and -index, as well as the moving location, country, gender, age, and symptoms.

Figure 10. User interface of case tracking by block information.

IPS File of COVID-19 Case

A COVID-19 case report is to be used as an example in this study. On February 5, 2020, a female patient 52 years of age presented with a fever and went to a hospital [28]. The patient had type 2 diabetes and had visited Wuhan on January 20. She developed a fever and myalgia 5 days after her return to Taiwan. She self-reported that she did not have dyspnea, a cough, chest pain, or diarrhea. The diagnosis of COVID-19 was made by a real-time reverse transcription polymerase chain reaction. The treatment for this patient was supportive care. The patient received the antipyretic therapy, which consisted of 300 mg of ibuprofen every 6 hours and 400 mg of acetaminophen every 6 hours for symptom management. The patient also received approximately 6 liters of normal saline and 300 mg of guaifenesin for her continued cough. An example of a COVID-19 IPS file is shown in Figure 11.

Figure 11. An example of a COVID-19 IPS file. IPS: International Patient Summary.



The following is additional information about the patient:

- Medication Summary
  - 6 liters of normal saline ondansetron
  - 300 mg of guaifenesin
  - 300 mg of ibuprofen
  - 400 mg of acetaminophen
- Problem List
  - Dry cough
  - Fever
  - 2-day history of nausea
  - 2-day history of vomiting
- Immunizations
- History of Procedures
  - Patient received 300 mg of guaifenesin for her cough and approximately 8 liters of normal saline during the first 6 days of hospitalization.
- Vital Signs
  - Body temperature of 37.5 °C
  - Blood pressure of 138/82 mm Hg
  - Pulse of 105 beats per minute
  - Respiratory rate of 15 breaths per minute
  - Oxygen saturation of 95% while the patient was breathing ambient air
- History of Illness
  - Hypertriglyceridemia
  - Hypertension

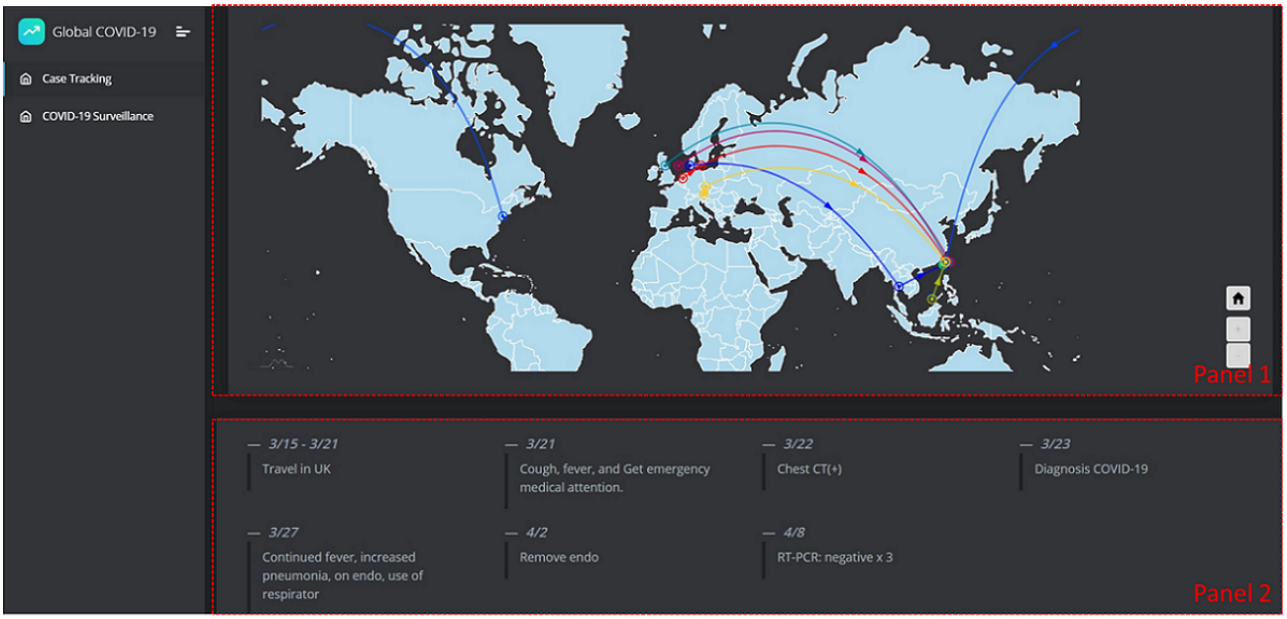
Based on other IPS files, international physicians can refer to the care plans of other patient, as well as their disease history,

medication, and therapy, and give their own patients the appropriate therapy. Our system provides a new architecture for the exchange of IPS files.

Case Tracking of COVID-19

From the establishment of the infectious disease case-tracking module, and by using the location information in the IPS file of the patient, we can track the moving paths of infectious disease cases. The location information of the patient is recorded in the block contents on the blockchain and is not protected as personal clinical data. Therefore, the location information can be retrieved and used by the system for the purpose of tracking the moving paths for different cases. The Case Tracking module has been established for CDC members to track cases and prevent the spread of a disease. Based on this module, CDC members can identify the moving paths of cases and design a case tracking plan for the epidemic investigation. The UI of the COVID-19 case tracking system is shown in Figure 12. The UI is divided into two panels. Nine cases that were diagnosed as COVID-19 in Taiwan were sampled as an example to show the case tracking function of the system. Their data were uploaded onto the blockchain, and the distributions of the moving paths of all cases is shown in Panel 1, where we can see all the worldwide cases as well as their moving paths in different colors on the map. The detailed case moving path information and history record is shown in Panel 2, with the locations, time stamps, and possible activities. In Panel 2 of Figure 12, we show the detailed information of one case. We can see that from March 15-21, 2020, the case had travelled in the United Kingdom. The case came back to Taiwan on March 21, showed some symptoms, and went to the emergency room. The case was confirmed as COVID-19 on March 23, and respirator use was started on March 27. From these nine samples, we can see that all of the cases were imported from outside of Taiwan.

Figure 12. The COVID-19 Case Tracking system. Panel 1 shows the distributions of the moving paths of all the cases. Panel 2 shows the detailed moving path information and history record with the locations, time stamps, and possible activities. CT: computed tomography; RT-PCR: reverse transcription polymerase chain reaction.



Discussion

After the outbreak of infectious diseases such as SARS, MERS, and COVID-19, it is well-known that international cooperation for disease treatment is critical, especially due to the current high frequency of travel between countries around the world. Diseases such as SARS and MERS not only affect people's health but also seriously affect the world economy [29]. Although the deterioration of a disease condition depends on many variables, when facing unknown diseases, experience sharing and the exchange of advice are still key points. The control and treatment of any disease needs to be found as soon as possible. To control and treat the disease, a global case study sharing system must be established, not only for clinical data sharing but also for the development of treatment methods.

Through the system designed by this study, minimal and useful patient summary data can be shared. Physicians only need to focus on essential clinical data that can be followed up on, and they can try a specific treatment or medicine when facing unknown diseases such as COVID-19. Data from other countries or other patients can be taken as a reference for patient care and treatment. According to published studies, having a fever and a cough are the dominant symptoms of COVID-19, while gastrointestinal symptoms are uncommon [5,30,31]. One report presents the first confirmed case of COVID-19 in the United States, including the process of identification, diagnosis, clinical course, management, and the patient's symptoms [3]. Overall, there is an important need for coordination between clinicians and public health authorities, as well as for the rapid transfer of clinical information relating to the care of patients with COVID-19.

One case study of the first-known imported case of COVID-19 infection in Taiwan describes how the doctor gave the patient supporting treatment for all her symptoms. However, there is still a lack of details on the clinical information about the patient [28]. Another study of numerous cases was conducted by Chan

et al [32] at Hong Kong University. They found that the outbreak of COVID-19 in Wuhan, China was similar to the 2003 SARS outbreak in Guangzhou, China. Both outbreaks initially happened in the animal-to-people transmission model and not by person-to-person transmission in the community. The case study exchange from the model and the subsequent knowledge exchange, analysis, conclusion, planning, and evaluation will provide a basis for understanding the experiences of previous epidemics, like SARS and MERS, and help to streamline the disease prevention and control measures (eg, regulations for animal and wet markets, patient isolation and tracking, contact quarantine, and public health and hygiene education) to prevent any rapid spread. As their system was helpless against SARS, Hong Kong later developed the Communicable Disease Information System to provide real-time and intelligent syndromic and communicable disease surveillance; to enable rapid intervention and quicker outbreak and emergency responses via field investigations, outbreak control, responsive risk communication, ongoing analysis, alert generation, predictive capability, and early outbreak detection; and to offer a framework for strategic planning and program evaluation. We can rapidly gather information for COVID-19 through international channels, but the information is still not clear enough to use as a reference for treating patients. Lipsitch et al [33] showed that viral testing should not be used just for clinical care, and public health efforts should use it to target the trajectory and severity of the disease. Guan et al [34], from the State Key Laboratory of Respiratory Diseases, noted the limitations of COVID-19 research due to the collection of data from different structures of electronic databases and the urgent timeline for data extraction. Some cases, therefore, have incomplete clinical data of the patients' exposure history and laboratory testing [34].

The main challenge of COVID-19 is that we do not have enough knowledge of the therapy, control methods, and full spread route of the virus, which can only be obtained from the patient. Based on the experience of rapid virus transmission and the burden on the health care system, a global information system is essential. When analyzing the development of COVID-19, it seems that an effective global communicable disease surveillance system has not yet been developed. The disease data are not timely or effectively linked. Physicians and scientists around the world are unable to obtain sufficient disease information in a thorough and timely manner to control the epidemic. Currently, the exchange of case data for clinical research on COVID-19 is incomplete and not quick enough, which limits the development of a treatment design. Even if many case reports were to be submitted, the goals of real-time tracking, data exchange, and referencing could not be achieved. Therefore, to reduce the restrictions on COVID-19 research, an EHRs–based information communication system is necessary, as it can quickly achieve such goals for the public.

This study created the IPS of infectious diseases that physicians can access when treating patients with COVID-19. We have also established a secure blockchain architecture for the protection of the IPS, and we have completed the application of tracking patients' moving path. The IPS case studies can be exchanged through our system and verified through the blockchain architecture. Over the past few years, blockchain has been used in many different fields, not only with regard to medical records (EHRs and personal health records) but also to medical data exchange issues. Benil and Jasper [35] introduced blockchain architecture for managing EHRs. In its design, the EHR is stored in the cloud, and its integrity in the cloud will be checked through the blockchain. This is a similar architecture to our study and proves that the blockchain can protect and verify EHRs. Fan et al [36] proposed a blockchain-based consensus mechanism for medical information data security and privacy in the medical system. Sun et al [37] presented a distributed signature scheme for medical systems with a record-sharing protocol that is based on blockchain. Yang and Li [38] designed an architecture for securing the EHR system, which is based on distributed ledger technology, to improve the interoperability of health record exchanges between different organizations. Chen et al [39] introduced a searchable encryption scheme for EHRs by using blockchain. Blockchain architecture can ensure data security and verify that the information is correct, and it is therefore a suitable architecture for global IPS file exchange.

The results of this study can help health authorities respond quickly to the transmission and spread of any unknown disease, and it can provide a good system for information retrieval on disease transmission. Another benefit of this system is that it can help public health researchers form study trials and analyze data from different countries. A trial on medication treatment in patients with COVID-19 found that the lopinavir–ritonavir treatment added to the standard supportive care, but it was not significant for clinical improvement or mortality in patients with COVID-19 [40]. Other research on the use of chloroquine and hydroxychloroquine in COVID-19 shows that the use of these drugs is premature and potentially harmful [41].

However, the clinical observation details of patients were not described by the authors. It is hard to identify which supportive care works best for patients in different situations. Another effective means for fighting an unknown virus could be using a common forum to facilitate the mutual sharing of experiences, best practices, therapies for patients, and the possible useful medications and outcomes from clinical interventions being trialed in various countries in a secure, trustworthy manner. The system designed by this study can become an effective tool for facilitating global collaboration and cooperation, and for promoting collective evidence-based efforts to address the unprecedented situation created by COVID-19. However, this study has some limitations. At present, there is no optimal treatment, and complete information about this disease has not yet been found. Governments, medical institutions, and physicians from all over the world should cooperate in the study of this virus. Without international cooperation, global interests will have significant losses. This study has completed the design and development of a global infectious disease surveillance and case tracking system for COVID-19, and found that it has a stable foundation and is a balanced system. However, there is still a need to test the effectiveness of a large number of users uploading and exchanging data simultaneously. In the future, our team will have discussions with governments, international medical service providers, and medical institutions to activate

this system and to promote international cooperation and development during the COVID-19 outbreak.

Acknowledgments

This study received funding from the Ministry of Science and Technology, Taiwan, under the project No. 108-3011-F-075-001 and the Ministry of Education, Taiwan, under the project No. 107EH12-22.

Authors' Contributions

This study was carried out in collaboration among all the authors. HAL and CYH conceptualized the research and designed the architecture of the system. YJL and JCJC provided the IPS use case for testing and demonstration. JGU and KKN contributed to the conceptualization. HAL, HHK, JGU, HCF, and KKN carried out the literature review. HAL, HHK, and YKC were instrumental in the implementation of the system. HAL drafted the manuscript, and CYH and JGU made significant revisions. CYH, YJL, JCJC, JGU, YKC, BK, ABM, and LRC supervised the methodology of implementing a global COVID-19 infectious disease surveillance and case tracking system, and suggested valuable improvements. All authors approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Lu H, Stratton CW, Tang Y. Outbreak of pneumonia of unknown etiology in Wuhan, China: the mystery and the miracle. J Med Virol 2020 Apr;92(4):401-402 [FREE Full text] [doi: 10.1002/jmv.25678] [Medline: 31950516]
2. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. JAMA 2020 Mar 17;323(11):1061-1069 [FREE Full text] [doi: 10.1001/jama.2020.1585] [Medline: 32031570]
3. Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, Washington State 2019-nCoV Case Investigation Team. First case of 2019 novel coronavirus in the United States. N Engl J Med 2020 Mar 05;382(10):929-936 [FREE Full text] [doi: 10.1056/NEJMoa2001191] [Medline: 32004427]
4. Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. Lancet 2020 Feb 15;395(10223):470-473 [FREE Full text] [doi: 10.1016/S0140-6736(20)30185-9] [Medline: 31986257]
5. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020 Feb 15;395(10223):497-506 [FREE Full text] [doi: 10.1016/S0140-6736(20)30183-5] [Medline: 31986264]
6. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet 2020 Feb 15;395(10223):507-513 [FREE Full text] [doi: 10.1016/S0140-6736(20)30211-7] [Medline: 32007143]
7. Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, et al. Severe acute respiratory syndrome-related coronavirus: the species and its viruses – a statement of the Coronavirus Study Group. BioRxiv. Preprint posted online February 11, 2020. [doi: 10.1101/2020.02.07.937862]
8. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med 2020 Mar 26;382(13):1199-1207 [FREE Full text] [doi: 10.1056/NEJMoa2001316] [Medline: 31995857]
9. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020 May;20(5):533-534. [doi: 10.1016/S1473-3099(20)30120-1] [Medline: 32087114]
10. Hamade MA. COVID-19: how to fight disease outbreaks with data. World Economic Forum. 2020 Mar 19. URL: https://www.weforum.org/agenda/2020/03/covid-19-how-to-fight-disease-outbreaks-with-data [accessed 2020-04-10]
11. Chang D, Lin M, Wei L, Xie L, Zhu G, Dela Cruz CS, et al. Epidemiologic and clinical characteristics of novel coronavirus infections involving 13 patients outside Wuhan, China. JAMA 2020 Mar 17;323(11):1092-1093 [FREE Full text] [doi: 10.1001/jama.2020.1623] [Medline: 32031568]
12. Jajosky RA, Groseclose SL. Evaluation of reporting timeliness of public health surveillance systems for infectious diseases. BMC Public Health 2004 Jul 26;4:29 [FREE Full text] [doi: 10.1186/1471-2458-4-29] [Medline: 15274746]
13. Hay SI, George DB, Moyes CL, Brownstein JS. Big data opportunities for global infectious disease surveillance. PLoS Med 2013;10(4):e1001413 [FREE Full text] [doi: 10.1371/journal.pmed.1001413] [Medline: 23565065]
14. Lederberg J, Shope RE, Oaks SC, editors. Emerging Infections: Microbial Threats to Health in the United States. Washington, DC: National Academies Press; 1992.
15. Feldmann H, Czub M, Jones S, Dick D, Garbutt M, Grolla A, et al. Emerging and re-emerging infectious diseases. Med Microbiol Immunol 2002 Oct;191(2):63-74. [doi: 10.1007/s00430-002-0122-5] [Medline: 12410344]

16. Chretien J, Burkom HS, Sedyaningsih ER, Larasati RP, Lescano AG, Mundaca CC, et al. Syndromic surveillance: adapting innovations to developing settings. PLoS Med 2008 Mar 25;5(3):e72 [FREE Full text] [doi: 10.1371/journal.pmed.0050072] [Medline: 18366250]
17. Chretien J, Lewis SH. Electronic public health surveillance in developing settings: meeting summary. BMC Proc 2008 Nov 14;2 Suppl 3:S1 [FREE Full text] [doi: 10.1186/1753-6561-2-s3-s1] [Medline: 19025678]
18. Woodall J. Official versus unofficial outbreak reporting through the internet. Int J Med Inform 1997 Nov;47(1-2):31-34. [doi: 10.1016/s1386-5056(97)00079-8] [Medline: 9506388]
19. Hung HC, Chan DW, Law LK, Chan EH, Wong ES. Industrial experience and research into the causes of SARS virus transmission in a high-rise residential housing estate in Hong Kong. Building Services Eng Res Technol 2016 Jul 27;27(2):91-102. [doi: 10.1191/0143624406bt145oa]
20. Global surveillance for COVID-19 caused by human infection with COVID-19 virus: interim guidance, 20 March 2020. World Health Organization. 2020. URL: https://apps.who.int/iris/handle/10665/331506 [accessed 2020-12-17]
21. COVID-19 novel coronavirus pandemic. Interoperability Standards Advisory. 2020. URL: https://www.healthit.gov/isa/covid-19 [accessed 2020-07-07]
22. Logica. Table of contents. Logica COVID-19 FHIR Profile Library IG. 2020. URL: https://covid-19-ig.logicahealth.org/toc.html [accessed 2020-07-07]
23. Dedalus COVID-19 Solution. Simplifier.net. 2020. URL: https://simplifier.net/guide/dedaluscovid-19solution/home [accessed 2020-07-07]
24. Cangioli G, Hausam R, Macary F, Geßner C, Dickinson G, Heitmann KU, et al. International Patient Summary Implementation guide. HL7 FHIR. 2020. URL: https://build.fhir.org/ig/HL7/fhir-ips/ [accessed 2020-04-10]
25. Holmgren A, Apathy NC, Adler-Milstein J. Barriers to hospital electronic public health reporting and implications for the COVID-19 pandemic. J Am Med Inform Assoc 2020 Aug 01;27(8):1306-1309 [FREE Full text] [doi: 10.1093/jamia/ocaa112] [Medline: 32442266]
26. HAPI FHIR. 2020. URL: https://hapifhir.io/ [accessed 2020-07-07]
27. Wood G. PoA private chains. GitHub. 2015. URL: https://github.com/ethereum/guide/blob/master/poa.md [accessed 2020-07-07]
28. Liu YC, Liao CH, Chang CF, Chou CC, Lin YR. A locally transmitted case of SARS-CoV-2 infection in Taiwan. N Engl J Med 2020 Mar 12;382(11):1070-1072 [FREE Full text] [doi: 10.1056/NEJMc2001573] [Medline: 32050059]
29. Keogh-Brown MR, Smith RD. The economic impact of SARS: how does the reality match the predictions? Health Policy 2008 Oct;88(1):110-120 [FREE Full text] [doi: 10.1016/j.healthpol.2008.03.003] [Medline: 18436332]
30. Lei H, Li Y, Xiao S, Lin C, Norris SL, Wei D, et al. Routes of transmission of influenza A H1N1, SARS CoV, and norovirus in air cabin: comparative analyses. Indoor Air 2018 May;28(3):394-403 [FREE Full text] [doi: 10.1111/ina.12445] [Medline: 29244221]
31. Otter J, Donskey C, Yezli S, Douthwaite S, Goldenberg S, Weber D. Transmission of SARS and MERS coronaviruses and influenza virus in healthcare settings: the possible role of dry surface contamination. J Hosp Infect 2016 Mar;92(3):235-250 [FREE Full text] [doi: 10.1016/j.jhin.2015.08.027] [Medline: 26597631]
32. Chan JFW, Yuan S, Kok K, To KK, Chu H, Yang J, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. Lancet 2020 Feb 15;395(10223):514-523 [FREE Full text] [doi: 10.1016/S0140-6736(20)30154-9] [Medline: 31986261]
33. Lipsitch M, Swerdlow DL, Finelli L. Defining the epidemiology of covid-19 - studies needed. N Engl J Med 2020 Mar 26;382(13):1194-1196. [doi: 10.1056/NEJMp2002125] [Medline: 32074416]
34. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, China Medical Treatment Expert Group for Covid-19. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med 2020 Apr 30;382(18):1708-1720 [FREE Full text] [doi: 10.1056/NEJMoa2002032] [Medline: 32109013]
35. Benil T, Jasper J. Cloud based security on outsourcing using blockchain in E-health systems. Computer Networks 2020 Sep;178:107344. [doi: 10.1016/j.comnet.2020.107344]
36. Fan K, Wang S, Ren Y, Li H, Yang Y. MedBlock: efficient and secure medical data sharing via blockchain. J Med Syst 2018 Jun 21;42(8):136. [doi: 10.1007/s10916-018-0993-7] [Medline: 29931655]
37. Sun Y, Zhang R, Wang X, Kaiqiang G, Liu L. A decentralizing attribute-based signature for healthcare blockchain. 2018 Presented at: 27th International Conference on Computer Communication and Networks (ICCCN); 2018; Hangzhou, China. [doi: 10.1109/icccn.2018.8487349]
38. Yang G, Li C. A design of blockchain-based architecture for the security of electronic health record (EHR) systems. 2018 Presented at: IEEE International Conference on Cloud Computing Technology and Science (CloudCom); December 2018; Nicosia, Cyprus. [doi: 10.1109/cloudcom2018.2018.00058]
39. Chen L, Lee W, Chang C, Choo KR, Zhang N. Blockchain based searchable encryption for electronic health record sharing. Future Generation Computer Syst 2019 Jun;95:420-429. [doi: 10.1016/j.future.2019.01.018]
40. Cao B, Wang Y, Wen D, Liu W, Wang J, Fan G, et al. A trial of lopinavir-ritonavir in adults hospitalized with severe covid-19. N Engl J Med 2020 May 07;382(19):1787-1799 [FREE Full text] [doi: 10.1056/NEJMoa2001282] [Medline: 32187464]

41. Ferner RE, Aronson JK. Chloroquine and hydroxychloroquine in covid-19. BMJ 2020 Apr 08;369:m1432. [doi: 10.1136/bmj.m1432] [Medline: 32269046]

Abbreviations

API: application programming interface
CDC: Centers for Disease Control and Prevention
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
GUID: globally unique identifier
HL7: Health Level 7
IPS: International Patient Summary
MERS: Middle East respiratory syndrome
PoA: Proof of Authority
SARS: severe acute respiratory syndrome
SEIPS: Systems Engineering Initiative for Patient Safety
UI: user interface
WHO: World Health Organization

Edited by G Eysenbach; submitted 26.05.20; peer-reviewed by Y Yu, E Chukwu, T Ueno; comments to author 29.06.20; revised version received 17.07.20; accepted 15.12.20; published 22.12.20.

Please cite as:
Lee HA, Kung HH, Lee YJ, Chao JCJ, Udayasankaran JG, Fan HC, Ng KK, Chang YK, Kijsanayotin B, Marcelo AB, Hsu CY
Global Infectious Disease Surveillance and Case Tracking System for COVID-19: Development Study
JMIR Med Inform 2020;8(12):e20567
URL: http://medinform.jmir.org/2020/12/e20567/
doi:10.2196/20567
PMID:33320826

Original Paper

ISO/IEEE 11073 Treadmill Interoperability Framework and its Test Method: Design and Implementation

Zhi Yong Huang[1], PhD; Yujie Wang[1], BE; Linling Wang[2], BE

[1]School of Microelectronics and Communication Engineering, Chongqing University, Chongqing, China

[2]Bioengineering College, Chongqing University, Chongqing, China

Corresponding Author:
Zhi Yong Huang, PhD
School of Microelectronics and Communication Engineering
Chongqing University
No 174 Shazhengjie, Shapingba
Chongqing, 400044
China
Phone: 86 02365103544
Email: zyhuang@cqu.edu.cn

Abstract

Background: Regular physical activity is proven to help prevent and treat noncommunicable diseases such as heart disease, stroke, diabetes, and breast and colon cancer. The exercise data generated by health and fitness devices (eg, treadmill, exercise bike) are very important for health management service providers to develop personalized training programs. However, at present, there is little research on a unified interoperability framework in the health and fitness domain, and there are not many solutions; besides, the privatized treadmill data transmission scheme is not conducive to data integration and analysis.

Objective: This article will expand the IEEE 11073-PHD standard protocol family, develop standards for health and fitness device (using treadmill as an example) based on the latest version of the 11073-20601 optimized exchange protocol, and design protocol standards compliance testing process and inspection software, which can automatically detect whether the instantiated object of the treadmill meets the standard.

Methods: The study includes the following steps: (1) Map the data transmitted by the treadmill to the 11073-PHD objects; (2) Construct a programming language structure corresponding to the 11073-PHD application protocol data unit (APDU) to complete the coding and decoding part of the test software; and (3) Transmit the instantiated simulated treadmill data to the gateway test software through transmission control protocol for standard compliance testing.

Results: According to the characteristics of the treadmill, a data exchange framework conforming to 11073-PHD is constructed, and a corresponding testing framework is developed; a treadmill agent simulation is implemented, and the interoperability test is performed. Through the designed testing process, the corresponding testing software was developed to complete the standard compliance testing of the treadmill.

Conclusions: The extended research of IEEE 11073-PHD in the field of health and fitness provides a potential new idea for the data transmission framework of sports equipment such as treadmills, which may also provide some help for the development of sports health equipment interoperability standards.

(JMIR Med Inform 2020;8(12):e22000) doi:10.2196/22000

KEYWORDS

ISO/IEEE 11073-PHD; treadmill; standard frame model; test standard; sports health data

Introduction

In order to prevent noncommunicable diseases, the World Health Organization recommends that the world establish special actions to encourage and guide people to participate in more sports, and therefore released the global action plan on physical activity 2018-2030 [1]. To achieve this goal, people need to carry out scientific and effective exercise. Health management service providers usually develop special and personalized training programs for users, and collect user's sports data through a series of sports and health equipment including treadmills, power cars, wearable devices, and so on. These data

can be incorporated into the personal health record [2], and the treadmill data can be integrated into a personalized health management service system along with data from other sports and health equipment.

Therefore, we need to customize a data flow interoperability protocol suitable for treadmills, and the protocol should preferably have the same semantic syntax as the exchange protocol of other sports and health equipment under the framework of a large protocol family. In this way, we can make multiple sports and health equipment conform to the same data exchange format, which greatly reduces the integration difficulty and cost of personal sports data, and facilitates the comprehensive analysis of multiple sports parameters.

The ISO/IEEE 11073 personal health data standard is a set of standards that address the interoperability of personal health equipment (such as scales, blood pressure meters, blood glucose meters). The 11073-PHD protocol family provides a unified semantic grammar data exchange framework for medical device and personal health equipment.

11073-PHD defines an agent device role, which represents a device that provides sports health data, and transmits the obtained data to the master device; a manager device role, which receives sports health data from one or more slave devices by wireless or wired transmission. Thanks to the 11073 protocol, personal health equipment has a unified data transmission protocol at the application layer.

In the 11073-PHD protocol family, 11073-20601 [3] is an optimal exchange protocol, which establishes an abstract logical connection framework between the manager and the agent. This general modeling framework is composed of 3 core models: domain information model (DIM), service model, and communication model, which are respectively used for the semantic description of information and its interrelation and the abstract expression of access interface, definition of data access service, description of interaction behavior, and definition of session synchronization mechanism.

The existing 11073-PHD [4,5] framework helps to provide interoperability for health equipment; unfortunately, compared with the designing and development of equipment and applications in the area of disease management, less efforts had been made to address the demand in the field of health and fitness, which has led to the fact that it cannot effectively support the richer personalized training applications, nor can fully respond to the potential capabilities of various equipment in the sports ecology centered on treadmills. Besides, there are a lot of legacy treadmill devices in the existing sports equipment market [6]. It is a major trend to intelligently transform these inventory devices. If a set of widely applicable interoperable standards can be properly applied, it will greatly reduce the difficulty of equipment transformation and system integration, and provide a unified and standardized interface for system integrators and third-party application developers.

In summary, it is necessary to develop suitable interoperability standards for treadmills, but there is less research work in this field. The development of standards for treadmills based on the latest version of the 11073-20601 exchange protocol can fill the above gaps in a technically appropriate and cost-effective manner. At present, no related research or project implementation is available. Therefore, we plan to expand a set of data transmission protocols specifically suitable for treadmills based on the 11073-PHD protocol family, and design a set of data stream detection schemes that match the protocol.

Methods

Design of PHD-Based Treadmill Interoperability Framework

In the design of treadmill interoperability framework, the main work is to create a DIM. First, we determine the parameters that the treadmill may transmit, then map the data type to the 11073-20601 general framework, add the attribute type of the mapped object according to the parameter type, and finally, determine the corresponding attribute value. As for the service and communication models, there is not much difference from the definition in 11073-20601.

Personal information such as height, weight, and age, and also speed, heart rate [7], distance, and other data generated during exercise during the marked period are essential for the analysis of personal exercise conditions and the formulation of personalized exercise plans [8]. Through the design of the following treadmill objects, the user's movement process can be mainly described, and each concept is briefly explained in the following sections.

Session

A session is similar to an envelope and contains all measures related to an activity scenario or an exercise scenario. Each exercise set defines the start date and time of the scenario and the activities and duration of the activities that the user participates in during the scenario.

Subsession

A subsession is similar to an envelope and contains all the metrics related to the session. Each sport item defines the start date, start time, and duration of the sport item, and also includes the activities that the user participates in during the duration of the sport item.

Age

The age is usually entered manually by the user. The agent can use the age for derivative calculations (eg, calculating the maximum recommended heart rate).

Weight

Weight is usually a setting manually entered by the user, although the device can measure it directly. The weight setting may be used by the device to derive calculations; for example, to calculate the energy consumed during jogging.

Height

The height is usually a setting manually entered by the user. The altitude setting may be used by the device to derive calculations, for example, to calculate BMI.

Distance

The distance defines the total distance covered since the start of the session or event. Distance can be specified as an actual distance concept, for example, meters or feet; it can also be specified as a more abstract concept, for example, the number of steps or the number of stairs climbed. In the latter case, the distance represented by MDC_DIM_STEP (11520) is equal to the step measurement.

Energy Consumption

Energy consumption refers to the amount of energy consumed since the start of a session or event.

Dynamic Heart Rate

Heart rate can be observed as the maximum value, minimum value, and average value of a movement or action, and can also be expressed as an instantaneous value. This rate is a key indicator of physical exertion. In particular, the observed maximum heart rate is an important observation value that may be used to calculate the user's VO_{2max}.

Slope

Slope indicates the steepness of the slope, which can be expressed as the minimum value, average value, or maximum value in the session or subsession, or it can be expressed as the instantaneous value. Positive values indicate uphill and negative values indicate downhill. Therefore, the minimum slope value represents the steepest downhill slope during a session or item.

Maximum Recommended Heart Rate

The maximum recommended heart rate [9] is usually manually entered by the user (or doctor) or calculated. The simplest estimation method is $h = 220 - a$, where h is the maximum recommended heart rate and a is the age. The maximum recommended heart rate can be used to provide background information for other values, such as the maximum heart rate value, minimum heart rate value, and average observed heart rate value that can be reached during an exercise set.

Program Identifier

This measured value identifies the exercise program used by a person during a session or item.

Session–Subsession–Start–Indicator

"Session–Subsession–start–indicator" is used to mark the start position of the continuously monitored session or subsession.

Speed

Speed adds additional contextual information to the ongoing movement and is used to capture the speed of the user through a distance. Speed can be reported as the minimum speed value, average speed value, or maximum speed value in a session or subsession, or as an instantaneous speed report.

Target Heart Rate Range

The target heart rate range [10] is the recommended heart rate for a certain session or subsession. Users can try to keep their heart rate within this range to achieve the preset exercise goal. When the user's actual heart rate exceeds this range, the treadmill directly gives the user a prompt, or sends the corresponding event message to the manager. In a certain session or event, the user should try to keep his/her speed above the lower limit to reach the preset exercise goal.

Target Speed Lower Limit

The target speed lower limit is the minimum speed for a certain session or sport item. The user should try to keep his speed above the lower limit to reach the preset exercise goal. When the user's actual speed exceeds this range, the treadmill directly gives the user a prompt, or sends the corresponding event message to the manager.

Target Energy Consumption Lower Limit

It indicates the minimum energy that should be consumed in a certain session or item. The user should try to consume more energy than the target value to reach the preset exercise goal. When the user's energy consumption value exceeds this target value, the treadmill directly gives the user a prompt, or sends the corresponding event message to the manager.

User's Exercise Standard and Health Status

According to the training goal set by the user in advance, the treadmill will send some key information related to the user's exercise physiological state to the manager in the form of an event report, such as "exceeded the upper limit of the target heart rate range," "reached target energy consumption lower limit" and other information.

Target Heart Rate Distribution Plan

It is set by several "heart rate range + duration" parameter groups. The user's exercise goal is to control his/her heart rate within a specified heart rate range for a certain length of time. Each parameter group contains 3 elements in sequence: the lower limit of the target heart rate range, the upper limit of the target heart rate range, and the duration of the target heart rate range.

VO_{2max}

The maximal rate of oxygen uptake (VO_{2max}) is an important determinant of cardiorespiratory fitness and aerobic performance. VO_{2max} can be estimated indirectly based on the heart rate at rest (HR_{rest}) and the heart rate at maximal exercise (HR_{max})[11].

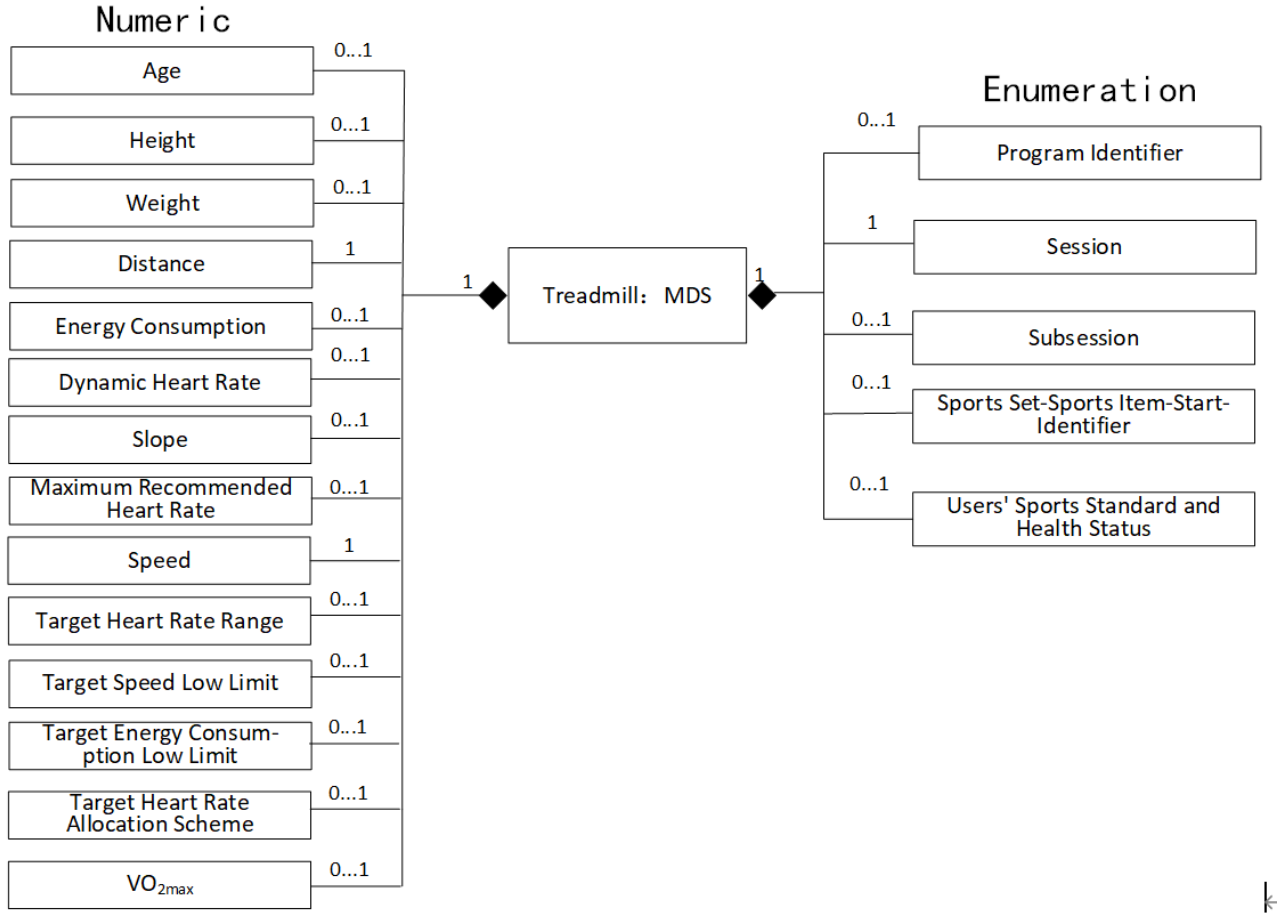
$$VO_{2max} = (15.0 \text{ mL min}^{-1} \text{ kg}^{-1}) \cdot (HR_{max}/HR_{rest})$$

Construction of Treadmill DIM

Treadmill Object Instantiation

Complete the mapping of the parameters mentioned above to the numeric objects and enumerated objects defined by 11073-20601. The object example diagram is illustrated in Figure 1.

Figure 1. The object instance diagram of the treadmill DIM. DIM: domain information model; MDS: medical device system.



Design of the Main Attributes of the Object

For the object instance model related to device information characteristics, it is necessary to further design the attributes of the object, and to achieve the semantic representation of the device information characteristics carried by the object through the definition of attribute values [12]. Instanced objects can be divided into 2 categories: the first category is medical device system (MDS) objects representing context information, and the other category is metric-derived objects representing treadmill user data parameters.

MDS Object

The *Dev-Configuration-Id* attribute holds a locally unique 16-bit identifier that identifies the device configuration. The *System-Id* attribute is an IEEE EUI-64 address, consisting of a 24-bit organizationally unique identifier and a 40-bit manufacturer-defined ID [13]. The agent sends the Dev-Configuration-Id and System-Id to the manager in the "associated state," so that the manager determines the configuration of the slave device during the association. If the manager has saved the configuration information related to *Dev-Configuration-Id* and *System-Id*, then it further identifies the *Dev-Configuration-Id* of the agent, and both agent and manager skip the "configuration state" and enter the "operating status." However, if manager cannot recognize the *Dev-Configuration-Id* of the *System-Id*, then both agent and manager enter the "configuration state" [14].

The attribute value design of the MDS object is shown in Table 1.

Table 1. Object MDS's attributes.

Attribute	The value of attribute
Handle	0
System-Model	{"Manufacturer","Model"}
System-Id	IEEE EUI-64 address
Dev-Configuration-Id	Extended configuration: 0x4000-0x7FFF
System-Type-Spec-List	Types and versions of device specifications: {MDC_DEV_SPEC_PROFILE_HF_CARDIO，3}; Device subtype and version: {MDC_DEV_SUB_SPEC_PROFILE_TREADMILL, 1}

Numeric Object

For the design of attribute values of numeric objects, the main aspects are the following:

- Handle: An unsigned, locally unique, 16-bit number, where each numeric object has a different nonzero handle value.
- Timestamp: All numeric object instances are associated with the session or subsession objects defined above. In the case of a session summary, only the session or subsession should have a timestamp attribute, whereas in the case of continuous monitoring of the session or subsession, the numerical object sampling instance not only reports the session summary attribute, but also each numerical object sampling instance brings its own timestamp attribute.
- Source-Handle-Reference: The session or subsession may contain associated numerical objects which represent observations that are generated throughout the session or subsession. Therefore, the Source-Handle-Reference attribute of a numeric object should identify whether the numeric object instance is associated with a session object or a subsession object. If the numeric object is an observation value at the session level, the Source-Handle-Reference attribute should be equal to the value of the handle of the session object. Similarly, if the numeric object is an observation value at the subsession level, the Source-Handle-Reference attribute should be equal to the value of the handle of the subsession object.
- BasicNuObsValue: In the numerical objects mentioned above, except for the target heart rate range and the target heart rate allocation plan, the basic numerical observations are all represented by the SFLOAT-Type type. Table 2 lists the design of Type, Metric-Spec-Small, and Unit-Code attribute values of other objects except the target heart rate range and target heart rate allocation scheme.

Table 2. Remaining attributes of numeric objects other than Target Heart Rate Range and Target Heart Rate Allocation Scheme.[a]

Object and type	Unit code
Age	
MDC_HF_AGE (126)	MDC_DIM_YR (2368)
Height	
MDC_LEN_BODY_ACTUAL (57668)	MDC_DIM_M (1280)
Weight	
MDC_MASS_BODY_ACTUAL (57664)	MDC_DIM_KILO_G (1731)
Distance	
MDC_HF_DISTANCE (144)	MDC_DIM_M (1280) | MDC_DIM_CENTI_M (1278) | MDC_DIM_STEP (11520)
Energy Consumption	
MDC_HF_ENERGY (196)	MDC_DIM_CAL (8352) | MDC_DIM_JOULES (3968)
Dynamic Heart Rate	
MDC_HF_HR (180)	MDC_DIM_BEAT_PER_MIN (2720)
Speed	
MDC_HF_SPEED (168)	MDC_DIM_M_PER_SEC (2816) | MDC_DIM_CENTI_M_PER_MIN (6577) | MDC_DIM_STEP_PER_MIN (11616) | MDC_DIM_KILO_M_PER_HR (11939)
Target, Speed, and Low Threshold	
MDC_HF_SPEED_TARGET_LOW (2105)	MDC_DIM_M_PER_SEC (2816) | MDC_DIM_CENTI_M_PER_MIN (6577) | MDC_DIM_STEP_PER_MIN (11616)
Target Energy Consumption and Low Threshold	
MDC_HF_ENERGY_EXPENDED_TARGET_LOW (2109)	MDC_DIM_CAL (8352) | MDC_DIM_JOULES (3968)
VO_{2max}	
MDC_HF_VO2_MAX (2112)	MDC_DIM_ML_PER_KG_MIN (4420)
Slope	
MDC_HF_INCLINE (176)	MDC_DIM_PERCENT (544) | MDC_DIM_ANG_DEG (736)

[a]Metric-Spec-Small: mss-avail-intermittent | mss-avail-stored-data | mss-updt-aperiodic | mss-msmt-aperiodic | mss-acc-agent-initiated | mss-cat-setting.

The *Target heart rate range* object uses the *Compound-Basic-Nu-Observed-Value* attribute to transmit the lower and upper limit values of the Target heart rate range. The value of this attribute is only transmitted through a fixed format event report. When the treadmill sends a configuration report, it will report the *Attribute-Value-Map* attribute value of the

target dynamic heart rate range. In the subsequent fixed format reports, the data content can be directly transferred according to that described in the *Attribute-Value-Map* without having to transfer the attribute Object Identifier [15] and the value length, which can reduce the length of the APDU to some extent. Here, the attribute sequence value of the *Attribute-Value-Map* is the attribute-id of the observation attribute, the timestamp attribute of the composite data, and the corresponding attribute value length. The Metric-Structure-Small attribute is used to identify each item of data in the observation list one by one. The order of the *Metric-Id-List* should correspond to the order of the observation items in the composite observation. Here, the first Object Identifier of the *Metric-Structure-Small* attribute value sequence is MDC_HF_HR_TARGET_LOW, and the second is MDC_HF_HR_TARGET_HIGH. For other attributes and their recommended attribute values, please refer to Table 3.

Table 3. Remaining attributes of the object Target Heart Rate Range.

Attribute	The value of attribute
Type	MDC_HF_HR_TARGET_RANGE (2100)
Metric-Spec-Small	mss-avail-intermittent | mss-avail-stored-data | mss-updt-aperiodic | mss-msmt-aperiodic | mss-acc-agent-initiated | mss-cat-setting
Metric-Id-List	First: MDC_HF_HR_TARGET_LOW (2101); Then: MDC_HF_HR_TARGET_HIGH (2102)
Metric-Structure-Small	ms-struct-compound(1)-multiple observations
Unit-Code	MDC_DIM_BEAT_PER_MIN (2720)
Attribute-Value-Map	MDC_ATTR_NU_CMPD_VAL_OBS_BASIC (2677) and MDC_ATTR_TIME_ABS (2439)
Compound-Basic-Nu-Observed-Value	It consists of 2 SFLOAT-Type dates: the first representing target heart rate low threshold and the other one representing high threshold.

The *Target heart rate allocation scheme* object is a data structure, which is set by several parameter groups of "heart rate range + duration + identifier." The user's exercise goal is to control his/her heart rate within a specified heart rate range for a certain length of time.

Each parameter group contains 3 elements in sequence: lower limit of the target heart rate range, upper limit of the target heart rate range, duration of the target heart rate range, and associated content identifier. The first 2 elements are provided by *Compound-Simple-Nu-Observed-Value*, the third element is provided by *Measure-Active-Period*, and the fourth element is provided by *Context-Key*. The value of this attribute is only transmitted via a fixed format event report. The following is an example of a heart rate distribution structure:

{

[70, 100, 180 seconds，"PLAN123"]

[100, 120, 240 seconds，"PLAN123"]

[120, 140, 120 seconds，"PLAN123"]

}

Table 4 illustrates the design of other attributes of the target heart rate allocation scheme.

Table 4. Remaining attributes of the object Target Heart Rate Aallocation Scheme.

Attribute	The value of attribute
Type	MDC_PART_PHD_HF|MDC_HF_HR_TARGET_ALLOC_PLAN
Metric-Spec-Small	mss-avail-intermittent | mss-avail-stored-data | mss-updt-aperiodic | mss-msmt-aperiodic | mss-acc-agent-initiated | mss-cat-setting
Metric-Id-List	First: MDC_HF_HR_TARGET_LOW; Then: MDC_HF_HR_TARGET_HIGH
Metric-Structure-Small	ms-struct-compound(1)-multiple observations
Unit-Code	MDC_DIM_BEAT_PER_MIN
Attribute-Value-Map	First: MDC_ATTR_NU_CMPD_VAL_OBS_SIMP; Second: MDC_ATTR_TIME_PD_MSMT_ACTIVE; Third: MDC_ATTR_CONTEXT_KEY (2680)
Compound-Simple-Nu-Observed-Value	Refer to the text description above.
Measure-Active-Period	The length of the period that each target range in the Target Heart Rate Aallocation Scheme lasts.
Context-Key	The value of this attribute is used to encode and identify different *Target Heart Rate Aallocation* to indicate the difference. Each target range that belongs to the same set of target heart rate allocation schemes uses the same identifier.

Enumeration Object

Table 5 illustrates the attribute value design of enumerated objects, and Table 6 lists the observed values of enumerated objects.

Table 5. Attributes of enumeration objects.

Object and attribute	The value of attribute
Program Identifier, Session, Subsession, Session-Subsession-Strat-identifier, Users' Sports Standard and Health Status	
Handle	An unsigned locally unique 16-bit number.
Type	MDC_HF_PROGRAM_ID (108); MDC_HF_SESSION (123); MDC_HF_SUBSESSION (124); MDC_HF_STRT (125); MDC_HF_USER_FITNESS_HEALTH_STAT (126)
Metric-Spec-Small	mss-avail-intermittent | mss-avail-stored-data | mss-updt-aperiodic | mss-msmt-aperiodic | mss-acc-agent-initiated.
Absolute-Time-Stamp	See the description of the timestamp attribute of the previous numeric object.
Measure-Active-Period	A FLOAT-Type that defines the length of the observation period (in seconds).
Enum-Observed-Value-Simple-Oid (only Object Program Identifier owns)	The value is a free string type and is not restricted by any nomenclature.
Enum-Observed-Value-Simple-Oid (This attribute is owned by all objects except Program Identifier.)	Refer to Table 6.
Source-Handle-Reference	Refer to the footnote.[a]

[a]Source-Handle-Reference: For objects such as Program Identifier, Session-Subsession-Strat-identifier, Users' Sports Standard and Health Status, their Source-Handle-Reference attribute value is the handle of Session or Subsession related to themselves; Subsession's Source-Handle-Reference attribute value is the handle of the Session associated with itself; Session does not have this attribute.

Table 6. Observations of enumeration object.

Object and identifier	Semantic
Session, Subsession, Session-Subsession-Strat-identifier	
MDC_HF_ACT_REST (1001)	Rest
MDC_HF_ACT_UNKNOWN (1007)	Unknown
MDC_HF_ACT_MULTIPLE (1008)	Mix of multiple types of sports
MDC_HF_ACT_RUN (1011)	Jogging
MDC_HF_ACT_WALK (1017)	Walk
MDC_HF_ACT_WATER_WALK (1028)	Walking under water
Users' Sports Standard and Health Status	
MDC_HF_STAT_LT_HR_TARGET_LOW (2200)	The user's heart rate is below the lower limit of the target heart rate range.
MDC_HF_STAT_HT_HR_TARGET_HIGH (2203)	The user's heart rate is above the upper limit of the target heart rate range.
MDC_HF_STAT_HT_SPEED_TARGET_LOW (2207)	The user's speed is higher than the target speed lower limit.
MDC_HF_STAT_HT_ENERGY_EXPENDED_TARGET_LOW (2217)	The user's energy consumption has exceeded the target energy consumption lower limit.

Standard Compliance Testing Process

Because the above data transmission framework is derived from the 11073-20601 optimization exchange protocol, it is necessary to determine whether the data stream sent by the instantiated object that implements this standard meets the 20601 standard [16]. If the instantiated object of the treadmill interoperability framework passes the test, it indicates that the content it sends can have the same semantic grammar as the information sent by other devices that have met the 11073-PHD protocol family [17]. The testing content of this article will focus on the 3 models [18,19] of 11073-PHD, namely, (1) PHD DIM, (2) PHD service model, and (3) PHD communication model.

The test of DIM is mainly based on the events of MDS.

- *MDS-Configuration-Event*: If the manager cannot learn the current agent configuration information from the associated historical records, the agent sends the event to the manager during the startup of the "configuration" state. This event provides static information about the measurement functions supported by the agent.

- *MDS-Dynamic-Data-Update-Var*: This event provides dynamic data (usually measurement data) from the agent for the objects supported by the agent, and reports the object's data in the format of a common attribute list variable.
- *MDS-Dynamic-Data-Update-Fixed*: Use the fixed format defined by the Attribute-Value-Map attribute of the measured object or MDS object to report data. The specific test items are shown in Multimedia Appendix 1 (see the "DIM test" section).

The service model provides the basic function of data access sent between the agent and the manager, and is used to exchange data derived from the DIM. The inspection items mainly include the command to obtain MDS device information (GET) and data report (Event Report). The specific test items are shown in Multimedia Appendix 1 (see the "SER test" section) [20].

The connection state machine defines a series of states and substates experienced between the agent and manager, including states related to connection, association, and operation. The communication model also defines the entry, exit, and error conditions of various states during the various running processes of measurement data transmission, which should be detected. The specific test items are illustrated in Multimedia Appendix 1 (see the "COM test" section).

Test Software Framework Design

Module Design

The test software is mainly divided into 5 modules: Abstract Syntax Notation One (ASN.1) [21] module, encoding module, decoding module, communication module, and test module.

- The ASN.1 module, which defines all data types and data structures of C struct, reuses the ASN.1 code block in the Continua Enabling Software Library (CESL) [22] open source software package provided by Continua in the test software we designed.
- The encoding module generates an APDU binary data stream according to the instantiated APDU object and the Medical Device Encoding Rules used in 11073-20601.
- The decoding module, which refers to the ASN.1 module, converts the binary data stream of the data buffer into an instantiated APDU structure.
- The communication module adopts the abstract factory pattern, calls different subclass factories to produce and initialize instantiated objects of different underlying connection methods, and establishes data connections under the application layer.
- The test module will carry out the testing procedures according to the instantiated object returned by the decoding module, and generate a test result report.

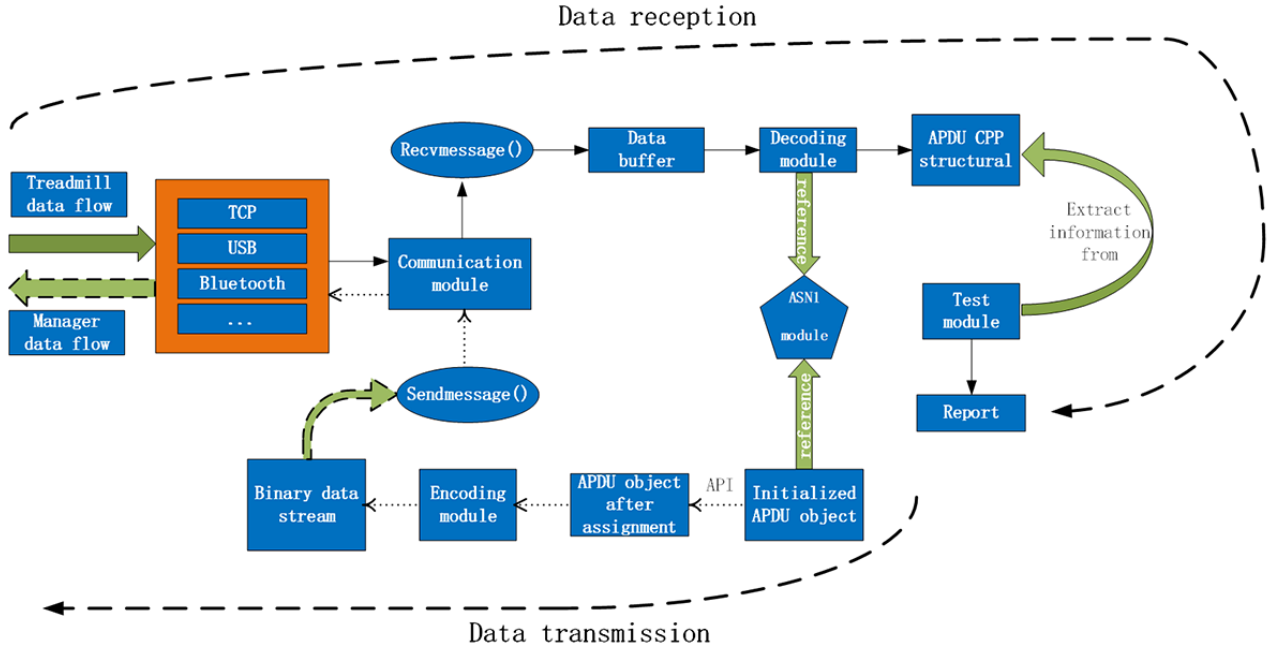
Data Receiving and Testing Process

The data stream sent by the treadmill is transmitted to the application layer listening port of the test software via transmission control protocol (TCP)/USB/Bluetooth/Zigbee or other methods, and then the instantiated object produced by the communication module abstract factory [23] calls the message receiving function to store the binary stream into the data buffer. The decoding module refers to the APDU structure of the ASN.1 module and decodes the binary stream, and then generates the C++ instantiated object of the APDU. The test module calls application programming interface functions according to the designed test items, extracts the data related to the test items from the APDU instantiated objects for testing, and finally generates a test report.

Data Transmission Process

According to the APDU to be sent, refer to the ASN.1 module to establish the initialization APDU object, and then call the application programming interface function to assign the initialization object. The encoding module uses the Medical Device Encoding Rules to encode the assigned APDU object and generate a binary data stream. The communication module calls the message sending function to send the data to the simulated treadmill. The entire workflow of the test software is shown in Figure 2.

Figure 2. The process of receiving and sending data streams in the test software. APDU: application protocol data unit; TCP: transmission control protocol.



Results

Implementation of Treadmill Interoperability Framework

To verify the feasibility of the above standards, we built a simulated treadmill device based on the CESL open source software package. The treadmill device transmits the age, height, weight, maximum recommended heart rate, and other information once using the MDS-Dynamic-Data-Update-Var method (variable format data report); the MDS-Dynamic-Data-Update-Fixed (fixed format data report) method is used to transfer the Session and Subsession, dynamic heart rate, speed, energy consumption, and other information multiple times. The fixed format data report eliminates the description information such as data length and attribute ID. This is because the treadmill includes its own data format context in the configuration report and sends it to the test software before reaching the operating state. For fixed data sent periodically, fixed format data reports can save some byte streams. Figure 3 shows the data sent to the test software by the simulated treadmill acting as an agent.

Figure 3. Information sent by simulated treadmill.

```
Waiting 30 seconds for the transport to connect...
Accepting connection to device ''
[\vascagenttreadmill.cpp(287):Vasc::Treadmill::setMeasurement] MDS-Dynamic-Data-Update-Var method.
[\vascagenttreadmill.cpp(299):Vasc::Treadmill::setMeasurement] User age: 32.
[\vascagenttreadmill.cpp(314):Vasc::Treadmill::setMeasurement] User height: 181cm.
[\vascagenttreadmill.cpp(324):Vasc::Treadmill::setMeasurement] User weight: 82kg.
[\vascagenttreadmill.cpp(335):Vasc::Treadmill::setMeasurement] Maximum recommended heart rate:190 beat/min
[\vascagenttreadmill.cpp(343):Vasc::Treadmill::setMeasurement] Starting session 1,duration:3600s.
[\vascagenttreadmill.cpp(352):Vasc::Treadmill::setMeasurement] MDS-Dynamic-Data-Update-Fixed method.
[\vascagenttreadmill.cpp(361):Vasc::Treadmill::setMeasurement] Setting heart rate to 140 beat/min.
[\vascagenttreadmill.cpp(369):Vasc::Treadmill::setMeasurement] Setting distance to 7000m.
[\vascagenttreadmill.cpp(377):Vasc::Treadmill::setMeasurement] Setting energy to 350CAL.
[\vascagenttreadmill.cpp(385):Vasc::Treadmill::setMeasurement] Setting speed to 3 m/s.
[\vascagenttreadmill.cpp(393):Vasc::Treadmill::setMeasurement] Starting subsession 1,duration:600s.
[\vascagenttreadmill.cpp(402):Vasc::Treadmill::setMeasurement] Setting heart rate to 110 beat/min.
[\vascagenttreadmill.cpp(410):Vasc::Treadmill::setMeasurement] Setting distance to 3000m.
[\vascagenttreadmill.cpp(418):Vasc::Treadmill::setMeasurement] Setting energy to 120CAL.
[\vascagenttreadmill.cpp(426):Vasc::Treadmill::setMeasurement] Setting speed to 2 m/s.
[\vascagenttreadmill.cpp(437):Vasc::Treadmill::setMeasurement] Starting subsession 2.
[\vascagenttreadmill.cpp(445):Vasc::Treadmill::setMeasurement] Setting heart rate to 155 beat/min.
[\vascagenttreadmill.cpp(452):Vasc::Treadmill::setMeasurement] Setting distance to 2000m.
[\vascagenttreadmill.cpp(459):Vasc::Treadmill::setMeasurement] Setting energy to 200CAL.
[\vascagenttreadmill.cpp(466):Vasc::Treadmill::setMeasurement] Setting speed to 4 m/s.
[\vascagenttreadmill.cpp(476):Vasc::Treadmill::setMeasurement] Starting subsession 3,duration:600s.
[\vascagenttreadmill.cpp(484):Vasc::Treadmill::setMeasurement] Setting heart rate to 110 beat/min.
[\vascagenttreadmill.cpp(491):Vasc::Treadmill::setMeasurement] Setting distance to 2000m.
[\vascagenttreadmill.cpp(498):Vasc::Treadmill::setMeasurement] Setting energy to 130CAL.
[\vascagenttreadmill.cpp(505):Vasc::Treadmill::setMeasurement] Setting speed to 1 m/s.
```

Testing Software

Here, the test software also plays the role of a manager, receiving the data stream sent by the treadmill to the binding port through the socket communication method of TCP, completing the test work according to the process, and then generating the final test result set report. The test software

provides TCP, user datagram protocol, Zigbee, and other low-level interface connection methods, and provides optional MDS test attributes in the initial interface. Figure 4 shows the initial interface of the test software, selecting the connection method and test attributes.

Figure 4. Test software start interface.

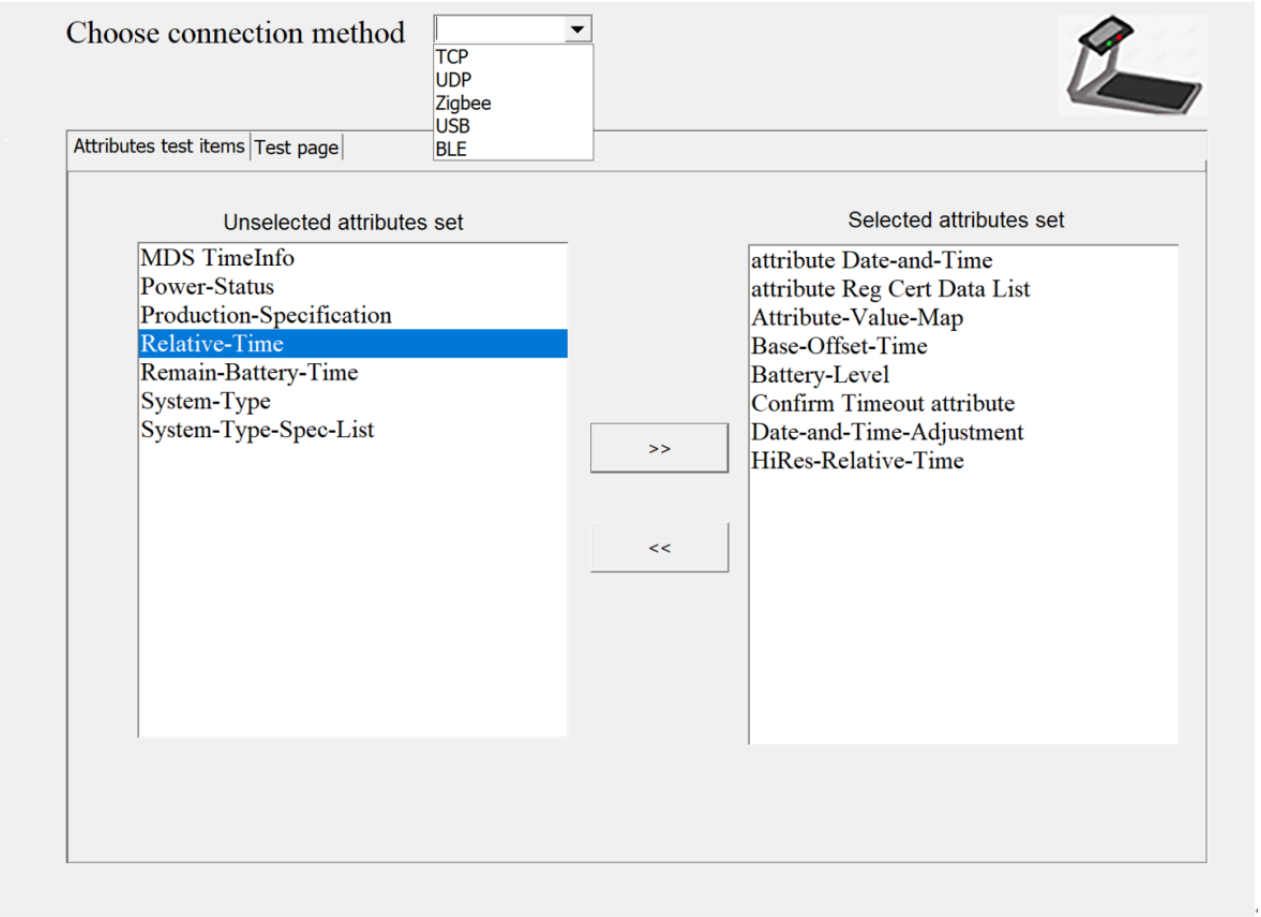


Figure 5 illustrates the test result of device configuration ID. During the association between an agent and a manager, the value of dev-config-id in the “Association Request” message indicates the configuration that the agent wants to use. In the subsequent “Configuration Information Report” and “GET Response” APDU, dev-config-id value should be consistent. In the APDU sent by the simulated treadmill, we deliberately set the value of the dev-config-id in “Association Request” and “GET Response” to 0x4001, and set the value of the dev-config-id in the “Configuration Information Report” to 0x4000. As can be seen in the test report generated by the test software, the consistency check item of dev-config-id has not passed, and it is given its value in the respective APDU.

Figure 6 shows the ongoing communication process between the test software and the treadmill. In the large box on the left side of the interface, we can see the binary data stream and partial decoding information of each APDU in real time; the first small box on the right side of the interface is the objects and attributes contained in the configuration report sent by the treadmill; the second small box is the attribute information of the MDS object; the third small box presents the observation value sent by the treadmill and the corresponding timestamp in real time. After the routine test is completed, the state machine test button is clicked to perform the state machine test. After all the test items are tested, a test report will be generated, and the results will be displayed in a list. Figure 7 demonstrates a small part of the results of the final test report.

Figure 5. The value of dev-config-id in different APDUs and its consistency test results. APDU: application protocol data unit.

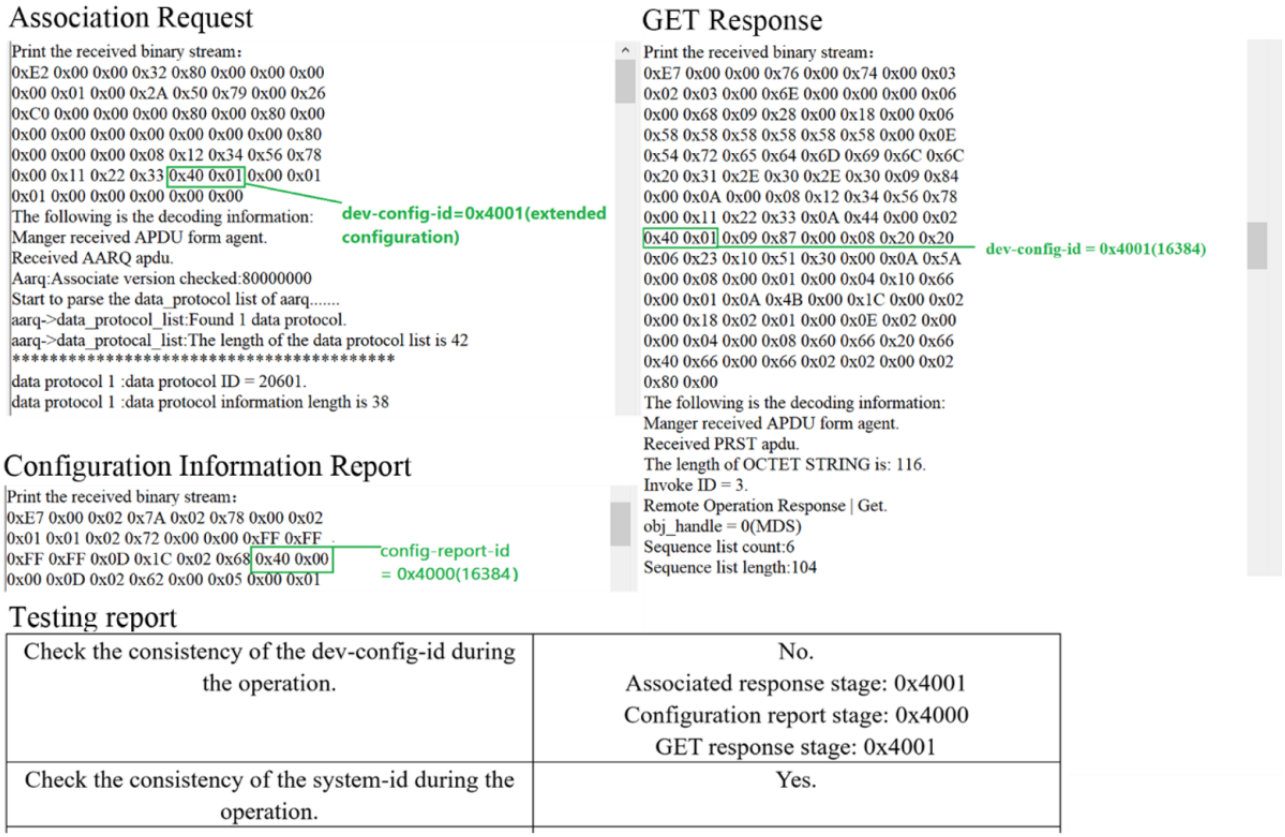


Figure 6. Data transmission between test interface and treadmill.

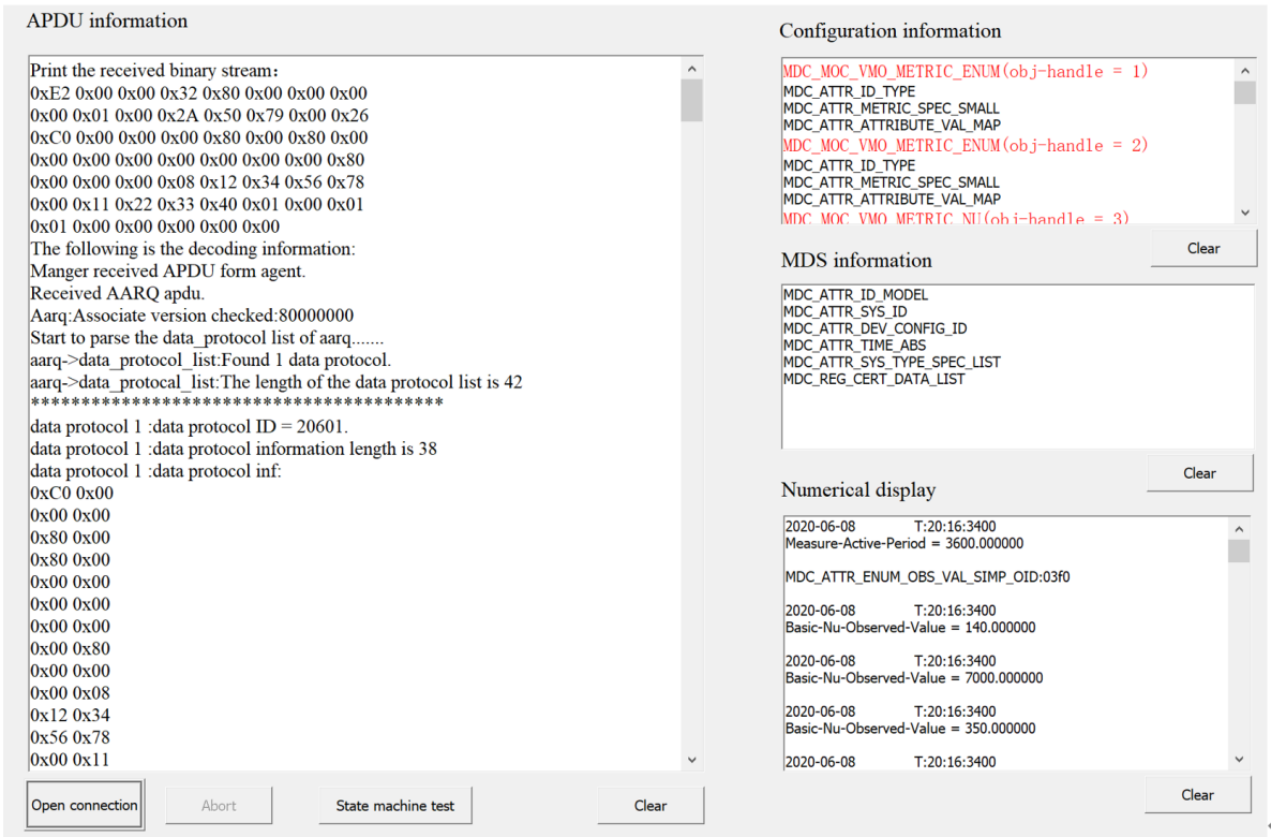


Figure 7. Part of the test results.

Is the type of apdu PrstApdu?	Yes
Does the obj-handle equal to 0x00 0x00(MDS_obj)?	Yes
Does PHD support relative time clock?	Yes
Is the event type (EventReportArgumentSimple) MDC_NOTI_CONFIG?	Yes
Is the config-report-id between 0x00 0x01 and 0x7f 0xff?	Yes
Is each obj-classes numeric or enumerated?	Yes
Does each obj-handle have a unique and non-zero value?	Yes
Check each object's each attribute has the attribute-id between 0x0913 (2323) and 0x0A77 (2679)> or <between 0xF000(61440) and 0xFBFF(64511)>.	Yes
Verify if the invoke-id is mirrored from the Get request.	Yes
Verify if the DataApdu contains the SEQUENCE GetResultSimple.	Yes
Verify if the GetResultSimple.obj-handle = 0x00 0x00.	Yes
Is the number of implemented attributes that are included in the GET response greater than 3?	Yes
Does it contain mandatory attribute System-Model?	Yes
Does it contain mandatory attribute System-Id?	Yes
Does it contain mandatory attribute Dev-Configuration-Id?	Yes

Discussion

In this article, we propose a treadmill data interoperability protocol based on 11073-PHD, and design a set of standard compliance testing methods that match it. Using the testing software, we tested the data stream sent by the simulated treadmill equipment and generated a corresponding test result report.

In previous work, most manufacturers of sports and health equipment such as treadmills have their own set of data transmission standards, which is very unfavorable for data integration analysis and processing between different manufacturers and different sports and health equipment. In our work, through tailoring and customizing the existing 11073-PHD, we designed a set of protocol standards suitable for the transmission of treadmill data. This not only provides a possibility to unify the data transmission standards of treadmill equipment among various manufacturers, but more importantly, it also provides an idea for unifying the application layer data format of other sports and health equipment. Sports health equipment is designed based on the 11073-PHD-based customized design, so that they have the same semantic syntax, making it possible for a gateway device to integrate multiple sports health data.

We have investigated 4 popular treadmill private protocols used in the market to transmit key data (Table 7), and compared all their functions with the standard protocols we developed. While Hlink's running posture detection data have no corresponding functional objects, the key data-bearing function objects established by our interoperability framework can cover all the main data of the 4 devices. A unified semantic syntax can help expand and upgrade service capabilities, which may greatly facilitate remote data capture, thereby enhancing the remote interaction between service providers and users.

Table 7. Comparison of proprietary protocols and standards.

Private standard and key data	Standard object
SOLE	
Pulse (beats/minute)	Dynamic heart rate (beats/minute)
Distance (km)	Distance (km)
Calories	Energy consumption (kcal)
User profile	Age (years), weight (kg), height (cm), user's exercise standard and health status
Program name	Program identifier
Speed (km/h)	Speed (km/h)
Slope (degree)	Slope (degree)
Hlink (HUAWEI)	
Calories	Energy consumption (kcal)
Heart rate (beats/minute)	Dynamic heart rate (beats/minute)
Distance	Distance (km)
Speed	Speed (km/h)
Steps	Distance (steps)
Program	Program identifier
Running posture	—[a]
Keep	
Maximum heart rate (beats/min)	Maximum recommended heart rate (beats/min)
Sports set	Session
Calories	Energy consumption (kcal)
Step frequency	Speed (steps/min)
Speed	Speed (km/h)
Distance	Distance (km)
IOT (XIAOMI)	
Speed	Speed (km/h)
Distance	Distance (km)
Steps	Distance (steps)
Calories	Energy consumption (kcal)
Mode	Program identifier
Slope (%)	Slope (%)

[a]—: not available.

However, this work plan only supports some common data information functions of treadmills in the usual sense. Some plans, such as Hlink's running posture detection, are not completely covered. This requires a more comprehensive arrangement and improvement in the next step. In addition, the treadmill we define is just acting as an agent. However, if you add some additional equipment that can be connected to a treadmill, such as a sports watch, the treadmill plays a dual role. When the treadmill is responsible for receiving data from the sports watch, it acts as a master device; at the same time, the treadmill transmits all its data to the gateway device. At this time, it acts as an agent device. The above situation covers only a small number of applications in the treadmill market, and our standard is only applicable for treadmills with common features at this stage. Finally, there is a lack of information expression regarding the working state of the treadmill itself (the working state of the electronic control board and the working state of the sensing components). Further information describing whether the speed and slope adjustment unit is working properly can be added.

Acknowledgments

The main support for this study came from the National Key Research and Development Program of China (No. 2018YFC2000804), the National Natural Science Foundation of Chongqing (No. cstc2020jcyj-msxmX0641), and the Fundamental Research Funds for the Central Universities (No. 2020CDJ-LHZZ-025).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Testing method.
[DOCX File , 19 KB - medinform_v8i12e22000_app1.docx]

References

1. World Health Organization. Global action plan on physical activity 2018-2030: more active people for a healthier world. Geneva, Switzerland: World Health Organization; 2018:10-20.
2. Hu H, Elkus A, Kerschberg L. A Personal Health Recommender System incorporating personal health records, modular ontologies, and crowd-sourced data. 2016 Presented at: IEEE/ACM International Conference on Advances in Social Networks Analysis & MiningACM; 2016; Austin, TX. [doi: 10.1109/asonam.2016.7752367]
3. Lim JH, Park C, Park SJ. Home healthcare settop-box for senior chronic care using ISO/IEEE 11073 PHD standard. Annu Int Conf IEEE Eng Med Biol Soc 2010;2010:216-219. [doi: 10.1109/IEMBS.2010.5627845] [Medline: 21097184]
4. Yuksel M, Dogac A. Interoperability of medical device information and the clinical applications: an HL7 RMIM based on the ISO/IEEE 11073 DIM. IEEE Trans Inf Technol Biomed 2011 Jul;15(4):557-566. [doi: 10.1109/TITB.2011.2151868] [Medline: 21558061]
5. Trigo JD, Chiarugi F, Alesanco A, Martínez-Espronceda M, Serrano L, Chronaki CE, et al. Interoperability in digital electrocardiography: harmonization of ISO/IEEE x73-PHD and SCP-ECG. IEEE Trans Inf Technol Biomed 2010 Nov;14(6):1303-1317. [doi: 10.1109/TITB.2010.2064330] [Medline: 20699215]
6. Nguyen TN, Su S, Celler B, Nguyen H. Advanced portable remote monitoring system for the regulation of treadmill running exercises. Artif Intell Med 2014 Jun;61(2):119-126. [doi: 10.1016/j.artmed.2014.05.002] [Medline: 24877618]
7. Stevens SL, Morgan DW. Heart rate response during underwater treadmill training in adults with incomplete spinal cord injury. Top Spinal Cord Inj Rehabil 2015;21(1):40-48 [FREE Full text] [doi: 10.1310/sci2101-40] [Medline: 25762859]
8. Kawai T. An attempt to design optimal personalized exercise prescriptions using the KEIO-SENIOR treadmill protocol for patients with type 2 diabetes. Personalized Medicine Universe 2016 Jul;5:27-31. [doi: 10.1016/j.pmu.2015.12.001]
9. Vandenberk T, Stans J, Mortelmans C, Van Haelst R, Van Schelvergem G, Pelckmans C, et al. Metadata Correction: Clinical Validation of Heart Rate Apps: Mixed-Methods Evaluation Study. JMIR Mhealth Uhealth 2018 Mar 14;6(3):e19 [FREE Full text] [doi: 10.2196/mhealth.9509] [Medline: 29537967]
10. Sebastian LA, Reeder S, Williams M. Determining target heart rate for exercising in a cardiac rehabilitation program: a retrospective study. J Cardiovasc Nurs 2015;30(2):164-171. [doi: 10.1097/JCN.0000000000000154] [Medline: 24866048]
11. Uth N, Sørensen H, Overgaard K, Pedersen PK. Estimation of VO2max from the ratio between HRmax and HRrest--the Heart Rate Ratio Method. Eur J Appl Physiol 2004 Jan;91(1):111-115. [doi: 10.1007/s00421-003-0988-y] [Medline: 14624296]
12. IEEE Draft Standard for Health Informatics - Personal Health Device Communication - Part 20601: Application Profile - Optimized Exchange Protocol. In: IEEE P11073-20601/D7. New York, NY: IEEE; Nov 5, 2013:1-252.
13. Carot-Nemesio S, Santos-Cadenas JA, Quirós PH, Bustos J. OpenHealth - The OpenHealth FLOSS Implementation of the ISO/IEEE 11073-20601 Standard. 2010 Presented at: HEALTHINF 2010 - Proceedings of the Third International Conference on Health Informatics; January 20-23, 2010; Valencia, Spain. [doi: 10.5220/0002766705050511]
14. Caranguian LP, Pancho-Festin S, Sison LG. Device interoperability and authentication for telemedical appliance based on the ISO/IEEE 11073 Personal Health Device (PHD) Standards. New York, NY: IEEE; 2012 Presented at: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society; August 28 to September 1, 2012; San Diego, CA p. 1270-1273. [doi: 10.1109/EMBC.2012.6346169]
15. Lee Y. Personal Medical Monitoring System: Addressing Interoperability. IT Professional 2013 Sep;15(5):31-37. [doi: 10.1109/mitp.2012.90]
16. Trigo JD, Chiarugi F, Alesanco A, Martínez-Espronceda M, Chronaki CE, Escayola J, et al. Standard-compliant real-time transmission of ECGs: harmonization of ISO/IEEE 11073-PHD and SCP-ECG. Annu Int Conf IEEE Eng Med Biol Soc 2009;2009:4635-4638. [doi: 10.1109/IEMBS.2009.5332677] [Medline: 19963856]
17. Dingler M, Dietz C, Pfeiffer J, Lueddemann T, Lüth T. A framework for automatic testing of medical device compatibility. New York, NY: IEEE; 2015 Jul 13 Presented at: 2015 13th International Conference on Telecommunications (ConTEL); 13-15 July 2015; Graz, Austria p. 1-8. [doi: 10.1109/contel.2015.7231211]

18. Park CY, Lim JH, Park S. ISO/IEEE 11073 PHD adapter board for standardization of legacy healthcare device. 2012 Jan Presented at: 2012 IEEE International Conference on Consumer Electronics (ICCE); January 13-16, 2012; Las Vegas, NV p. 482-483. [doi: 10.1109/icce.2012.6161953]
19. International Telecommunication Union. Conformance of ITU-T H.810 personal health system: Personal Health Devices interface Part 1: Optimized Exchange Protocol: Personal Health Device. Geneva, Switzerland: International Telecommunication Union; 2018:25-88.
20. ISO/IEEE Health informatics — Personal health device communication — Part 10417: Device specialization — Glucose meter. In: IEEE/ISO 11073-10417-2010. New York, NY: IEEE; May 01, 2010:21-38.
21. Christoph P, Czerwonka R. Method for Converting Initial Data Into Target Data According to ASN.1. 2013. URL: https://www.freepatentsonline.com/WO2013182634.html [accessed 2020-03-18]
22. Benner-Wickner M, Schope L. Using Continua Health Alliance Standards - Implementation and Experiences of IEEE 11073. In: 2011 IEEE 12th International Conference on Mobile Data Management. 2011 Presented at: Mobile Data Management (MDM), 2011 12th IEEE International Conference; 2011; Lulea, Sweden p. 40-45. [doi: 10.1109/MDM.2011.25]
23. Ellis B, Stylos J, Myers B. The Factory Pattern in API Design: A Usability Evaluation. New York, NY: IEEE; 2007 May Presented at: 29th International Conference on Software Engineering (ICSE'07); May 20-26, 2007; Minneapolis, MN p. 302-312. [doi: 10.1109/ICSE.2007.85]

Abbreviations

APDU: application protocol data unit
ASN.1: Abstract Syntax Notation One
CESL: Continua Enabling Software Library
DIM: domain information model
HR_{max}: heart rate at maximal exercise
HR_{rest}: heart rate at rest
MDS: medical device system
TCP: transmission control protocol

Edited by C Lovis; submitted 02.07.20; peer-reviewed by A Samuel; comments to author 26.09.20; revised version received 28.10.20; accepted 07.11.20; published 09.12.20.

Please cite as:
Huang ZY, Wang Y, Wang L
ISO/IEEE 11073 Treadmill Interoperability Framework and its Test Method: Design and Implementation
JMIR Med Inform 2020;8(12):e22000
URL: http://medinform.jmir.org/2020/12/e22000/
doi:10.2196/22000
PMID:33295293

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org